Proceedings of the National STEM Assessment Conference

October 19–21, 2006

Omni Shoreham Hotel
Washington, D.C.

Hosted by the National Science Foundation and Drury University

Chief Editors: Donald Deeds and Bruce Callen, Drury University

This project supported by the National Science Foundation under Grant DUE-0243207

Proceedings of the National STEM Assessment Conference

October 19-21, 2006

Omni Shoreham Hotel
Washington, D.C.

Hosted by the National Science Foundation
and Drury University

Chief Editors: Donald Deeds and Bruce Callen,
Drury University

Editing and Layout: Open Water Media, Inc.
(www.openwatermedia.com)

©Drury University 2007

Table of Contents

The Intersection of STEM Assessment, Accountability, and Education Policy: A "Gathering Storm" for Higher Education? 3
Labov (Opening Plenary Speaker)

From the Grassroots: Provocative Insights & Compelling Research on the Learning of Undergraduates in STEM Fields 12
Narum (Closing Panel Moderator)

Assessment and Accountability from a Scientific Point of View: What is the Evidence? 16
Oxtoby (Opening Plenary Speaker)

Curriculum Improvement in Practice-based Biology Programs Using Student E-Portfolios: A Progress Report 27
Bergman, Porter, Poklop, Aman, Noyes, Woolfson

A Self-regulated Learning Assessment System for Electromechanical Engineering Technology Students 37
Blank, Hudesman, Morton, Armstrong, Moylan, White, Zimmerman

Rapid Assessment for Improved Student Learning and Satisfaction 46
Chen, Whittinghill, Kadlowec

Transferable Assessments for Capstone Engineering Design 53
Davis, Beyerlein, Harrison, Thompson, Trevisan

The Calculus Concept Inventory 60
Epstein

Developing and Assessing Student Scientific Abilities 68
Etkina, Brookes, Murthy, Karelina, Villasenhor, Van Heuvelen

The Design and Validation of the E-NSSE and E-FSSE Surveys of Student Engagement in Engineering 81
Drewery, Bjorklund, Fortenberry

Assessment Resource Tools for Assessing Students' Statistical Literacy, Reasoning, and Thinking 93
Garfield, DelMas, Ooms, Chance

Using Self-assessment Within an Electronic Portfolio as a Framework for Development of Student Problem-solving and Analysis Abilities Across Courses and Majors Within Science and Mathematics 100
Guilbault, Dollhopf, Pustejovsky, Thompson, Truchan, Young

Assessing Student Learning and Evaluating Faculty Capacity Development in the NSF-funded Regional Workshops Project NSF/DUE/ASA 0127725 110
Himangshu, Luli

Assessing Problem-solving Strategies in Chemistry Using the IMMEX System 118
Cooper, Stevens, Holme

Ed's Tools: A Web-based Software Toolset for Accelerated Concept Inventory Construction 130
Garvin-Doxas, Doxas, Klymkowsky

ClassAction: A Model Rapid-feedback and Dynamic Formative Assessment System 141
Lee, Guidry, Schmidt, Slater, Young

Development of the Geoscience Concept Inventory 148
Libarkin, Anderson

Measuring What Students Know about How to Learn Chemistry 159
Grove, Lowery Bretz

National Dissemination and Research on Professional Development: Supporting Assessment in Undergraduate Mathematics 166
Madison

Formative Assessment Materials for Large-enrollment Physics Lecture Classes 173
Meltzer

Calibrated Peer Review for ABET Assessment 190
Berry, Carlson, Millard

Computer Science Assessment Instrument Development: Evaluating Attitudes and Outcomes 194
Moskal, Behrens, Guzdial, Tew, Dann, Cooper

Designing a Peer Evaluation Instrument that is Simple, Reliable, and Valid: The Comprehensive Assessment of Team-Member Effectiveness 202
Ohland, Loughry

Development of Faculty Collaboratives to Assess Achievement of Student Learning Outcomes in Critical Thinking in Biology Core Courses 209
Ono, Casem, Hoese, Houtman, Kandel, McClanahan

Tracing Matter and Energy Through Biological and Geological Systems: Fundamental Problems of Students' Understanding 219
Wilson, Anderson, Heidemann, Long, Merrill, Merritt, Richmond, Sibley, Urban-Lurain, Parker

ChemQuery: An Assessment System for Mapping Student Progress in Learning Chemistry 227
Scalise, Claesgens, Wilson, Stacy

Discovering Functional Role Specialization: The Effect on Undergraduate Student Learning of Engineering Project Teams 240
Leonard, Schmidt, Schmidt, Smith

Assessing Pedagogical Content Knowledge of Inquiry Science Teaching: Developing an Assessment Instrument 247
to Support the Undergraduate Preparation of Elementary Teachers to Teach Science as Inquiry
Schuster, Cobern, Applegate, Schwartz, Vellom, Undrelu

Calibrated Peer Review™ in Math and Science: Research and Practice 266
Simpson, Russell, Gunersel, Layne

Developing Assessments That Help Students Learn to Develop Open-ended Designs 274
Sims-Knight, Upchurch, Fortier

Project CAT: Assessing Critical Thinking Skills 290
Stein, Haynes, Redding

Developing an Outcomes Assessment Instrument for Identifying Engineering Student Misconceptions in Thermal and Transport Sciences 300
Streveler, Miller, Nelson, Geist, Olds

The Signals and Systems Concept Inventory 307
Wage, Buck, Hjalmarson

Assessing Students' Value for Science and Math Literacy 314
Wood, Luttrell, Callen, Allen, Deeds

Collaborative Project: Using Students' Design Tasks to Develop Scientific Abilities 326
Zou

Abstracts for Assessment of Student Achievement Awards, 2001-2005 332

Index 359

Lead Author Contact Information

*An * beside a name in the title bar denotes a lead author*

Kostia Bergman
Biology Department
Northeastern University
(k.bergman@neu.edu)

Seymour Blank
Department of Computer/Electromechanical Engineering Technology
New York City College of Technology
City University of New York
(sblank@citytech.cuny.edu)

John Chen
Department of Mechanical Engineering
Rowan University
(jchen@rowan.edu)

Denny Davis
Washington State University
(davis@wsu.edu)

Jerome Epstein
Mathematics Department
Polytechnic University
(jepstein@duke.poly.edu)

Eugenia Etkina
Graduate School of Education
Rutgers University
(etkina@rci.rutgers.edu)

Norman Fortenberry
Center for the Advancement of Scholarship on Engineering Education
National Academy of Engineering
(nfortenb@nae.edu)

Joan Garfield
University of Minnesota
(jbg@umn.edu)

Lauralee Guilbault
Alverno College
(Lauralee.Guilbault@alverno.edu)

Sumi Himnagshu
Macon State University
(sumitra.himangshu@maconstate.edu)

Tom Holme
Department of Chemistry and Biochemistry
University of Wisconsin - Milwaukee
(tholme@csd.uwm.edu)

Mike Klymkowski
The Bioliteracy Project
University of Colorado, Boulder
(Michael.Klymkowsky@Colorado.edu)

Kevin Lee
University of Nebraska
(klee6@unl.edu)

Julie Libarkin
Geological Sciences & Division of Science and Mathematics Education
Michigan State University
(libarkin@msu.edu)

Brad Lister
Anderson Center for Innovation in Undergraduate Education
Rensselaer Polytechnic Institute
(listeb@rpi.edu)

Stacey Lowery-Bretz
Department of Chemistry & Biochemistry
Miami University
(bretzsl@muohio.edu)

Bernie Madison
University of Arkansas
(bmadison@uark.edu)

David Meltzer
Department of Physics
University of Washington
(dmeltzer@u.washington.edu)

Don Millard
Academy of Electronic Media
Rensselaer Polytechnic Institute
(millard@rpi.edu)

Barbara Moskal
Mathematical and Computer Sciences
Colorado School of Mines
(bmoskal@mines.edu)

Matt Ohland
Engineering Education
Purdue University
(ohland@purdue.edu)

Joyce Ono
Department of Biological Science
California State University, Fullerton
(jono@fullerton.edu)

Kathleen Scalise
Educational Leadership — Applied Measurement
University of Oregon
(kscalise@uoregon.edu)

Linda Schmidt
Mechanical Engineering
University of Maryland
(lschmidt@umd.edu)

David Schuster
Western Michigan University
(david.schuster@wmich.edu)

Nancy Simpson
Center for Teaching Excellence
Texas A&M University
(n-simpson@tamu.edu)

Judith E. Sims-Knight
Psychology
University of Massachusetts
Dartmouth
(jsimsknight@umassd.edu)

Barry Stein
Tennessee Technological University
(BStein@tntech.edu)

Ruth Streveler
Engineering Education
Purdue University
(rastreve@purdue.edu)

Kathleen Wage
George Mason University
(k.e.wage@ieee.org)

Christopher Wilson
Michigan State University
(wilso250@msu.edu)

Mark Wood
Department of Chemistry
Drury University
(mwood@drury.edu)

Xueli Zou
Physics
California State University, Chico
(xzou@csuchico.edu)

The Intersection of STEM Assessment, Accountability, and Education Policy: A "Gathering Storm" for Higher Education?

Jay B. Labov[1] - *National Research Council*

Plenary Speaker

Jay Labov is senior staff member of the National Research Council's Center for Education. Dr. Labov is the Senior Advisor for Education and Communications, splitting his time between the Center for Education and the National Academies' Office of Communications. In this capacity, Dr. Labov leads an institution-wide effort to leverage the National Academies' work in education by helping to make more deliberate connections between the work of the Center for Education, the National Academy of Sciences, National Academy of Engineering, and the program units of the National Research Council. He is the principal liaison on education activities between the program units of the National Academies and its Office of Communications, with the goal of enhancing communication with outside stakeholders about the Academies' work in education and the public's understanding of science and technology.

Dr. Labov also has been the study director for several NRC reports, including: *Evaluating and Improving Undergraduate Teaching in Science, Mathematics, Engineering, and Technology* (2003); *Learning and Understanding: Improving Advanced Study of Mathematics and Science in U.S. High Schools* (2002); *Educating Teachers of Science, Mathematics, and Technology: New Practices for the New Millennium* (2000); *Transforming Undergraduate Education in Science, Mathematics, Engineering, and Technology* (1999); *Serving the Needs of Pre-College Science and Mathematics Education: Impact of a Digital National Library on Teacher Education and Practice* (1999); and *Developing a Digital National Library for Undergraduate Science, Mathematics, Engineering, and Technology Education* (1998). He has been Director of the Center's Committee on Undergraduate Science Education and oversees the National Academy of Science's efforts to improve the teaching of evolution in the public schools. Prior to assuming his position at the NRC, Dr. Labov was a member of the biology faculty for 18 years at Colby College in Waterville, Maine.

INTRODUCTION

I was honored to be asked to deliver one of the opening keynote addresses for this conference in October, 2006. I believe I was asked to speak because I have had the good fortune of seeing higher education from several perspectives. Prior to beginning my work for the National Research Council (NRC) and the National Academy of Sciences, I had been a member of the faculty in the Department of Biology at Colby College in Maine, so I had direct experience with some of the issues confronting higher education faculty and institutions around changing expectations for teaching, student learning, and balancing faculty responsibilities. My work at the NRC has allowed me to examine higher education from more of a policy perspective. Here I was able to view higher education's role and responsibilities (including accountability) in contributing to the nation's well-being and competitiveness from a level that many individual faculty typically do not think about.

---

1. All statements and opinions expressed in this essay are the author's alone and do not necessarily represent any position of the National Research Council.

My assignment for this presentation was to integrate these perspectives for faculty who have focused primarily on science, technology, engineering, or mathematics (STEM) but who are now being supported by the National Science Foundation to devote a significant portion of their professional lives to undertaking research and implementation projects that would make assessment a more fundamental component of STEM education.

A CONFLUENCE OF "GATHERING STORMS"

At the time of that conference, several "storms" were building that I tried to summarize from my vantage point as a staff member at the National Research Council. They included:

Assessment and Accountability in K-12 and Higher Education:

There was increasing discussion and concern expressed from within a number of circles in higher education about a growing emphasis on accountability for student learning and what that might actually mean (see for example [1,2]). These concerns were catalyzed by a series of statements from Secretary of Education Margaret Spellings (Table 1) and a then-to-be released report from a national commission that she appointed to address various issues in higher education.[2]

Table 1. The Nature of Recent Policy and Political Storms: Statements from U.S. Education Secretary Spellings (September 22, 2006)

"Believe it or not, we can't answer the most critical and basic questions about student performance and learning at colleges and that's unacceptable," … "Information will not only help with decision -making —it will also hold schools accountable for quality. We want to work with Congress, states and institutions to build a system that is more useful and widely available to every student.	"Right now, accreditation is the system we use to put a stamp of approval on higher education quality. It's largely focused on inputs, more on how many books are in a college library, than whether students can actually understand them. Institutions are asked "Are you measuring student learning?" And they check yes or no. That must change. Whether students are learning is not a yes or no question ... It's how? How much? And to what effect? "

http://www.ed.gov/news/speeches/2006/09/09262006.html

2. The final version of the Commission's report, *A Test of Leadership: Charting the Future of U.S. Higher Education,* is available at http://www.ed.gov/about/bdscomm/list/hiedfuture/reports/final-report.pdf. Additional information about the Commission and its activities is available at http://www.ed.gov/about/bdscomm/list/hiedfuture/index.html.

Whether the kinds and levels of accountability that are currently in place in the nation's K-12 education system can or should be applied to higher education is continuing to receive a great deal of attention from both governmental and private sectors, including higher education itself. [3,4,5] Currently, the federal No Child Left Behind Act (NCLB) mandates that every child in public education in the United States must be tested annually in reading and mathematics while they are in Grades 3 through 8, and once in high school. Data from these tests must be disaggregated to measure the performance of students by gender, ethnicity, socioeconomic status, English language learners, and disability. Under the Act, responsibility for producing and administering the tests is delegated to the individual states. However, each state also must have developed a plan for student advancement as part of what the Act calls Adequate Yearly Progress (AYP); by law all students are supposed to be proficient in reading and mathematics by 2014.

Science will be tested under NCLB beginning in the 2007-8 school year. NCLB mandates that all students be tested at least once in every grade band (elementary, middle, and secondary grades). As of now, achievement in science will not be considered as part of the formula for measuring AYP although there have been several attempts in Congress to change this aspect of the law. At the request of the National Science Foundation, the National Research Council produced a report that examined the potential effects of this testing program on science education [3].

NCLB also calls for states to have aligned their content standards for the aforementioned subjects with their assessment systems by 2006 (at the time of my presentation in October 2006, only nine states had developed plans that experts commissioned by the U.S. Department of Education certified as satisfying all six criteria outlined in the guidance provided by the Department. Thirty-nine states had submitted plans that met some of the criteria, and four states' plans were judged as not meeting any of the criteria[6]). Forty-nine of 50 states now have their own

---

3. For example, the American Enterprise Institute convened a conference on March 13, 2007, *Higher Education, after the Spellings Commission: An Assessment*. For more information see http://www.aei.org/events/eventID.1476,filter.all/event_detail.asp.

4. On June 20, 2007, *the Chronicle of Higher Education* reported [http://chronicle.com/news/index.php?id=2534?=atnb] that the U.S. Department of Education plans to set aside $2.5-million in the budget for the Fund for the Improvement of Postsecondary Education (FIPSE, http://www.ed.gov/about/offices/list/ope/fipse/index.html), for a focused competition on student-learning assessment. The money will help in developing methods to "measure, assess, and report student achievement and institutional performance at the postsecondary level," according to a notice that appeared in the Federal Register - http://a257.g.akamaitech.net/7/257/2422/01jan20071800/edocket.access.gpo.gov/2007/E7-11860.htm. When combined with another set aside of $500,000, this amount will account for approximately half of FIPSE's budget. Unrestricted funds will enable the awarding of 18-18 grants for the 2007 budget year, approximately one third of the number of grants available in the previous year.

5. On June 26, 2007 the *Chronicle of Higher Education* reported (http://chronicle.com/daily/2007/06/2007062602n.htm?=attn) that two higher education organizations (the National Association of State Universities and Land-Grant Colleges and the Association of American Colleges and Universities) are planning to vote this autumn on whether their member institutions should implement a voluntary system of accountability that would allow for easier and more useful comparisons of institutions by the public. Part of this system would include each participating institution posting measurements of student learning outcomes from one of three assessments (the Collegiate Assessment of Academic Proficiency, the Collegiate Learning Assessment, and the Measure of Academic Proficiency and Progress).

6. For additional information see http://www.ed.gov/news/pressreleases/2006/08/08162006a.html. As of May 8, 2007, ten states have received full approval from the Department of Education's standards/assessment peer review process; the rest of the states are in the process of submitting additional evidence for approval. Source: Council of Chief State School Officers (personal communication).

set of standards in science and other subjects. In science, these state standards are based to some extent on national standards documents published by the American Association for the Advancement of Science [4] or the National Research Council [5]. These documents call for fundamentally different approaches to what constitutes content in science (Table 2). The NRC's *National Science Education Standards* also provide standards for how science should be taught and assessed (Tables 3 and 4), and for professional development of teachers of science (Table 5). In the STEM disciplines, national standards also have been produced in mathematics [6,7] and for technology education [8]. Analyses of these standards suggest that, at least for science, there is considerable variance in what students are expected to know and be able to do (e.g., [9,10]). Concerns also have been expressed about how science will be tested; if these tests do not assess higher order thinking and learning, the vision for new approaches to science education that are expressed in the national standards could be severely compromised (e.g., [3]).

The law also calls for a "highly qualified" teacher to be in every U.S. classroom. For NCLB, "highly qualified" means that a teacher must be fully certified or licensed, hold a bachelor's degree in a subject area, and show competence in subject knowledge and teaching skills (generally demonstrated by passing a rigorous state test). The timetable for achieving this mandate was the 2005-2006 school year, but the Department of Education has pushed back the date because so many states had failed to comply by that time.

Table 2. Content Topics in the *Benchmarks for Science Literacy and the National Science Education Standards**

AAAS BENCHMARKS:	NATIONAL SCIENCE EDUCATION STANDARDS:
• The Nature of Science	
• The Nature of Mathematics	
• The Nature of Technology	• Science and Technology
• The Physical Setting	• Physical/Earth/Space Sciences
• The Living Environment	• Life Sciences
• The Human Organism	• Science in Personal/Social Perspectives
• Human Society	
• The Designed World	
• The Mathematical World	
• Historical Perspectives	• History and Nature of Science
• Common Themes	• Unifying Concepts and Processes
• Habits of Mind	• Science as Inquiry

*Topics are organized such that similar topics in each column are displayed on the same row.

Table 3. The *National Science Education Standards* Stress a Changing Emphasis on Scientific Content

LESS EMPHASIS ON:	MORE EMPHASIS ON:
• Knowing scientific facts and information	• Understanding science processes and developing abilities of inquiry
• Studying subject matter disciplines (e.g., physics, earth sciences) for their own sake	• Learning subject matter disciplines in the context of inquiry, technology, science in personal and social perspectives, and history and nature of science
• Separating science knowledge and science process.	• Integrating all aspects of science content
• Covering many science topics	• Studying a few fundamental science concepts
• Implementing inquiry as a set of processes	• Implementing inquiry as instructional strategies, abilities, and ideas to be learned
	([5], p. 113)

Table 4. The *National Science Education Standards* Stress a Changing Emphasis on Assessment of Scientific Knowledge and Understanding

LESS EMPHASIS ON:	MORE EMPHASIS ON:
• Assessing discrete knowledge	• Assessing what is most highly valued
• Assessing scientific knowledge	• Assessing rich, well-structured knowledge
• Assessing to learn what students do not know	• Assessing scientific understanding and reasoning
• Assessing what is easily measured	• Assessing to learn what students do understand
• Assessing only achievement	• Assessing achievement and opportunity to learn
• End-of-term assessments by teachers	• Students engaged in ongoing assessment of their work and that of others
• Development of external assessments by measurement experts alone	• Teachers involved in the development of external assessments

([5], p. 100)

Table 5. Excerpts from the *National Science Education Standards* of Standards for the Professional Development of Teachers of Science

• STANDARD A: The professional development of teachers of science requires learning science content through the perspectives and methods of inquiry …
• STANDARD B: Professional development of teachers of science requires integrating knowledge of science, learning, pedagogy, and students applying that understanding to science teaching…
• STANDARD C: The professional development of teachers of science enables them to build the knowledge, skills, and attitudes needed to engage in lifelong learning…
• STANDARD D: Pre-service and in-service professional development programs for teachers are coherent and integrated…
Excerpted and modified from [5]

The Growth of Advanced Placement and International Baccalaureate Programs

The number and diversity of high school students who are enrolled in Advanced Placement (AP) and International Baccalaureate (IB) programs, and the number of examinations being written, have been increasing almost exponentially over the past decade (e.g., [11]). One of the recommendations in *Rising Above the Gathering Storm* [12] is for the federal government to increase the number of teachers who are qualified to teach advanced courses, especially through the programs. A number of reports (e.g., [13,14]) have challenged various aspects of instruction, assessment, curriculum, and professional development in the science and mathematics components of AP based on emerging research on learning (e.g., [15,16,17,18,19]). In response, the College Board (which oversees the AP Programs), received $1.8 million from the National Science Foundation to restructure its AP courses in biology, chemistry, physics, and environmental science; that effort is currently underway[7].

While many students may be using high scores on AP and IB both for college credit and placement out of introductory courses, others are either electing or are being required to enroll in first-year STEM courses. Thus, it is incumbent on institutions of higher education to examine the effects of these high school programs on the structure, emphasis, learning goals, and actual learning of STEM subjects by their students. Evaluating the "value added" by introductory courses over and beyond the experiences that students acquired in high school is especially important.

---

7. For more information see http://www.nsf.gov/news/news_summ.jsp?cntn_id=106929. In the STEM disciplines,

American Competitiveness in an Increasing Global Economy

Concerns continue to be expressed virtually every day about the seeming declining competitiveness and skills of the U.S. labor force in the 21st century, and how K-12 and higher education in science, technology, engineering, and mathematics (STEM) must contribute to addressing these problems (e.g., [12,20,21,22,23]; for an alternate view see [24]). Such discussions inevitably lead to questions such as what students are learning in our colleges and universities (both two- and four-year institutions), whether higher education is imparting to students skills as well as knowledge, and the balance between liberal education and more directed knowledge and professional skills.

RELEVANCE OF THESE "GATHERING STORMS" TO INVESTIGATORS IN NSF'S *ASSESSMENT OF STUDENT ACHIEVEMENT IN UNDERGRADUATE EDUCATION* INITIATIVE

The original program announcement for the Assessment of Student Achievement (NSF 01-82) provides the following description for this initiative:

> *The Assessment of Student Achievement in Undergraduate Education (ASA) program supports the development and dissemination of assessment practices, materials (tools), and measures to guide efforts that improve the effectiveness of courses, curricula, programs of study, and academic institutions in promoting student learning in science, mathematics, engineering, and technology (SMET). Assessment tools and measures can inform educational materials development, teaching practices (pedagogy), use of educational technology, laboratory practices, field experiences, curricular decisions, co-curricular activities, advice offered by academic advisors, departmental decisions about major course requirements, and the organization of educational activities, both internally and externally (such as links with other academic sectors--high schools, other two- and four-year colleges and universities, and distance education opportunities-- and with employers).*

The research that is reported in this volume is supported through NSF's Assessment of Student Achievement in Undergraduate Education initiative. The work is innovative and, in many ways, unique because the studies have been developed and undertaken primarily by STEM faculty who are devoting at least part of their professional lives to understanding effective ways to encourage and catalyze student learning. Such knowledge is essential to finally understanding how young adults come to acquire knowledge, skills, and appreciation of these disciplines. The insights and perspectives that undergraduate faculty from the disciplines bring to bear on designing, implementing, and interpreting this research is groundbreaking and could open new horizons in understanding student learning and developing effective programs to foster that learning.

The 33 research papers in this volume focus on truly innovative ways to develop quality curriculum materials and concept inventories; and they assess student knowledge, learning and problem solving, and engagement in individual undergraduate STEM courses or programs. This research is clearly essential and welcomed.

However, in the evolving landscape of education, work at this level may not be sufficient. As undergraduate STEM education progresses and matures, we also will need to move beyond individual courses and programs as the primary units of analysis. We will need to ask questions about longer-term retention of knowledge, concepts, skills, and attitudes about STEM and how these translate into ways that students view science and how they apply them to the kinds of decisions that they make about their own health, their families, and the local

and global communities in which they will live. We will need to study in far greater depth whether particular approaches to teaching and learning increase students' abilities to transfer and apply knowledge, skills, and attitudes obtained in one course to other courses and to other disciplines. Given the increasing body of research that shows that the most important factor in improving student learning is an effective teacher, we will need to design more research that examines how undergraduate STEM courses and programs enable and empower students to become effective practitioners. And we need to determine whether new approaches to teaching and learning that are directed toward STEM majors are retained and utilized as some of those students become involved with undergraduate teaching and learning as graduate students, postdocs, and future faculty.

The tasks are daunting and vitally important. We are at the beginning of a new era of serious scholarship by STEM faculty in finding out how to impart their knowledge to students and measuring whether those new approaches are effective. It is my hope that this work will also make clearer the importance of teaching and learning for longer-term purposes and loftier ends.

ACKNOWLEDGMENTS

I thank Dr. Rolf Blank, Council of Chief State School Officers, for providing updated data about compliance with the No Child Left Behind Act.

REFERENCES

[1] Dwyer, C., Millett, C.M., and Payne, D.G., *A Culture of Evidence: Postsecondary Assessment and Learning Outcomes*, (Educational Testing Service, Princeton, N.J.), 2006. Available at: http://www.ets.org/Media/Resources_For/Policy_Makers/pdf/cultureofevidence.pdf

[2] McPherson, P., and Shulenburger, D., *Elements of Accountability for Public Universities and Colleges*, discussion draft, (National Association of State Universities and Land-Grant Colleges, Washington, D.C.), 2006. Available at: http://www.neasc.org/cihe/assessment/NASULGC_national_accountability_agreenent.pdf

[3] National Research Council, *Systems for State Science Assessment*, (National Academies Press, Washington, D.C.), 2005. Available at: http://www.nap.edu/catalog/11312.html

[4] American Association for the Advancement of Science (AAAS), *Benchmarks for Science Literacy,* (AAAS, Washington, D.C.), 1993. Available at: http://www.project2061.org/publications/bsl/online/bolintro.htm

[5] National Research Council (NRC), *National Science Education Standards*, (National Academies Press, Washington, D.C.), 1996. Available at: http://www.nap.edu/catalog/4962.html

[6] National Council of Teachers of Mathematics (NCTM), *Curriculum and Evaluation Standards for School Mathematics*, (NCTM, Reston, Va.), 1989.

[7] National Council of Teachers of Mathematics (NCTM), *Principles and Standards for Components of AP Based on Emerging Research on Learning School Mathematics*, (NCTM, Washington, D.C.), 2000. Available at: http://www.nctm.org/standards/

[8] International Technology Education Association (ITEA), *Standards for Technological Literacy,* (ITEA, Reston, Va.), 2000.

[9] Gross, P.R., Goodenough, U., Lerner, L.S., Haack, S., Schwartz, M., Schwartz, R., and Finn, Jr., C.E., *The State of State Science Standards 2005*, (Thomas B. Fordham Foundation, Washington, D.C.), 2005.
Available at: http://www.edexcellence.net/foundation/publication/publication.cfm?id=352

[10] Labov, J.B., "National and State Standards in Science and Their Potential Influence on Undergraduate Science Education," *CBE Life Sciences Education*, vol. 5, pp. 204-209, 2006.
Available at: http://www.lifescied.org/cgi/content/full/5/3/204?maxtoshow=&HITS=10&hits=10&RESULTFORMAT=1&author1=Labov&andorexacttitle=and&andorexacttitleabs=and&andorexactfulltext=and&searchid=1&FIRSTINDEX=0&sortspec=relevance&resourcetype=HWCIT

[11] College Board, *Advanced Placement Report to the Nation*, (College Board, New York), 2007.
Available at: http://apcentral.collegeboard.com/apc/public/repository/ap07_report_nation.pdf

[12] National Academy of Sciences, National Academy of Engineering, and Institute of Medicine (NAS, NAE, IOM), *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future*, (National Academies Press, Washington, D.C.), 2006.
Available at: http://books.nap.edu/catalog.php?record_id=11463

[13] National Research Council, *Learning and Understanding: Improving Advanced Study of Mathematics and Science in U.S. High Schools*, (National Academies Press, Washington, D.C.), 2002.
Available at: http://books.nap.edu/catalog.php?record_id=10129

[14] Sadler, P.M. and Tai, R. H., "Advanced Placement exam scores as a predictor of performance in introductory college biology, chemistry, and physics courses," *Science Educator*, vol. 16, no. 1, in press.

[15] National Research Council, *How People Learn: Brain, Mind, Experience, and School* (expanded ed.), (National Academies Press, Washington, D.C.), 2000.
Available at: http://books.nap.edu/catalog.php?record_id=9853

[16] National Research Council, *Knowing What Students Know: The Science and Design of Educational Assessment*, (National Academies Press, Washington, D.C.), 2001.
Available at: http://books.nap.edu/catalog.php?record_id=10019

[17] National Research Council, *Adding It Up: Helping Students Learn Mathematics*, (National Academies Press, Washington, D.C.), 2001.
Available at: http://books.nap.edu/catalog.php?record_id=9822

[18] National Research Council, *How Students Learn: History, Mathematics, and Science in the Classroom*, (National Academies Press, Washington, D.C.), 2005.
Available at http://books.nap.edu/catalog.php?record_id=10126

[19] National Research Council, *Taking Science to School: Learning and Teaching in Grades K-8*, (National Academies Press, Washington, D.C.), 2007.
Available at: http://books.nap.edu/catalog.php?record_id=11625

[20] Business Roundtable, *Tapping America's Potential: The Education for Innovation Initiative*, (Business Roundtable, Washington, D.C.), 2005.
Available at: http://64.203.97.43/pdf/20050803001TAPfinalnb.pdf

[21] National Academy of Engineering, *Educating the Engineer of 2020: Adapting Engineering Education to the New Century*, (National Academies Press, Washington, D.C.), 2005.
Available at: http://books.nap.edu/catalog.php?record_id=11338

[22] Project Kaleidoscope, *Transforming America's Scientific and Technological Infrastructure: Recommendations for Urgent Action*, A Report on Reports, Vol. II, (Project Kaleidoscope, Washington, D.C.), 2006.
Available at: http://pkal.org/documents/ReportOnReportsII.cfm

[23] Association of American Colleges and Universities (AACU), *College Learning for the New Global Century*, (AACU, Washington, D.C.), 2007.
Available at: http://aacu-secure.nisgroup.com/acb/stores/1/product1.cfm?SID=1&Product_ID=122

[24] Wadhwa, V., Gereffi, G., Rissing, B., and Ong, R., "Where the Engineers Are," *Issues in Science and Technology*, vol. 23, no. 3, pp. 73-84, 2007.
Available at: http://www.issues.org/23.3/wadhwa.html#

From the Grassroots: Provocative Insights & Compelling Research on the Learning of Undergraduates in STEM Fields

Jeanne Narum - *Project Kaleidoscope*

Closing Panel Moderator

Jeanne Narum is the Director of Project Kaleidoscope (PKAL), where her responsibilities are focused on building leadership at the institutional and national levels to ensure that American undergraduates have access to robust learning experiences in STEM fields. A major responsibility is coordinating the volunteer efforts of a national cadre of change agents, whose experiences shape PKAL institutes, seminars, workshops, and publications, illustrating best practices in the work of reform. Another responsibility is editing PKAL publications (electronic and print) on leadership in reform. PKAL's goal is to encourage the design and development of an intellectual, physical, and organizational infrastructure that supports strong learning in STEM fields. PKAL's Faculty for the 21st Century network is a significant dimension of PKAL. Narum is the project director for the grants from NSF and other funders that support the work of PKAL.

In response to the question, "What is PKAL?" Jeanne answers: "Our world is one in which science and technology have a profound impact on every aspect of life. PKAL is part of the growing national effort, using the energies and expertise of leaders within the undergraduate STEM community, to prepare coming generations for that world, for lives that are self-fulfilled, productive, and of service to society."

Narum was educated as a musician and her prior experience includes administrative positions at Augsburg College, where she was vice president for advancement; Dickinson College, where she served as director of development; and St. Olaf College, where she was director of government and foundation relations. From these positions she learned about the role of careful planning and creative thinking, both in shaping institutional futures and in securing the resources needed to realize that future, and developed a clear understanding of how strong undergraduate programs serve the interests of students and society. Narum has published widely on the subjects of external fundraising and grantsmanship.

All of us–citizens and senators and shopkeepers and scholars–need to review the principles of "critical thinking." In 1990, psychologists Carole Wade and Carol Tavris listed eight elements of critical thinking:

1. Ask questions; be willing to wonder.
2. Define your problem correctly.
3. Examine the evidence.
4. Analyze assumptions and biases.
5. Avoid emotional reasoning.
6. Don't oversimplify.
7. Consider other interpretations.
8. Tolerate uncertainty.

These words are not from a college or university catalogue, but from an article in the July 1, 2007 Weekend Commentary section of the *Washington Post* with the opening sentence, "Here's who we need in Washington: Socrates. The Greek fella. We need him not because of what he knew, but because of what he knew he didn't know, which was pretty much everything." Article author Joel Achenbach quotes Loyal Rue, professor of science and religion at Luther College in Iowa, describing Socrates like this: "He would say things like, 'How do you know that? What's the evidence for that? What do you really mean when you say that?'"

This publication summarizes lessons learned and presents documented research findings from recent work by the growing community of pioneers in assessment of student learning within the undergraduate STEM learning environment. Its timeliness cannot be emphasized strongly enough. The stories, reflections, and data in these pages are valuable on many counts. They will inform the increasing public discussions about what students should know and be able to do during and beyond their undergraduate years. They confirm other public calls to action, such as:

> *People are not born with inherent innovation skills, but they can learn them. They can acquire the social skills to work in diverse, multidisciplinary teams, and learn adaptability and leadership. They can develop communication skills to describe their innovation. They can learn to be comfortable with ambiguity, to recognize new patterns within disparate data, and to be inquisitive and analytical. They can learn to translate challenges in opportunities and understand how to complete solutions from a range of resources* [1].

The experiences of this cadre of academic leaders spotlight the emerging culture of evidence within the broader undergraduate STEM community. *What works?, why?*, and *how do we know?* are questions that always drive faculty concerned about the quality and character of the learning of students in their classrooms and labs. In years past, addressing these questions was mostly the province of the single faculty member; he or she had only his or her personal experience from which to extract answers that could potentially inform how they shaped the learning of their students. That said, these were precisely the questions that sparked the work of the pedagogical pioneers who, more than two decades ago, began experimenting with approaches that are now becoming widely practiced: collaborative problem-based learning, studio classrooms, Just-in-Time Teaching, etc.

One interpretation I've made of the tapestry of educational innovation and reform in undergraduate STEM that has emerged since 1985 (remembering Socrates' caution about interpretations) is the critical trifecta of:

1) The research on learning that was captured in the seminal publication, *How People Learn,* from the National Research Council. [2]

2) The explorations, successes, and failures of today's generations of pedagogical pioneers within disciplinary communities in STEM fields.

3) The shift on campuses across the country on the learner and learning from attention to the teacher and teaching, and the attention at the national level to student learning in STEM fields.

The value of the work described in these pages is that it illustrates how connected these threads are. Research on how people learn outlines the theoretical basis for shaping robust learning environments. The many and diverse contemporary pedagogical explorations and innovations illustrate how to translate those theories into practice. The

very real interest in and concern about student learning by leaders on campuses from all sectors of higher education reveal a community eager to learn about and to adapt approaches that help them address—in ways that serve their institutional circumstances and context—questions about the how and why of student learning.

Even a cursory review of the ASA abstracts confirms the potential impact of these projects on the larger community of leaders in undergraduate STEM. As the following examples suggest, they are designed:

– for dissemination and adaptation in a variety of disciplines and institutional settings

- "... the Science Perceptions Survey will be disseminated in a web-based format for student self-evaluation...." (Angelica Stacy, University of California, Berkeley: *"ChemQuery: An Assessment System for Mapping Student Progress in Learning Chemistry")*
- "Diagnostic assessment is the foundation for instructional improvement. It alerts instructors to student difficulties in learning particular concepts....we are piloting a procedure for developing diagnostic questions that can be used in any field." (Joyce M. Parker, Michigan State University: "*Diagnostic Question Clusters...*")

– around explicit awareness of research on learning

- "... students are using the instrument to self-assess their learning after each assignment, to clarify expectations for their performance, particularly with regard to conceptual understanding and problem solving, and to identify strategies for improving their achievement...." (David Hanson, Stony Brook University: "*Real-time Multi-dimensional Assessment of Student Learning*")
- "... the ultimate products will be web-based pedagogical assessment materials for introductory STEM courses that will be disseminated nationally." (John Orr, University of Nebraska–Lincoln, "*Assessment of Student Achievement in Undergraduate Education*")

– for collective action, within departments, interdisciplinary communities, and/or at the institutional level

- "... planning a major curricular revision of the biology major's program, faculty in the department of biological sciences determined student learning outcomes for the entire curriculum and for the four new core courses, which were to incorporate active learning and inquiry-based activities in both lecture and lab sections." (Joyce Ono, CSU Fullerton: "*Development of Faculty Collaboratives to Assess Achievement of Student Learning Outcomes in Critical Thinking...*")

– as mechanisms for faculty development

- "... this project is studying...the capacity [of faculty] to use research on learning and assessment practices in ways that result in changes in their approaches and attitudes toward teaching...." (Richard Iuli, University of Rochester: "*Assessing Student Learning and Evaluating Faculty Capacity Development...*")

– address both learning issues of concern for those shaping programs for all students and programs serving majors in STEM fields

- "... the Science and Math Value Inventory provides a means to determine if their general education math and science programs have truly made a difference in the lives of their students." (Donald G. Deeds et al., Drury University: "*Valuing Literacy: The Science and Mathematics Value Inventory*")

Further examples of these 'threads' are presented in the reports from individual projects that follow.

At the end of the first meeting of ASA grantees in late 2006, participants were invited to think about lessons they had learned that would be of value to colleagues and peers interested in exploring, adapting, and extending their work. Their advice included:

• *There is no need to reinvent the wheel.* Many well-established theories, methods, and analytical techniques exist, based on work in cognitive science and psychology, and they can (must) be used to inform and understand assessment, including how to develop and use instruments, and shape and monitor learning outcomes.

• *You are not alone.* There are a growing number of conversations about how difficult it is to teach certain concepts; there are many faculty struggling with student learning. Although 'assessment' may be scary to some, a good start is to be actively reflective about the teaching and learning that happens in your classroom.

In addition, ASA grantees thought STEM faculty should know that:

1) Students know and understand less when they emerge from courses than most faculty think they do.

2) What we teach despite our best efforts, is not what students learn or how they learn.

3) Student achievement can be increased with effective assessment.

4) You can teach better and enjoy it more if your students are demonstrably learning better.

Increasingly, our work in Project Kaleidoscope is based on the realization that systemic and sustainable transformation of the undergraduate STEM learning environment will only happen if and when there is a collective vision at the campus level about what students should know and be able to do upon graduation, and a parallel commitment to collaborate in realizing that vision.

REFERENCES

[1] Council on Competitiveness, *National Innovation Initiative Summit Report: Thriving in a World of Challenge and Change*, 2005.

[2] National Research Council, *How People Learn: Brain, Mind, Experience, and School* (expanded ed.), (National Academies Press, Washington, D.C.), 2000.
Available at: http://books.nap.edu/catalog.php?record_id=9853

Assessment and Accountability from a Scientific Point of View: What is the Evidence?

David Oxtoby - *Pomona College*

Plenary Speaker

David Oxtoby is president of Pomona College in Claremont, California, and Professor of Chemistry. He received his bachelor's of arts from Harvard University and received his Ph.D. from the University of California, Berkeley.

As President of Pomona College, Dr. Oxtoby is interested in liberal arts colleges and the future of education in the 21st century. He has particular expertise in science education and active learning approaches, and in the international dimensions of education. His work in environmental chemistry has focused on atmospheric chemistry and the science of global climate change. His research interests in chemistry involve thermodynamics and statistical mechanics, including the study of phase transitions and nucleation, crystal growth, and amphiphile fluids and self-organization. As a research chemist, he is author or co-author of more than 165 scientific articles on such subjects as light scattering, chemical reaction dynamics, phase transitions, and liquids. He has been invited as a guest lecturer at conferences and institutions around the globe.

Dr. Oxtoby also co-authored two nationally popular textbooks in chemistry, *Principles of Modern Chemistry*, with H. Pat Gillis, now in its fifth edition and in use at institutions across the country; and *Chemistry: Science of Change*, with Wade Freeman and Toby Block, which is now in its fourth edition and has been translated into Italian and Korean. He has received fellowships from the Guggenheim, von Humboldt, Dreyfus, Sloan, Danforth and National Science Foundations.

He is a fellow of the American Physical Society and a member of the American Chemical Society, the Materials Research Society, the Royal Society of Chemistry (Great Britain), and the American Association for the Advancement of Science.

The alarm bells are ringing for higher education these days. In the past year alone, a clamor calling for major changes in the funding, management, and even the fundamental structure of colleges and universities has risen to record levels. Even the general public has taken notice, although the range of issues for them goes all the way from the pressures on top high school students from SATs and Early Decision programs all the way to the challenges our lowest income students face simply to find places in local colleges.

It is noteworthy that we have gone from a state of complacency about postsecondary education (and a focus on the ills of K-12 education) to one in which, for both good and bad reasons, we are recognizing that preschool through professional school education is one interlocking system. With this said, we still must not blindly apply common solutions across the boundaries, or ignore the special issues in each sector.

In some ways, this new attention arose first, and appropriately, from within the academy itself, with the release last fall of the National Academy of Sciences report *Rising above the Gathering Storm* [1]. Stressing the

importance of competitiveness in a flat world economy, this report argued that the United States is falling behind in many measures in preparing its citizens for the world of the 21st century, and that significant investment in the science, technology, engineering, and mathematics (STEM) sectors was essential to regaining our position of leadership. The American Competitiveness Initiative announced by the White House in January was a direct outgrowth of this call.

In parallel, a less-noticed discussion was taking place in Congress in the course of the Higher Education Reauthorization process, led by John Boehner, now the majority leader of the House. His committee hearings were hammering higher education for our costs, lack of efficiency, and failure to monitor outcomes in terms of student success. In principle, all these are valid criticisms of our sector, and ones worth examining in a thoughtful fashion, but those of us in higher education could not help noticing the heavy funding to Mr. Boehner and others from the for-profit sector of higher education and the prominent way in which the agenda of proprietary schools was entering the discussion. If the primary issue for higher education is efficiency and cost, there is no question that an institution using primarily adjuncts and online instruction can deliver a degree for a lower cost than a traditional college or university. But that requires us to leave quality out of the equation.

A third strand focusing on higher education issues is what I would refer to as the upper middle class anxiety about getting into the best colleges. Twenty or thirty years ago, outstanding students went to a large range of colleges, from the nearby state university or college to more distant institutions. These days, with ratings of colleges so prominent in the month of August (mostly aimed at selling magazines) there is a great deal of hyper-ventilation about the challenges even top students have in getting into the "best" colleges, in part because those lists of best colleges tend now to be more national than regional. Many options that offer excellent educations to college students can be forgotten in the admissions frenzy.

The culmination of all these inputs was the establishment early this year by the Secretary of Education, Margaret Spellings, of the Commission on the Future of Higher Education. The Spellings Commission, which included representatives from all sectors of education and from the general public, held hearings, drafted a report, and published it in September 2006 as *A Test of Leadership: Charting the Future of U.S. Higher Education* [2]. I would like to make a few comments about this report, including both its value and its limitations, and then turn to the core of my talk today, which is the role that the STEM community in higher education (represented by all of you) can play in furthering the dialog about the role of liberal education in the 21st Century, and ways in which you can be leaders in using assessment to improve the quality of higher education. Before proceeding, I would like to acknowledge the central role of conversations with fellow board members of the Association of American Colleges and Universities (AAC&U), as well as written materials they have produced. I consider the AAC&U to be leaders in rethinking the role of higher education today, and I frankly wish the Commission had paid more attention to their recent work.

The Spellings Commission report begins by stating that this is a critical juncture for higher education in this country. While acknowledging that "American higher education has been the envy of the world for years," the report goes on to point out that by some measures we are falling behind other countries and that this is not a time for complacency. The variety in our higher education system, ranging from community colleges, to small liberal arts institutions, to large research universities, provides many entry points, allowing the most motivated students to find a place that is right for them and to succeed, or even to have a second or a third chance after failing on the first

attempt. But the concern is with the number of students who are not achieving access to the system, or who enter but fail to make progress toward a degree. The report states that recent data places us only twelfth among major industrialized countries in higher education attainment.

In the past, there has been a tendency on the part of higher education to point fingers at the K-12 system and blame our problems on them. Give us better prepared students out of high school, we would say, and everything will be fine. Or we focus on issues of culture and society: too much television and too little reading, too much emphasis on having fun. I think that in a healthy way we are beginning to recognize the connections across the levels of education and to acknowledge that a good share of the fault is our own. This is, in my view, a positive development. Even students coming from excellent high schools are not succeeding at the level we would hope in our colleges and universities. We must ask ourselves: Why?

Much of the Spellings Commission report is right on target. In a discussion of access, they point out that even high-achieving students in the lowest income quartile are going to college in far lower percentages than those from high-income groups; the linkage between income and access is deeply disturbing if we truly think of our higher education system as a meritocracy. They correctly deplore the low rates of graduation from our colleges and universities, saying that we have failed our students if we have not made it possible to succeed once they enter our gates. They call for a significant increase in need-based financial aid, with the Pell Grants as the cornerstone of this program—a welcome statement, but one unfortunately short on details. They applaud innovation and ask for more, in areas ranging from new methods of teaching to appropriate uses of technology in our institutions.

To me, one of the most shocking results described in the report is the low levels of adult literacy revealed among not high school dropouts but among college graduates. Now, I cannot say that I have looked at the National Assessment of Adult Literacy, so I don't know what types of questions it asks, but this test showed that in 2003 the percentage of adult college graduates who are proficient in prose literacy (the ability to understand narrative texts such as newspaper articles) stood at 31 percent, the percentage proficient in document literacy (the ability to understand practical information such as instructions for taking medicine) was only 25 percent, and the percentage proficient in quantitative literacy was 31 percent. The first two percentages had decreased significantly since 1992. Of course, many of the individuals in the sample are far from their college years, but these low numbers are still disturbing.

Another Commission recommendation is of direct relevance to the U.S. STEM community. They call for an increased federal investment in education and research in STEM fields, as well as a relaxing of visa requirements so that international students would no longer have to provide proof that they will not stay in this country after receiving their degree. Instead, the Spellings Commission calls for expedited visa processing for such students to keep their talents in this country and help build our intellectual capital.

Other proposals from the Commission are much more problematic, in my opinion. The National Association of Independent Colleges and Universities (NAICU) has focused on one in particular, the so-called unit record system. According to this, a large-scale federal database would be constructed that would contain a record of every course taken by every student in the country. That would allow one to study how students follow paths from one institution to another, whereas at present the government only tracks graduation rates from single institutions. What is wrong with this idea? Invasion of privacy is a major issue, since every time a student drops or adds a course it would need to

be reported to the federal government. Can you imagine the use of such a database to see what students had taken courses on Islam? And what kind of bureaucracy will be needed to set up and run a classification scheme for courses? Although the report states that at the federal level the system "would not include individually identifiable information such as student names or Social Security numbers," how would this be possible in practice? How does a college enter information without an identifier that will follow the student throughout his or her life? If our goal is to understand better the pathways followed by typical students, why not just do a large-scale survey, rather than assemble this information on every one of the millions of students in the country? These questions have never been answered in a satisfactory fashion by the members of the Commission, nor by the Secretary of Education herself.

I am also concerned about the naïve discussion of the costs of higher education presented in the report. As someone who is faced on a daily basis with choices on budgets, I can assure you that neither Pomona College nor other institutions increase tuition simply because the market will allow it. Faculty members' salaries have not kept up with those for other skilled professionals in this country; the provision of need-based financial aid has soared and impacts our tuition income; the costs of technology in education, or of offering first-class science facilities and equipment to excite our students about these fields, do not track with the consumer price index. Of course there are excesses in any free-market system, such as the building of expensive new athletics centers or subsidies to large-scale competitive football programs. But to suggest, as the Commission does, that "the growth in college tuition [should] not exceed the growth of median family income over a five-year period" is naïve. Even more absurd is their suggestion that private contributions from donors "insulate . . . colleges and universities from the consequences of their own spending decisions." As someone who spends a great deal of time in fundraising, I can say that our institutions would be far worse off if it were not for the generosity of donors. On the other hand, I agree that prestige (such as *U.S. News & World Report* rankings) is often measured by how much we spend, not by what we deliver for the money, and this can have negative impacts. In Claremont, we are asked by *U.S. News* about how much money we spend on the library, not on the extent of library resources made available to our students.

To me, however, the most disappointing aspect of the Spellings report is what it does not say. The single mention of the word "faculty" comes on p. 23 in the statement that "faculty must be at the forefront of defining educational objectives for students and developing meaningful, evidence-based measures of their progress toward these goals." What happened to faculty actually teaching students? Nowhere in the report is there a hint of the goal of intellectually challenging our nation's students through the encounter with new ideas, of bringing them into the life of the mind through their time on our campuses, of opening their horizons through the encounter with other students from a range of different backgrounds. Does all this count for anything?

The Spellings report calls for efficiency at our colleges and universities in order to lower costs. I am all for reductions in waste, but there is much about the effort of higher education that can never be, and should not be, efficient. Instruction costs are clearly lower when part-time adjunct faculty members are used instead of regular tenure-track faculty. Does this mean we should move in that direction? It would be more efficient for us to eliminate a small department such as Latin and replace those faculty members with others to teach high-enrollment areas such as economics, but is that the right thing for a college or university to do? Of course those of us at liberal arts colleges believe strongly in the value of small, discussion-based classes, but what is less efficient than asking a faculty member being paid $80,000 per year to lead a discussion class of 15 students? Why not put that faculty member in front of a 500-person lecture or, better yet, broadcast her words over the internet to a population of distance education students?

Perhaps most seriously of all, the report ignores the possibility of aiming for coherence in the education we offer. It points out, correctly, that many students do not stay at a single institution for their full college education, instead transferring from one college to another in that time. But rather than raising flags of concern about whether students who move from one institution to another are getting a balanced, broad, and coherent education, the Commission instead puts all its emphasis on smoothing these connections and making it as easy as possible to graduate. Courses are thought of not as parts of a whole, but rather as numerical credits that should be put together until there are enough to receive a degree. Too many of our own students think too much this way, regarding the bachelor's degree as a credential to be won by accumulating enough credits and checking the right boxes. We do not need national commissions to encourage this thinking.

I even worry about the implications of this approach for our high schools. Although I welcome the offering of challenging courses for students in high school, the advanced placement programs are rapidly becoming the latest way in which schools are "teaching to the test" rather than using creativity to excite and challenge students. Too much of the high school curriculum is turning into a pale imitation of college courses instead of providing the solid foundation that students need to build on in the future. I would rather that a chemistry course in high school present a thorough treatment of periodic properties of the chemical elements than that students learn about s-, p-, and d-orbitals. Yet in the name of being "advanced" students are thrown theoretical interpretations when they have no facts to connect them with, nor the math skills to understand the theory. I worry that in the future, students will try to get their broadening requirements (whether they are math, science, or literature) "out of the way" in high school so they can concentrate immediately upon entering college on more specialized material for their majors. In my view, we do not want to move toward a European type of system in which higher education is highly specialized and there is little opportunity for cross-fertilization between fields, and little flexibility for students to move from one area to another.

We read articles about the emphasis being placed in countries such as China and India on educating scientists and engineers. New universities are being established at a great pace to produce students with degrees in critical fields such as computer science or engineering. It is this fear of falling behind that has helped to motivate our own nascent competitive initiative, and I welcome the attention to the problem. However, we should recognize that much of the teaching in these new mega-universities abroad is done in rote fashion in classes of hundreds of students. Earlier this week the *New York Times* reported that only one in four students graduating with engineering degrees in India is employable. Surely we do not want to rush headlong in this direction, processing large numbers of our own students efficiently through large classes to obtain degrees of little value even more rapidly.

Princeton economist Alan Blinder has written eloquently about the "third industrial revolution" that he sees coming as the nature of the economy changes through international competition. The First Industrial Revolution was the move of workers from the farm (producing food) to the factory (producing industrial goods). The Second Revolution was the move since the World War II from manufacturing to services. Blinder quotes data suggesting that by 2004 only about one-sixth of America's non-agricultural jobs were in goods-producing industries, and five-sixths in the service sector. The next, and third, revolution, which is underway right now, stems from the new Information Age and involves international competition and the resulting flow of jobs that can be done remotely from high-cost regions (such as the U.S.) to low-cost regions such as China and India.

What Blinder points out is that there are certain jobs that need to be done locally and others that can be done remotely. In fact, these do not divide into "high-skilled" and "low-skilled" jobs. Jobs that are unlikely to be off-shored include those of carpenters, taxi drivers, nurses, and (we hope) college professors, while jobs that are easily done remotely and thus subject to international competitive pressures include those of typists, reservations clerks, computer programmers, and tax preparers. The important thing is to prepare future job-holders for jobs that require personal contact, rather than even sophisticated jobs that can be done remotely.

Blinder worries that we are retooling our higher education system to respond to the previous industrial revolution, not the third industrial revolution that is upon us now. Getting more students to graduate from college is good, and granting more degrees in engineering or computer science may be a useful goal, but only if there is real value added, and our engineers are able to do things that an engineer overseas cannot. Following the path of the Spellings Commission toward having more students complete a college education that has specific, routinely testable goals will prepare a nation of workers whose jobs can all be done more easily overseas. If our mark of success is high scores on standardized tests such as the Collegiate Learning Assessment, India and China can probably teach more effectively to that same test and we will have gained no real advantage.

Blinder talks directly about the issue of efficiency. Efficiency in manufacturing has increased significantly, just as it earlier increased in the field of agriculture. Many fewer people are required now to make a car or produce a bushel of wheat. The same efficiency is moving into what he refers to as the "impersonal services" area. An example is telecommunications, where many fewer people are required to operate a much more versatile and extensive system than in the past. In relative terms, personal services are not subject to the same improvements in efficiency. He gives two simple examples: the number of person-hours to play a Mozart string quartet or to drive students to school on a bus has not significantly changed in several generations. Teaching is a "personal service" occupation as well, so calls to improve efficiency and productivity have real limits; that is one important reason why college costs (just like medical care) increase at a greater rate than the average represented by the Consumer Price Index.

Our education must instead be focused on preparing students for roles in a competitive environment in which qualities that cannot be off shored are developed in our students. These include creativity, the ability to work effectively with others one-on-one and in groups, and the ability to address complex issues from multiple perspectives. It is these skills that our educational system needs to be focusing on, and I regret to say they are not always subject to increased efficiency, nor to simple testing on a national basis. Instead, they require a recommitment in this country to the core principles of liberal education.

A liberal education has nothing to do with liberal or conservative political beliefs. Rather, "liberal" is used in the sense of "freeing"; the Latin term "artes liberales" can be translated as "the skills of freedom." The types of skills we need to be teaching in our colleges and universities include practical skills such as critical inquiry, written and oral communication, quantitative literacy, and problem solving. We need to make sure our students develop a sense of personal and social responsibility, including civic engagement locally and globally, intercultural knowledge, and ethical reasoning. Finally, students need to learn how to address big questions from multiple points of view, and to collaborate with others to synthesize different perspectives. An education such as this will encourage creativity and enable students to move across areas to new problems and unfamiliar challenges. A liberal education, combined with deep knowledge and understanding of some particular field, is the best preparation for the world-scale competition we all face in the 21st Century, and the best way to educate our citizens to play active roles in a democratic society.

Last June I attended a meeting in Strasbourg sponsored by the Council of Europe, entitled "Higher Education and Democratic Culture: Citizenship, Human Rights, and Sustainability." A particularly interesting dialog took place there between presidents of liberal arts colleges in the United States and rectors and ministers of education from the new democracies in Eastern Europe. Those new countries, many of them carved out of nations torn apart by civil wars or standing on their own feet after years of Soviet domination, are eager to understand how higher education can help to educate citizens for democracy, a task that we sometimes take for granted in this country. There is a place for specialized training in technical fields, but it must be accompanied by a liberal education, helping all students to think about issues that cross boundaries, and about the civic responsibilities that they face as citizens. This country can and should be a leader in spreading this message in the world.

A liberal education cannot be the province of only those elites who can afford to send their children to the top small residential liberal arts colleges in the country. A two-tier system in which a few gain the benefits of close attention in small classes while the majority end up in those "efficient" online and for-profit institutions so enthusiastically endorsed by the Commission would be a step back for this country. Someone has suggested that we should take a look at the college choices made by Commission members for their own children; the Secretary of Education herself sends her child to an elite (and expensive) liberal arts college in the South.

It is thus imperative that all institutions, from liberal arts colleges to research universities to comprehensive regional institutions to two-year colleges, adopt the core guiding principles of liberal education, even though the implementation may differ considerably from one institution to another. This is the call that has recently been made by the Association of American Colleges and Universities (AAC&U) through its new program, Project LEAP (Liberal Education and America's Promise). One of their core principles is that we should "Aim high – and make excellence inclusive." There should not be one kind of broad education suitable for the elites and a narrow vocational training for everyone else.

The AAC&U points out, in a draft statement, that in 1947 the Truman Commission on Higher Education connected liberal education with the broad education in the first two years of college, what we might call general education. A distinction was thus made between liberal education and vocational education. In today's world, such a distinction is to be avoided; the broad issues raised in a liberal education need to come up throughout our curricula, including in culminating interdisciplinary projects that build on both specialized and general education.

What is the role of scientists and engineers in this conversation? It is critical that we not think of liberal education as something that "others" on campus lead. The STEM fields need to be seen not as specialized areas to be "complemented" by general education, but rather as areas right at the heart of a liberal education for all our students. Too often, in curricular discussions, each side is fighting only for its own turf, so the STEM community sometimes works for meaningful requirements in mathematics and science but leaves to the rest of the faculty the broader conversation about what we mean by civic engagement or moral reasoning. This is very shortsighted, as critical issues such as environmental impact on surrounding communities and the ethics of modern genetic research inevitably bring science together with the social sciences and humanities. Educating creative students who will be leaders in society requires all our best efforts, and scientists need to reach out across campus to their colleagues in these discussions.

It is always puzzling to me that, as a scientist who is president of a liberal arts college, I encounter members of the general public who think that such colleges are good places for future humanists, but not for future leaders in science and technology. We ourselves all know the type of science and math education our campuses can provide, and the role of engaged students and early involvement in student research to excite students about science and keep them in the field. We all need to communicate the excitement of our work more effectively outside the walls of our colleges, to emphasize that scientists and engineers are not narrowly trained specialists but individuals who have thought deeply about the major issues that face our world.

Let me turn now in the closing minutes specifically to the subject of assessment, the subject of this conference. What role can we as scientists play, and how can we help to move forward the national agenda on improving the accountability of our higher education institutions?

Let me give a few words of personal background to set the stage. As you know, I am a chemist by background, with a passion for teaching as well as for research. (I am able to continue teaching a small amount as president, but regretfully have given up my research program.) My first interest in issues of assessment stemmed from my role as co-author of two general chemistry textbooks, first published in the early 1980s and now in their fifth and fourth editions. In the course of revising these texts, the numerous comments from reviewers (both official and unofficial) on what worked and what did not work in their own classrooms helped to shape changes as each book was revised for its new edition. That was an extraordinarily useful form of assessment of our work.

In the early 1990s, I became involved in the ChemLinks coalition of colleges and universities, developing new materials for teaching chemistry with the support of a National Science Foundation grant. My own particular interest was in ways of bringing environmental chemistry into first-year chemistry, and I reshaped my own teaching not just to use environmental themes as illustrations of chemical principles, but rather to turn the course around: starting with questions posed by environmental issues and developing methods of chemistry to answer those questions. A major value of that consortium was the presence of liberal arts colleges and two research universities (including my own, the University of Chicago) from the Midwest, together with state universities and two-year colleges from California. The conversations across institutional type were extremely valuable, and it was exciting to see materials developed on a small scale at Grinnell College move to large-scale implementation at Berkeley. A key aspect of that project was the quantitative as well as qualitative assessment of outcomes.

Later in the 1990s I became involved in Project Kaleidoscope, a wonderful organization familiar to most of you that focuses on "What Works?" in STEM education. This network of change agents again showed me the value of independent assessment of our new ideas. It is not enough simply to say that something new is better; instead, it is critical to have evidence for the truth of such statements.

I have moved from assessment of individual efforts to collaborative projects and now, as president, to institutional assessment. I am currently chairing two different accreditation teams for institutions of higher education: one is the Graduate Theological Union in Berkeley (I'm not sure what my exact qualifications are in the area of theology) and the other is Bowdoin College in Maine. This has let me see at first hand two different sets of accreditation standards (California and New England) and take part in two different processes. Each is, however, focused on the central question of: "What is the evidence for student learning at your institution?"

As president, I go around the country making speeches to alumni and prospective students about the excellence of a Pomona College education. I am convinced of this from my own observations of students and faculty in our classrooms, but I have to say that I sometimes worry about the absence of quantitative or qualitative evidence for the statements that I am making. Could it just be that we have such bright students that whatever we do to them they would come out well? How can we demonstrate added value, and how can we learn from assessment to do even better? Pomona is itself entering its own reaccreditation process over the next several years, and I should say that I am the most active proponent of what we can learn from the process. My job is to inspire the rest of the campus to join with me in this effort. Most of our faculty members are doing what they are because they came to us with a passion for teaching, not a passion for assessment. Yet teaching and assessment need to be seen as more of a whole, not as separate efforts.

To me, the Spellings Commission approach to assessment is one of the most disappointing aspects of the report. They focus heavily on national tests, which have little bearing on the types of high-level skills we seek to develop in our students. For example, they call for the National Assessment of Adult Literacy to be given every five years, instead of every ten years, and to include samples of graduating students at two- and four-year colleges and universities. This is a remedial level test, surely one which we should worry if our graduates fail, but not a standard to be aimed for in higher education. They call for higher education institutions to use instruments such as the Collegiate Learning Assessment (CLA), "which measures the growth of student learning taking place in colleges." This is true, but only in a very limited sense. Scores on the CLA are closely correlated with SAT performance for entering students. Do we really want to spend the four years of college preparing students for yet another SAT-like test to evaluate what they have "learned" in college? Surely the high school years already focus too much on this very limited aspect of student learning.

The Commission report speaks quite positively about another type of assessment represented by the National Survey of Student Engagement (NSSE). This very useful survey asks students about how much time they spend on activities such as meeting with professors or engaging in other educational activities outside of class. Institutions that give this survey are able to compare their own institution anonymously with a group of comparable institutions and make changes to improve their practices. To me, the voluntary aspect of this assessment is its value. The Spellings Commission, however, suggests that such results be made public so that outside stakeholders can use the data to evaluate institutions. This worries me very much. I can imagine this turning into yet another "test" that institutions will seek to teach to in artificial fashions, where public relations will be as critical as actual practice in getting students to respond to this in the "optimal" fashion.

The central problem with the Spellings Commission recommendations is that they are too much of a "one-size-fits-all" approach. Although they do acknowledge that current instruments such as the CLA can be improved, it is not clear to me that there will ever be an appropriate test for the type of liberal education that I have described earlier in this talk, one that educates students to be creative and collaborative citizens of a democratic society. Neither will differences in preparation for a Latin major, a mathematics major, or an economics major ever be reflected in performance on any single standardized test, or barrage of tests. I think it is hard to argue that standardized testing has improved our K-12 schools, and the consequences for higher education (where even more advanced skills are developed) would be even worse.

Currently, we have an accreditation system that features, indeed thrives on, diversity. Although all approaches are focused on outcomes assessment and on student learning, having different accrediting agencies in different parts of the country fosters a spirit of creativeness and innovation that reflects our higher education system as well. I have noted the differences in approach (toward the same general goal) in New England and in California, and feel that this is healthy. It is incompatible, though, with the Spellings Commission's goal that institutional performance should be comparable across the country and that results of accreditation be fully public. I worry that these proposals are the first step toward the setting of overall national standards that will reduce the creative vitality of American higher education. Our strength is in our diversity and in the free-market system of higher education that lets many approaches flourish, not in a centralized system as is found in many countries of the world. A recent statement from the American Council on Education (ACE) quotes the British newsweekly *The Economist* as noting, "America's system of higher education is the best in the world. That is because there is no system."

Although I have been critical of the Spellings approach to assessment, in no way do I want to suggest that assessment is not important. Rather, I believe that assessment should be embedded in everything that we do, in order to see what works and improve student learning. Making comparisons across institutions is less important, however, than developing methods of assessment that lead to improvements in each institution; best practices can then be shared, but not with the goal of "rating" particular colleges and universities. Given the higher order goals of liberal education such as ability to integrate concepts or to apply principles of ethics, the corresponding methods of assessment will not be susceptible to simple quantitative analysis.

What role can scientists and engineers play in this process? By our very nature, we tend to be skeptical about assumptions. When broad statements are made we naturally ask: "What is the evidence?" The same questioning attitude we bring to our research should apply also to our teaching. Instead of just saying, "I'll teach it this way because I always have," or "My new approach is better than that used in this book," we should always be asking, "How do I know that this change is enhancing student learning?" This is a scientific question, and it is incumbent upon us to develop assessment methods to answer it.

But this does not mean simply using a few national tests to rate our colleges and universities, a la Spellings. Rather, it means that as scientists we are constantly doing experiments and testing hypotheses, and those experiments should be as different and as innovative as we can make them. The talks presented at this conference provide some examples of the range of experiments being carried out, and we should celebrate that diversity. As scientists, we should also be leading the way on our home campuses to show that evidence-based assessment of student learning is possible, encouraging our colleagues in the humanities and social sciences to follow us down this path.

At the same time, it is important to find effective ways to share and disseminate best practices in teaching and in assessment. We have long ago passed the era in which teaching innovations funded by the National Science Foundation were limited to the single college where they were developed. One method of dissemination is conferences such as this. But in a modern age of information, many other methods are possible as well.

I recently heard a talk by Uri Treisman, professor of mathematics at the University of Texas at Austin. He described new approaches to teaching mathematics, both at the high school and elementary calculus (college)

level. In Texas, and now in some other parts of the country, online networks are being set up in which teachers share their approaches to teaching and the problems that they have developed to test knowledge of concepts. Other teachers use these problems and rate their usefulness. Of course there is not one size that fits all; different approaches work better for different teachers. But all that is taken into account in the computer software that backs up the exchange. You all know how amazon.com works, don't you? Once you order a book, the program comes back with a suggestion of one or more related books that you might like as well. The same is true with this online exchange of teaching and problem suggestions, and it is having a measurable effect on student performance in the school systems where it has been employed. It is this type of "national" assessment that we need to work toward, one that is not imposed from the top but that organically integrates the best ideas from a large network of creative individuals.

My challenge to all of you is to be scientists in your teaching as well as in your research, to pose hypotheses and test them imaginatively, and to be leaders on your campuses in bringing evidence of student learning to the forefront of the educational process. We have major challenges in front of us, and there is much exciting work to be done.

REFERENCE

[1] National Academy of Sciences, National Academy of Engineering, and Institute of Medicine (NAS, NAE, IOM), *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future*, (National Academies Press, Washington, D.C.), 2006.
Available at: http://books.nap.edu/catalog.php?record_id=11463

[2] U.S. Department of Education, *A Test of Leadership: Charting the Future of U.S. Higher Education*, (Washington, D.C.), 2006.

Curriculum Improvement in Practice-based Biology Programs Using Student E-Portfolios: A Progress Report

Kostia Bergman*, Veronica L. Porter, Laurie Poklop, Susan Aman, Sara Noyes, and Judith F. Woolfson
Northeastern University

PROJECT OVERVIEW

This project began with the ambitious goal of assessing student learning in our biology/biochemistry undergraduate curriculum that incorporates traditional academic coursework; experiential education, including but not limited to experiences in Northeastern University's flagship Cooperative Education Program (Co-op); and the informal education that comes with student life. Our assessment method for this project includes expert review of student work according to well-defined rubrics with performance-based standards of measurable progress toward proficiency in desired learning goals. To facilitate the assessment process, we developed a web-based portfolio system that allows students to archive their work and present it in different views for different audiences, and a web-based performance evaluation process for use by employers of Northeastern students.

IMPLEMENTATION OF ASSESSMENT TOOLS

Rubric Development

For students to assess their own work and for reviewers to provide independent assessment, with the long range goal of curriculum improvement, we developed a scoring rubric that clearly identifies learning objectives and criteria as well as measurable performance-based levels of execution. The rubric development team included three employers, two academic faculty members, a co-op faculty member, an instructional designer, a graduate from the Northeastern biology program, and a graduate research assistant. Two of the employers were scientists working in the biotechnology industry, and the third was a human resource management specialist, also working in biotechnology. We included employer representatives on the team because we are particularly interested in their point of view. From previous surveys we had learned that our employers value "soft skills" such as dependability, ability to work effectively on a team, and effective communication, more highly than success in academic courses.

Over the past ten years, the Biology faculty at Northeastern University has engaged in several curriculum change initiatives, and for each of these efforts, a major task was the development of learning goals for the curriculum. The results were remarkably similar each time, leading the faculty to agree that future efforts should focus on assessment of student performance rather than defining learning goals. After collecting and analyzing information from all the initiatives, the rubric development team agreed on 21 measures, or criteria, grouped into five major learning objectives. Our aim was to develop a rubric that would translate the learning objectives of the program into behavioral measures of proficiency gleaned from student "artifacts" - examples of student work such as lab reports, evaluations of student performance on co-op experiences, research papers, slide presentations, and study abroad reports deposited into student portfolios.

Software Development

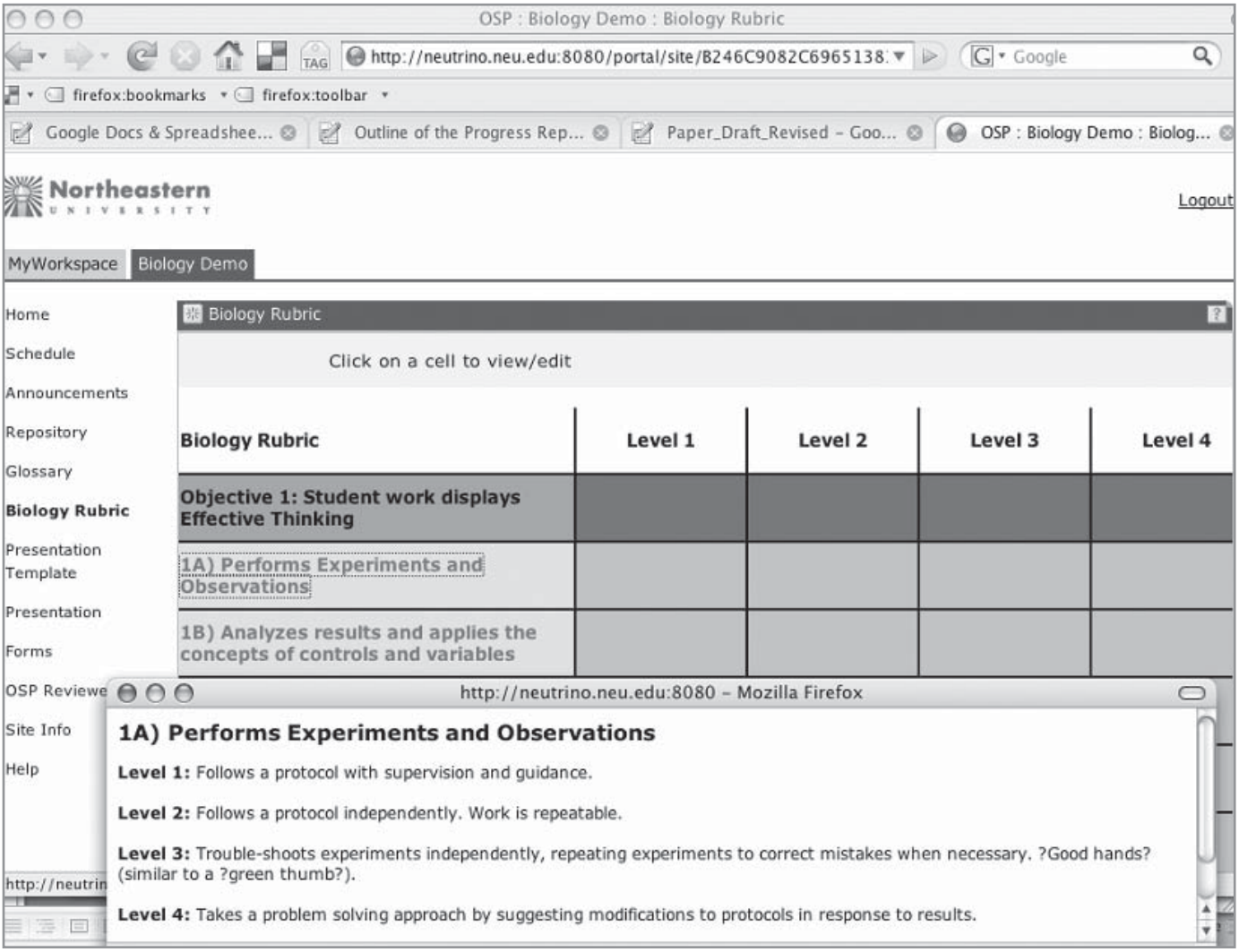
We selected the Open Source Portfolio (OSP) as our software platform for several reasons: 1) it offers features that fit well with curriculum assessment goals through the use of a rubric table; 2) it has the flexibility to

allow the student to choose different views for different audiences; and 3) we thought that using open source software would provide flexibility in customizing the solution and also provide an opportunity to assess the open source tool for wider adoption by the university. Finally, the software itself was free (although the effort required for configuration and development was significant).

The OSP 2.0.1 system used by students in this project has three primary components: a Repository, the Biology Rubric, and Presentations. The Repository is a personal storage area, with access restricted to the individual user (student). Items in the Repository may be organized in any way the user chooses using folders. Users may upload any type of file (e.g., documents, slide presentations, spreadsheets, images, movies, and pdf files).

OSP includes a matrix tool that allowed us to enter the Curriculum Assessment Biology Rubric described above. The five main objectives and associated criteria are listed in the left-hand column of the matrix. Each criterion is followed, horizontally, by four boxes representing levels of performance. When a user clicks on a criterion (e.g., "Performs experiments and observations"), a pop-up window displays the descriptions for each performance level (referred to as "expectations" by OSP) (See Figure 1).

Figure 1. Online Screen Shot – Rubric/Student Interface



To place an artifact in the Rubric, the student selects a file that provides evidence of meeting a specific criterion and uses the level descriptions to evaluate his/her level of performance for that criterion. The student then clicks on the appropriate rubric cell and associates the file (or files) with that cell. OSP provides a "wizard" to guide the student through a reflection process about their entry. The wizard asks the student to respond to three questions: What evidence demonstrates that you have met this expectation? Explain how your evidence demonstrates this expectation? and, How has your understanding of these criteria changed as a result of your having created the above evidence? The student's Rubric is visible to the student who created it and the instructors, but not to other students or reviewers.

OSP allows students to share their work through Presentations created using presentation templates. We created two presentation templates, written in xslt and xml, for this implementation: a Rubric Presentation and a Resume Presentation. The Rubric Presentation simply displays whatever content a student completes in his/her rubric by listing each criterion for which the student has made an entry, associated files, and the student's reflection. Rubric Presentations are updated automatically by the system to reflect any changes the students make to their rubric table. Students can choose to share these presentations with instructors, other students, and/or individuals outside the system. The Resume Presentation displays information about academic and experiential education, work history, and references. For each major section, a student may attach sample files from the repository, or links to material elsewhere on the internet. Like the Rubric presentation, the student can choose to share the resume with instructors, students, and/or individuals outside the system. The goal here was to create an *enhanced electronic resume* that the student is able to share when applying for a job or graduate school. We believe this can provide a strong motivation for student participation.

Web-Based Performance Evaluation - Co-op Education

For most biology/biochemistry students at Northeastern, the major source of experiential education is participation in the Cooperative Education Program (Co-op). Students earn experiential learning credit upon satisfactory completion of the three components of the Co-op learning process: preparation, activity, and reflection. During the activity component of the program, students typically spend six months in a planned career-related work experience where they have the opportunity to apply critical concepts and skills learned in class. While they are on Co-op, students are supervised and mentored by professionals in their field. Employers are expected to complete a performance evaluation for each student.

The Co-op student performance evaluation process provides employers with an effective method to assess the performance of Co-op students across multiple attributes (e.g., writing abilities, oral presentation skills, quality of overall performance). It requires active participation by students and by employers at each of three time points during the six month work experience: Goal Setting during the initial 2 weeks of employment (learning and performance goals); a Mid-Term Evaluation; and a Final Evaluation.

Early in this project, we realized that few students placed artifacts into the biology rubric that related to their experiential education. This is a serious problem for the process of assessment and curriculum improvement. In discussions, we realized that the major artifact from the Co-op experience would be the employer evaluation of student performance at the end of the Co-op period, and that this artifact was not readily available for placement into the rubric because it was handwritten.

In response, we initiated a side project using a web application to process employer evaluations. This application was designed to produce a pdf document that simulates the University-approved standard paper evaluation form. In keeping with the integrated learning model adopted by the university, our goal was to provide at least three opportunities for formal communication and feedback between the student and the supervisor during the six month Co-op period.

PRELIMINARY RESULTS/DISCUSSION

OSP Reporting Tool

The greatest limitation of the OSP 2.0 software in meeting the goals of this project was a lack of tools for reporting and analyzing data across the rubric table rather than student by student. Since OSP maintains all the data in an accessible SQL database, we were able to write a web tool that: 1) Displays the number of artifacts submitted for each level of each criterion; 2) Allows an administrator to access all the artifacts submitted for a specific criterion and performance level and then assign artifacts for review; 3) Allows reviewers to access assigned artifacts and enter an evaluation level and comments; and 4) Allows an administrator to view reviewers' evaluations and comments.

Student Self-Assessment with Portfolio System

The students in this study were enrolled in the Biology Department "Capstone" course, developed with the explicit goal of helping students integrate and assess the concepts and skills learned from the entire curriculum, including both experiential and classroom-based components. This course is a graduation requirement for all biology/biochemistry majors. Students in each of two successive classes (spring 2006 and fall 2006) were asked to deposit at least five artifacts into the rubric table, as a whole, but were given no further instruction about which of the 21 criteria should be selected. We collected more than 500 student artifacts with accompanying reflections and self-assessments. The results from the two offerings of the capstone class were remarkably similar. Table 1 shows the distribution of artifact placement report for the fall 2006 Capstone class (N=50).

Collectively, each class addressed all criteria, but the pattern of responses is clearly not uniform. For example, we had anticipated that artifacts generated as the result of experiential education such as the Co-op program would be placed within Objective #3. Since we have now implemented the web-based performance evaluation system we expect that students in future Capstone classes will place their final evaluations (in pdf format) into their ePortfolios under this objective.

Table 1. Summary Results: Distribution of Student Artifacts

	PERFORMANCE LEVEL				
Objectives and Criteria	1	2	3	4	TOTAL
Objective 1: Student work displays effective thinking					
1A) Performs Experiments and Observations	0	21	12	7	40
1B) Analyzes results and applies the concepts of controls and variables	0	10	16	3	29
1C) Relates biological structure and function	2	0	6	4	12
1D) Uses an evolutionary perspective	0	0	2	1	3
1E) Critically assesses the hypotheses, assumptions, and conclusions of others, including those from popular and scientific literature	1	6	8	14	29

Objective 2: Student work displays effective communication					
2A) Connects primary lit and results from experiments/field observations	2	1	10	5	18
2B) Presents organized experimental data	0	6	11	7	24
2C) Writes clearly and concisely	0	5	21	8	34
2D) Demonstrates effective oral presentation skills	0	3	9	14	26
2E) Participates and contributes when working with others	0	3	4	6	13
Objective 3: Student work displays a professional presence					
3A) Demonstrates time management skills	0	2	3	2	7
3B) Displays reliability	0	0	0	2	2
3C) Projects a professional attitude	0	2	6	4	12
3D) Exhibits responsible and competent judgment	0	2	4	1	7
Objective 4: Student work reflects an understanding of the social & cultural context					
4A) Considers historical, philosophical, economic, and governmental/political realities when making judgments	0	4	11	5	20
4B) Considers consequences of development and advancement	0	0	5	0	5
4C) Considers information in international and multicultural contexts	0	2	1	0	3
4D) Collects, records, and interprets data responsibly	0	3	10	1	14
4E) Identifies, analyzes, and justifies positions on ethical issues	0	2	11	0	13
Objective 5: Student work reflects an integrated educational experience					
5A) Reflects on undergraduate experiences through writing	4	4	7	7	22
5B) Connects ideas and experiences across disciplines and settings	0	3	7	11	21
Total Number of Student Artifacts: 354					

Student Reactions to Rubric/OSP Project

Students were asked to comment anonymously on the portfolio assignment. Two thirds of those who commented found the assignment useful; however, many expressed doubt that it would ever be useful in finding a job. Several students commented that they were impressed by how much work they had produced. Others were quite frustrated that they had not been told about the portfolio assignment earlier in their academic career, feeling it was unfair that they were now expected to produce old material (particularly since they had experienced numerous problems with computer upgrades and disk crashes during their college careers). From our point of view, the difficulties were more imagined than real since all students were able to produce at least 5 artifacts of their work. Nevertheless, we take this concern seriously and have started encouraging students to save their work earlier in their academic careers using the new campus-wide remote electronic repository system called "Myfiles" (Northeastern's implementation of Xythostm).

Expert Review of Student Artifacts

To assess the need for curriculum improvement, the next step in this project was to bring together a panel of expert reviewers to evaluate the artifacts that were placed into the Biology Rubric. To date, we have completed a preliminary pilot study in which three Northeastern University biology department faculty members and two employers (scientists at the Ph.D. level) participated. The primary goal of this pilot was to obtain data about inter-rater reliability. Reviewers were asked to rate individual student artifacts and reflections, according to the criteria and performance levels set by the rubric. We were encouraged that reviewers reported that they found the web-based review process using the rubric to be straightforward.

Seven of the eight artifacts selected for the pilot study represented artifacts placed into Objectives 1 and 2, because most of the artifacts rated by students fell within these two objectives. We also selected one artifact representing Objective 3. We deliberately selected artifacts with student ratings of 3 or 4 because we wished to have reviewers rate examples of students' "best" work. Finally, we decided that each artifact would be rated by at least two different reviewers and we selected two artifacts to be rated by all reviewers.

Table 2. Pilot Study – Student Artifacts: Expert Review vs. Student Self Evaluation

	Student Self Evaluation	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5
Objective 1: Student work displays effective thinking						
1A) Performs experiments and observations						
Student Artifact 1	4	4			4	
Student Artifact 2	3			1		2
Student Artifact 3	3	3	3	2	3	2
Student Artifact 4	3		4	3		
Objective 2: Student work displays effective communication						
2C) Writes clearly and concisely						
Student Artifact 5	3	2	2	2	2	2
Student Artifact 6	3	4				3
Student Artifact 7	3		3		2	
Objective 3: Student Work displays a professional presence						
3C) Projects a professional attitiude						
Student Artifact 8	3	2	2	2	2	2

Table 2 shows the distribution of the artifacts among the reviewers and performance levels. The student self-evaluation ratings are included for comparison. These results suggested that inter-rater reliability was strong enough for us to proceed to a large-scale study.

Figure 2. Distribution of Expert Reviewer Ratings independent of Objectives or Criteria

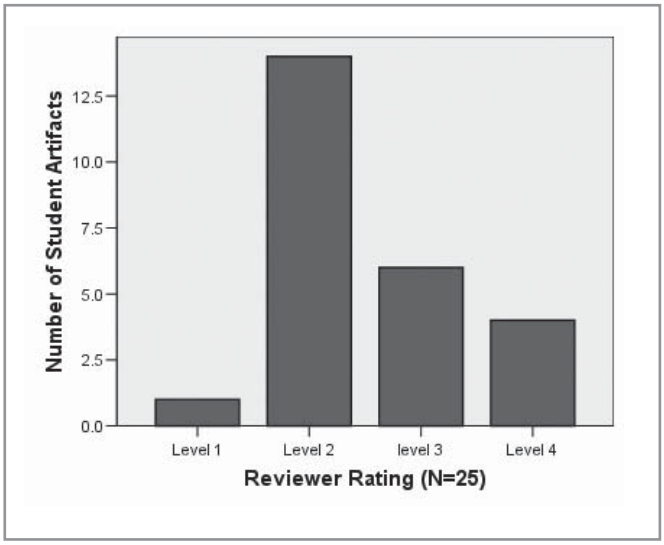


Figure 2 displays the overall distribution of performance ratings assigned by reviewers, independent of rubric objectives or specific criteria. Clearly, the expert reviewers had higher standards than the students. Although we do not intend to compare student ratings with expert reviewer ratings as a formal part of this project, we anticipated that student self-assessment ratings would be higher than those of the expert reviewers and were not surprised with this finding. Before this pilot we had some concerns that the distribution of expert ratings would not use the whole scale. So far that does not seem to be the case.

Web-Based Co-op Employer Performance Evaluation

The web-based performance evaluation process has resulted in many advantages for students and employers. It provides a centralized repository accessible to students and employers. It is available to students for use in building electronic portfolios to share with others (e.g. prospective employers), eliminates hand-written paper forms, and eliminates the possibility of lost or misplaced forms.

Table 3. Participation in Web-Based Performance Evaluation Process

	Spring 2006		Fall 2006	
	Percent Compliance	Number of Students	Percent Compliance	Number of Students
Goal Setting	10%	8	45%	39
Mid Term Evaluation	10%	8	43%	37
Final Evaluation*	84%	69	92%	80
[Submitted within 30 days of end of Co-op Period]*				

Interestingly, we have seen tremendous improvement in employer compliance in the submission of completed evaluations, as well as a substantial improvement in the time it takes for employers to submit completed evaluations after the Co-op period ends. As Table 3 shows, the return rate of final evaluations within 30 days of the end of the Co-op period was 84 percent for the spring 2006 Co-op period and 92 percent for the fall 2006 period. In prior years, it was not possible to document compliance with all three phases of the performance evaluation, but anecdotal reports suggest that it was spotty at best. The dramatic improvement in fall 2006 participation rates compared with the spring 2006 rates for Goal Setting and Mid-Term Evaluations lends support to the anecdotal reports and suggests that participation will improve as the users become more comfortable with the system. With paper forms, even with frequent reminders to the student and denial of credit for non-compliance, the typical rate for submission of the final evaluation form within 30 days of the end of the Co-op period was less than 50 percent.

Table 4. Co-op Employer Ratings - 10 Attributes (N=158 Students)

Attribute	Not Meeting Standard	Approaching Standard	Meets Standard	Exceeds Standard
Job Performance Level	0	9	51	99
Verbal Communication	0	4	73	81
Written Communication	0	11	78	68
Problem Solving Skills	0	11	84	62
Interpersonal Skills	0	3	46	110
Use of Technology	0	4	96	57
Professional Ethics	0	0	77	82
Professional Behavior	2	8	69	80
Work Ethics	0	12	44	103
Level of Supervision	1	6	52	100

With the implementation of web-based performance evaluation, we are able to easily assess student performance data across ten attributes, which were previously identified by Co-op employers as key indicators of workplace success. The combined results of the spring and fall 2006 Co-op terms are presented in Table 4. It seems apparent from the pattern of responses that Coop Employers are very satisfied, in general, with the level of performance provided by Co-op Students.

Figure 3. Co-op Employer Ratings – Select Attributes

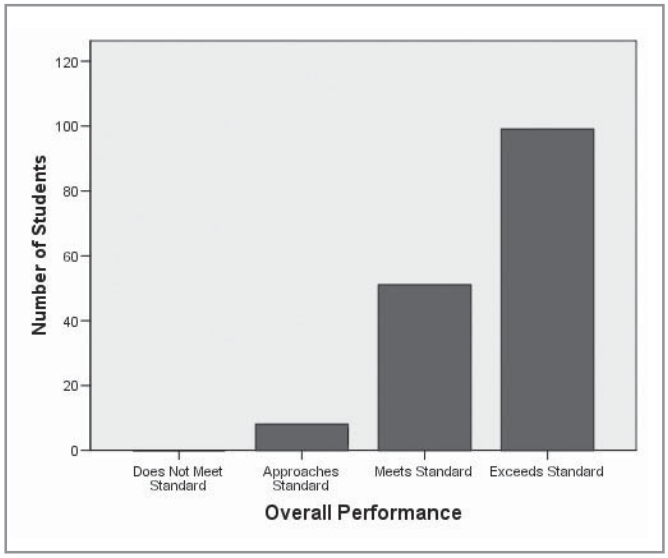


Figure 3A

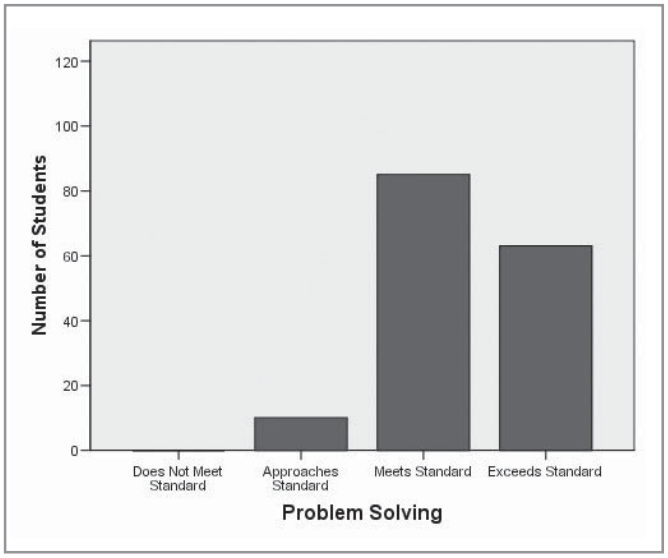


Figure 3B

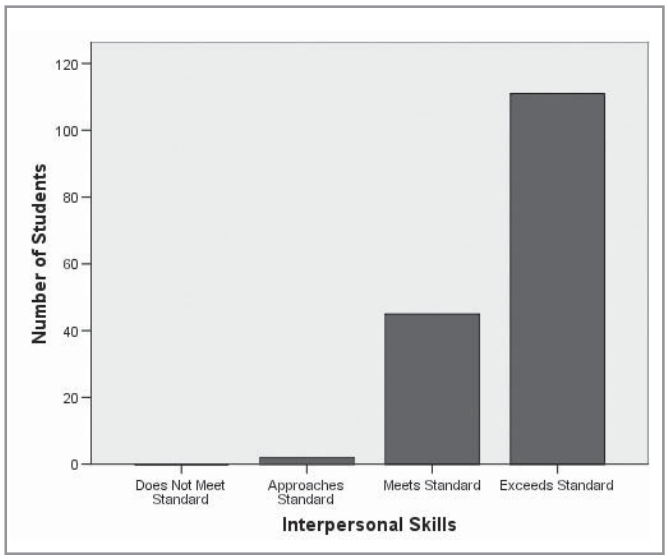


Figure 3C

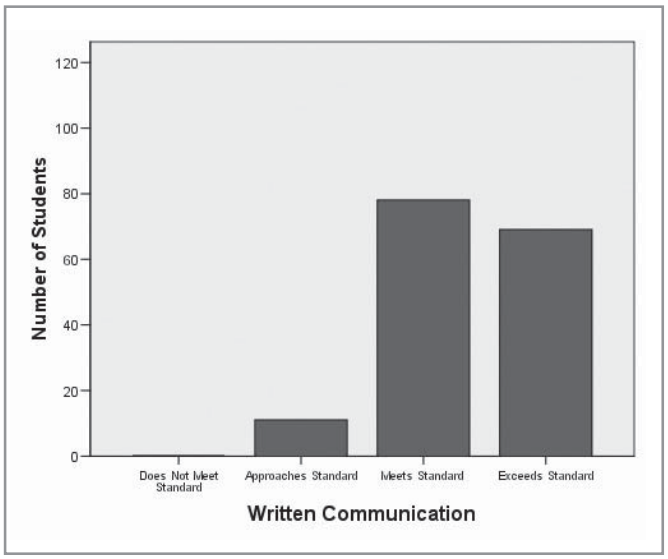


Figure 3D

However, further observation of these results, as shown in Figure 3, indicates that employers do evaluate students with deliberation, not just "sing their praises" indiscriminately. Although most students are rated as exceeding the standard for overall performance and interpersonal skills, they did not score as well in such areas as written communication and problem solving. Chi Square analysis, comparing employer scores for written communication and problem solving, using overall student scores as expected values, demonstrated that these differences were statistically significant (written communication: Chi Square= 24.89, 2 df; p<.001; problem solving: Chi Square= 36.42, 2 df; p<.001. Chi Square analysis comparing employer scores for interpersonal relationships against scores for overall performance indicated no significant differences, as was expected.)

LONG TERM GOALS

Rubric Assessment

The success of our initial pilot has given us the confidence to scale up the expert review of student artifacts. We have increased the number of expert reviewers and the number of artifacts for each reviewer. We

will continue to look at measures of inter-rater reliability as well as intra-rater reliability among the artifact ratings. We continue to see this review as a formative tool for curriculum improvement rather than as a graduation requirement for students.

An emphasis on writing in subject classes has become an important part of the new university-wide core curriculum implementation at Northeastern University. The Biology Department has designated certain courses as writing intensive and has agreed to make a major effort to develop writing assignments for the laboratory component of these courses. Results from our expert review will inform these efforts and, in turn, we will use the student portfolio/expert review system to assess the results.

Co-op Employer Evaluation

Although the current web-based process has resulted in substantial improvement in employer participation and provided better access to evaluation data, there is an opportunity to focus future work in this area on developing performance-based criteria for the work attributes that are important to employers.

Student Reflection

As the project progressed, the importance of the reflection process became apparent. The OSP architecture permits students to reflect upon their work when the work is placed into the Rubric. Students in the spring 2006 class struggled with this assignment, because they were not quite sure what was expected. For the fall 2006 class, we provided written guidelines for student reflection. We observed that students were able to complete this assignment with less difficulty. Because some of the student artifacts did not represent recent work, we began looking at the potential benefits of students reflecting upon their work as it is finished. Therefore, we plan to encourage faculty to require students to write a reflection piece on major assignments when they are completed and to store their work and reflections in the Myfiles repository. This effort will certainly facilitate portfolio building in subsequent years. We hope it will also help to create an atmosphere in which students value their past work and realize that it is worth saving after the semester is over and they have received a grade.

When students return from Co-op, they have individual meetings with their Co-op Faculty Coordinators, providing an opportunity to compare and contrast employer performance evaluations and student self evaluations, and discuss them from the perspective of academic and career goals. Student are expected to prepare post-meeting written reflections about what was discussed at the meeting and during his/her Co-op experience, after which, all three documents are deposited in the Myfiles repository. We expect that this process will lead to more artifacts from experiential education being included in future e-portfolios.

Implications for Student Learning and Curriculum Development

Student electronic portfolios that are managed using a powerful database can provide a wealth of data on student learning in various settings. Using expert review of student artifacts as an assessment tool we have just begun to process data we have collected over the course of a single academic year. We have already identified and acted on the need for more communication with Co-op employers and we can extend the web-based employer evaluation system to students in other experiential education settings. The Biology Department will use the methods developed in this project to assess the changes in instruction required by the new core curriculum that will start at Northeastern University for students entering in fall 2007.

For those who would like to have "hands-on" demonstration of our instance of the OSP Portfolio System, we have a live demo site you may visit http://neutrino.neu.edu:8080/portal. (Log in with user id "biology". The password is "demo". After the site opens, click on the "Biology Demo" tab.)

A Self-regulated Learning Assessment System for Electromechanical Engineering Technology Students

Seymour Blank*, John Hudesman, Ed Morton, and Robert Armstrong
New York City College of Technology / The City University of New York
Adam Moylan, Neisha White, and Barry Zimmerman
The City University of New York Graduate School and University Center

OVERVIEW

The purpose of this one-year exploratory project was to develop and implement a Self-regulated Learning Assessment System for students in a two-year associate's degree program in electromechanical engineering technology at the New York City College of Technology (NYCCT) of the City University of New York (CUNY). This was a continuation of previous work in Self-Regulated Learning (SRL) at NYCCT. Generally, the assessment intervention involved using classroom-based, formative assessments designed to help students better self-reflect on performance feedback via enhancement of sub-processes such as accurate self evaluation about goal progress and improved adaptation of learning or problem-solving strategies. Using a pre-posttest experimental design, students enrolled in either an Electrical Circuits course (N = 83), or a Digital Controls course (N = 55), were randomly assigned to either SRL or non-SRL classrooms. The salient outcomes in this experimental design included measures of attrition and academic achievement. Initial results provide support for the effectiveness of the SRL Assessment System. In addition, we began developing and pilot testing a computer-based approach to make the SRL Assessment System more effective and efficient.

INTRODUCTION TO THE PROBLEM: UNDERPREPARED STUDENTS, POOR RETENTION, AND LOW GRADUATION RATES

In the United States, the high attrition and poor academic achievement of many underprepared incoming students has greatly concerned two-year and technical colleges. First-year attrition rates have reportedly ranged from 36 percent to 50percent [1]. This has been attributed primarily to poor academic performance, which is correlated with the admission of those high-risk student populations that most two-year and technical colleges are committed to serving.

This national problem is also true for students enrolled in the Electromechanical Engineering (EM) Technology program at NYCCT, which is the bi-level technical college of the City University of New York (CUNY). NYCCT has an enrollment of more than 12,000 students, most of which enter the EM program via an open admissions policy. About 90percent of these students are from ethnic and racial minority groups. More than 50percent of our students were born outside of the United States and speak a language other than English at home. Only 29 percent of two-year college students report incomes as low as those of our students, and 58 percent of NYCCT students work at least 20 hours per week. The EM program is the largest of the college's two-year programs in engineering technology, with an enrollment of about 750 students. The associate-degree graduation rate for the program after six years ranges from 9 percent to 15 percent [2].

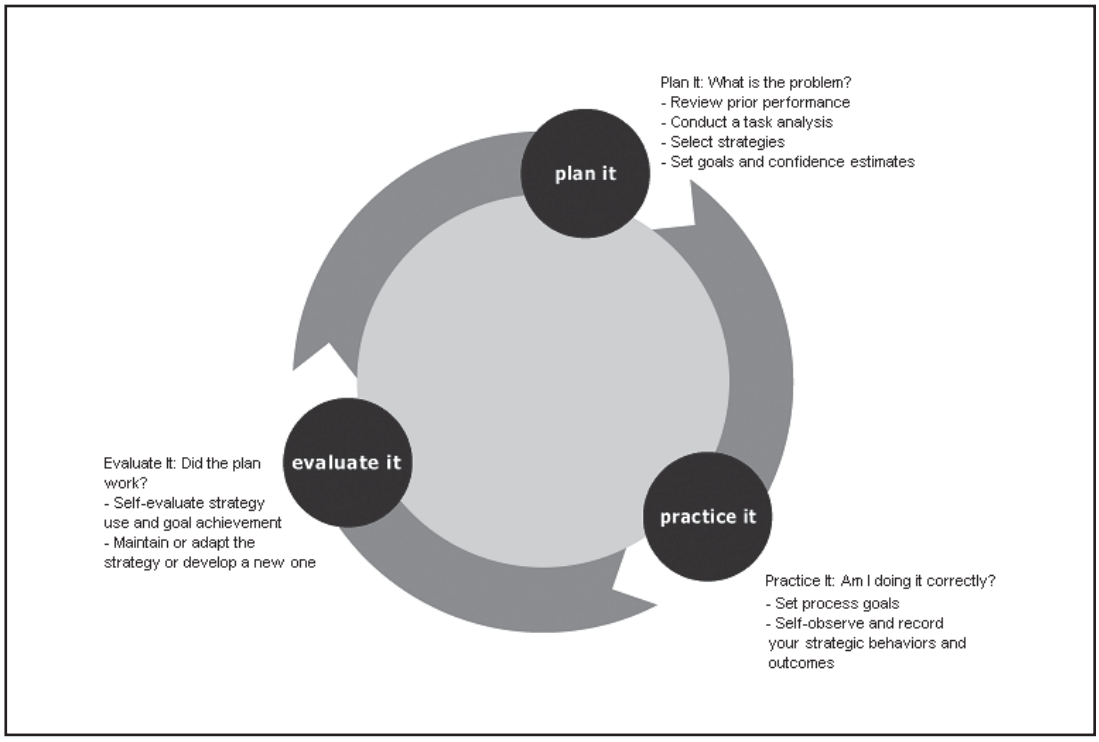
APPLICATION OF A SELF-REGULATED LEARNING MODEL

Most interventions designed to address the needs of these at-risk students have focused on teaching academic content together with a variety of academic/study skills, such as note taking and test taking. However, reviews of these intervention studies reveal that such programs do not sufficiently help students attain their

academic goals [3, 4]. It has often been found that students fail to demonstrate learning gains and do not sustain their use of these learning methods, largely because their effects are not well understood by the students or are difficult to use effectively.

What is missing from these interventions is an essential component of learning–the development of effective academic self-regulation. Self-regulated learning (SRL) involves metacognitive, behavioral, and motivational processes directed toward goal achievement [5,6,7]. The model of SRL used in this study [8,9] considers the learning process as a series of self-directed feedback cycles, each involving three phases: planning, practice, and evaluation (Figure 1).

Figure 1. SRL Model



During the planning phase, SRL students learn to assess their prior knowledge of an academic domain, breakdown the learning tasks into small steps, choose learning strategies that best address a specific academic challenge, and establish achievable process goals to regulate their learning. During the practice phase, students learn to monitor progress, including the implementation of their plans. In the evaluation phase, students evaluate their performance in relation to their goals and reflect on the effectiveness of each of the learning strategies they implemented earlier. The students' responses from the evaluation phase then become the basis for the planning phase in the next iteration of the SRL cycle. Thus, implicit in the SRL model is the recognition that learning is cyclical, involving a series of learning cycles that bring students successively closer to their academic goals. Through deliberate practice with these self-regulatory processes, students become more skilled at using both metacognitive and external feedback to continuously adjust and improve their learning efforts.

ASSESSMENT IN THE SERVICE OF LEARNING

There is an increasing awareness of the limits of traditional classroom grading practices on fostering students' ability to regulate their learning. Traditional academic assessment procedures involve having students

take an examination, which the instructor grades and then returns to the students with comments on their errors. The typical student reaction is to look at the grade and then mentally "file" the exam score. Many students interpret their test scores in terms of categorical "right or wrong" indices of their academic ability [10]. This kind of evaluation and reaction is associated with students viewing learning simply as memorization, and it discourages systematic analysis of their errors, which is an essential process for working scientists, engineers, and technologists. By contrast, "assessment for learning" approaches are designed to help students respond to their test scores and other grades as opportunities to learn. In this fashion, test grades are used to monitor and evaluate learning progress and make future adjustments. A major review by Black and Wiliam [11] concluded that achievement gains from assessment for learning approaches were among the largest effect sizes of any educational interventions.

The goal of this study was to develop an SRL Assessment System that combines the assessment-for-learning approach with an empirically grounded SRL model. Within this framework students learn to track and assess more effectively their academic learning and self-regulation skills. The process starts with quizzes (Figure 2) formatted so that, before answering each question, students make a self-efficacy judgment [12] about how confident they are that they can correctly answer the question. After answering the question, students make a self-evaluative judgment about how confident they are that they correctly answered the question. When the corrected quizzes are returned at the following class session, students have the opportunity to learn from their errors by completing a quiz reflection form, which is divided into three sections that correspond to the three phases of the SRL model (Figure 3).

Figure 2. Example quiz form

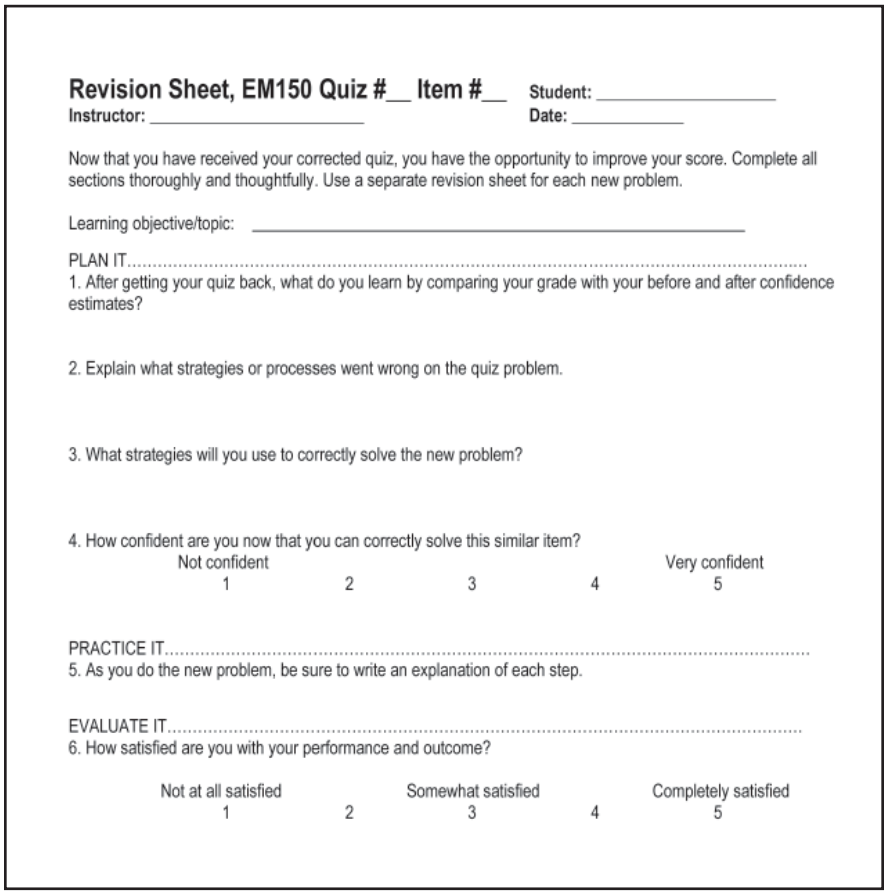
Name: _____

Date: _____

Quiz # – EM150
Topics:

Before solving each problem, circle the number that represents how confident you are that you can solve it correctly.	**Remember to show all your work.**	**After** you have solved each problem, circle the number that represents how confident are you that you solved it correctly.
1 · 2 · 3 · 4 · 5 1=not confident 5=very confident	1. Find the voltage in a circuit of 50 μA and 4.7 MΩ.	1 · 2 · 3 · 4 · 5 1=not confident 5=very confident
1 · 2 · 3 · 4 · 5 1=not confident 5=very confident	2. If a battery voltage equals 150 V and an ammeter reads 68.2 μA, what is the resistor value?	1 · 2 · 3 · 4 · 5 1=not confident 5=very confident
1 · 2 · 3 · 4 · 5 1=not confident 5=very confident	3. How are the current and resistance related?	1 · 2 · 3 · 4 · 5 1=not confident 5=very confident
1 · 2 · 3 · 4 · 5 1=not confident 5=very confident	4. If 200 W of power occurs for 40 s, how much energy in Joules is used?	1 · 2 · 3 · 4 · 5 1=not confident 5=very confident

Figure 3. Example reflection form



Revision Sheet, EM150 Quiz #__ Item #__ Student: _____
Instructor: _____ Date: _____

Now that you have received your corrected quiz, you have the opportunity to improve your score. Complete all sections thoroughly and thoughtfully. Use a separate revision sheet for each new problem.

Learning objective/topic: _____

PLAN IT.....
1. After getting your quiz back, what do you learn by comparing your grade with your before and after confidence estimates?

2. Explain what strategies or processes went wrong on the quiz problem.

3. What strategies will you use to correctly solve the new problem?

4. How confident are you now that you can correctly solve this similar item?

Not confident				Very confident
1	2	3	4	5

PRACTICE IT.....
5. As you do the new problem, be sure to write an explanation of each step.

EVALUATE IT.....
6. How satisfied are you with your performance and outcome?

Not at all satisfied		Somewhat satisfied		Completely satisfied
1	2	3	4	5

In the first of the three sections on the form, "Plan It", students compare their confidence ratings from the quiz with their graded performances. Students also describe what academic strategies or processes were not working, and then revise or establish new strategies, and rate their confidence that the new approach will, in fact, work. In the second section of the reflection form, "Practice It", students apply these new strategies to a new problem from the initial content domain. Furthermore, each step of the problem solving process has to be explained in the student's own words. Lastly, in the final section, "Evaluate It", students rate how satisfied they are with their performance on the reflection form. Instructors then review the self-reflection forms and award up to 80 percent of the difference between the original score and the full value of the question. Because of practical limitations, students could only submit one form per quiz, which had to be the problem with the lowest score.

COURSE DESCRIPTIONS AND STUDY DESIGN

The SRL Assessment System was piloted for one semester using an experimental pre-post test design. Students were registered for either a course in Electrical Circuits (N = 83) or a second-level course in Digital Controls (N = 55). Electrical Circuits is considered a first-semester gatekeeper course with a high level of student attrition and failure. Students in the second semester course in Digital Controls are successful survivors of the first course. Four sections of each course were included in this study. EM instructors trained in SRL theory and practice taught the two SRL sections of each course. After several initial training sessions, SRL instructors were observed four times during the semester and they attended weekly meetings to discuss the implementation of the program. Aside from the systematic use of the quiz and self-reflection forms, instructors also incorporated the SRL model and terminology throughout their classroom activities. Instructors with experience teaching the course, but with no

SRL training, taught the two control group sections of each course. Students were randomly assigned to either the SRL or non-SRL sections of each course.

At the start and end of the semester, students completed the short version of the Self-Efficacy for Learning Form that assesses students' beliefs about their ability to use various learning strategies. During the semester, instructors in each experimental and control group section gave at least eight quizzes, as well as uniform midterm and final examinations. These examinations were created and scored by EM instructors who were not otherwise involved in the study.

RESULTS

We collected data on a variety of academic and SRL-related measures that supported a number of important findings with respect to both attrition and academic achievement. Unfortunately, we were unable to gather complete data for all students on all salient experimental variables, such as student reflection form data.

Attrition

In the introductory Electrical Circuits course we found a significantly lower dropout rate in the SRL group (10 percent) than in the control group (26 percent), ($X2 = 3.785$, $p = .05$). Moreover, on average, students who dropped out of the course, whether in the SRL or control group, had lower prior GPA's and Elementary Algebra pretest scores, $t(76) = 1.66$, $p < .05$ (two-tailed). In the second level Digital Controls course there was only one student who dropped out. This is not surprising because this course serves somewhat more academically able students.

Achievement

In the Electrical Circuits course, the experimental and control groups were similar on a variety of pre-measures (Table 1). Likewise, as shown in Figure 4, both groups earned similar grades on the departmental midterm examination (SRL: $M = 64.08$, $SD = 18.18$; Control: $M = 64.82$, $SD = 21.17$, $p = .33$). However, by the end of the semester, the academic achievement of the SRL students as measured by their scores on a uniform final examination (SRL: $M = 54.80$, $SD = 22.48$) surpassed the control group students (Control: $M = 41.23$, $SD = 30.10$, $p = .04$), with a medium effect size, $d = .50$.

Figure 4. Achievement: Electrical Circuits

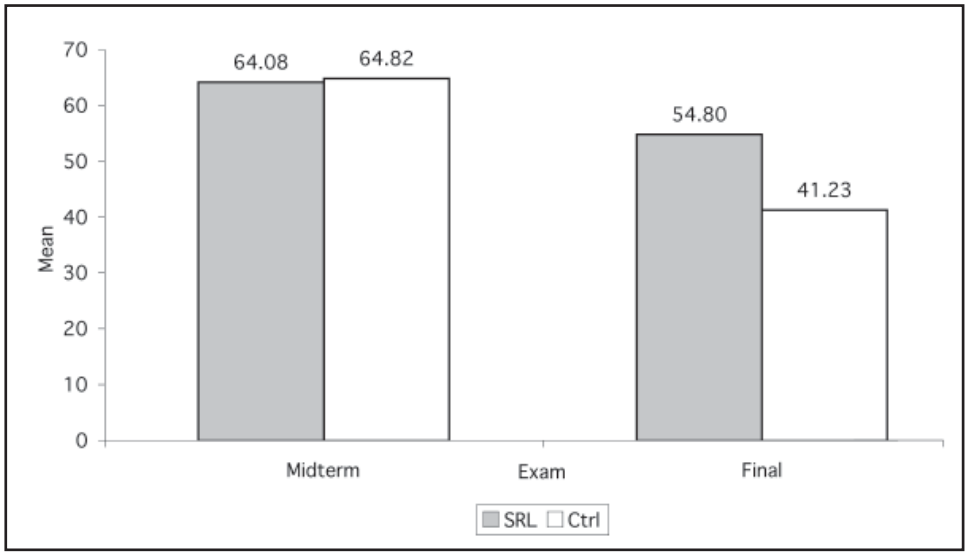
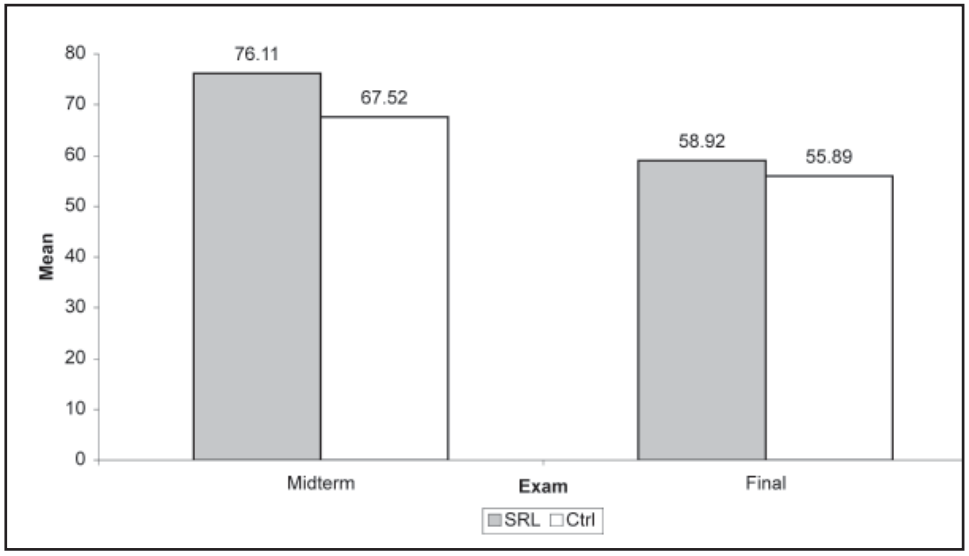


Table 1. Means and standard deviations for pre-measures by course and group

		Prior GPA			SELF pretest			EM pre-test		
Course	Group	n	M	SD	n	M	SD	n	M	SD
Electrical Circuits	SRL	36	2.31	.76	30	70.45	8.54	41	69.59	21.27
	Ctrl	34	1.95	1.01	12	63.79	11.66	37	76.92	19.30
Digital Controls	SRL	28	2.58	.86	28	73.66	11.75	22	73.45	26.66
	Ctrl	27	2.82	.56	27	72.98	14.05	27	69.59	23.64

In the second-level Digital Controls course, the SRL and control group students had equivalent scores on all of the pre-measures (Table 1). On the midterm examination, the SRL group had a mean of 76.11 (SRL: M=76.11, SD = 19.01), while the control group's mean was 67.52 (Control: M=67.52, SD = 18.23), see Figure 5. The difference in means was not statistically significant at the .05 alpha level, p =.10, d = .45. The observed gap in achievement diminished by the time of the final examination, at least based on the available evidence (SRL: M = 58.92, SD = 19.25; Control: M = 55.89, SD = 15.86, p =.58). Unfortunately, there were no final examination data for the control classroom that had scored the lowest on the midterm examination because the instructor mistakenly administered an examination that was different from the uniform final given to students in the other sections.

Figure 5. Achievement: Digital Controls



Self-regulated Learning Gains

The pre-post measure of SRL, the Self-Efficacy for Learning Form, did not show any significant change in scores from pretest to posttest. A possible explanation for this finding is presented in the discussion section.

Instructors' Comments

We asked the course instructors to give us their perspective on how teaching an SRL section differed from their traditional teaching experience. Their comments included the following:

- *There was a definite increase in confidence among the students.*

- *I was able to comfortably employ more group activities with the active involvement of the students.*

• Students became more active learners, they took greater responsibility for their learning, and students learned to function as team players.

• The overall benefit included the students' ability to explain a strategy for solving the problem and to reinforce to the rest of the class the importance of detailing this strategy.

• We constantly discuss the strategies involved in solving problems. We review methods that improve the students' ability to accurately determine their confidence in solving a particular problem correctly. Moreover, using SRL techniques, we constantly discuss resources available to students to increase their knowledge.

• The atmosphere is an encouraging one—working with their peers. It creates the feeling that the student can ask any question. This technique appears to keep all students engaged.

• Students have an understanding that they must conduct a task analysis, select strategies, set goals, and make confidence estimates. In addition, students learn to self-monitor and self-evaluate strategy use and goal achievement.

• The paper-and-pencil format for the class quizzes and self-reflection forms are very time consuming. It is also hard to keep track of all of the data.

The Computerization of the SRL Assessment System

As mentioned in the instructor comments, the use of the SRL formatted quizzes and reflection forms can become burdensome for instructors and students. The paper-and-pencil format also made data collection very difficult and data analysis inefficient. In order to make this process more efficient, we developed a preliminary computerized model of the SRL Assessment System that allows students to complete quizzes and reflection forms using a Tablet PC along with Tablet PC support software [13]. The Tablet PC enables students to use a writing stylus to perform complex calculations and to draw circuits directly on the screen. This system has been successfully pilot tested by the two SRL instructors with a small sample of students. Our future goal is to expand this effort by allowing students to upload their work to a Blackboard website that presents graphic feedback about their academic performance and the development of their self-regulated learning skills.

DISCUSSION

While some instructors allow students to earn additional credit by revising their original test responses, we are unaware of any other systematic attempts to develop and empirically test an assessment for learning process that is specifically designed to target key self-regulatory and motivational processes. Overall, the mastery approach to instruction framed by our SRL Assessment System resulted in significant reductions in student attrition in the Electrical Circuits course and gains in achievement in some areas.

It is important to note that our statistical analyses of achievement outcomes included students who, ultimately, dropped out, thus possibly skewing our data. Indeed, many, if not most, of the students who dropped out did so because they were doing poorly. Moreover, we also found that the SRL sections retained more of these

poorly performing students than the control group sections, likely making it more difficult to detect treatment effects in some cases. We believe it is also worth mentioning that this was our research teams' first involvement with a second-level course (the Digital Controls course). It is possible that the similarity in academic achievement we observed in the SRL and control groups is indicative of the need to design SRL interventions differently for second-level technical courses, and to attend more carefully to the content and construction of the examinations used in these courses.

Keeping in mind that this was an initial, exploratory study, we encountered a number of design constraints and methodological issues that will be addressed in future studies. It is apparent, for example, that the quality of the academic achievement and "SRL measures" needs improvement from a psychometric perspective. We are currently developing self-report instruments with greater specificity to the targeted behaviors of our intervention. Further, we would like to report additional process-related and qualitative data pertaining to students from intervention classrooms to provide a more detailed examination of self-regulatory processes and self-motivational beliefs. Second, more effective precautionary actions are required regarding missing data. In the future, we expect that computerizing the assessment process will enable us to gather all completed quizzes and self-reflection forms more efficiently and effectively. Lastly, future projects will address the issue of teacher effects by including more classroom instructors in subsequent research designs, and by comparing the instructors' achievement outcomes from prior semesters with their post-study outcomes.

In summary, SRL involves making learning strategy adjustments as students reflect on their performance feedback. Through an ongoing, interactive process, students are able to use the self-reflection form as a tool to discern and identify the sources of their errors on quizzes and examinations, and employ strategies to correct them. We are optimistic that an SRL approach to assessment-for-learning will help students to use test-based feedback systematically to improve their strategic problem-solving processes and, therefore, improve their academic achievement. The SRL approach, we suspect, will also have the added benefit of reducing attrition in these key scientific and technical courses.

ACKNOWLEDGEMENTS

This material is based on work supported by the National Science Foundation under Grant No. DUE 0512527. The authors thank Bert Flugman for his study design suggestions, Howard Everson for his assistance in evaluating the program, and Bob Albano for his help in the design and implementation of our computer project.

REFERENCES

[1] Wyman, F.J., "A predictive model of retention rate at regional two-year colleges," *Community College Review*, vol. 25, pp. 29-58, 1997.

[2] Ondrus, S., *Six-Year Graduation Rates*, New York City College of Technology, Office of Assessment and Institutional Research, 2006.

[3] Hattie, J., Biggs, J., and Purdie, N., "Effects of learning skills interventions on student learning: A meta-analysis," *Review of Educational Research*, vol. 66, pp. 99-136, 1996.

[4] Simpson, M., Hyned, C., Nist, S., and Burrell, K., "College academic assistance programs and practices," *Educational Psychology Review,* vol. 9, pp. 39-87, 1997.

[5] Zimmerman, B. J., "A social cognitive view of self-regulated academic learning," *Journal of Educational Psychology*, vol. 81, 329-339, 1989.

[6] Zimmerman, B. J., "Self regulating academic learning and achievement: The emergence of a social cognitive perspective," *Educational Psychology Review*, vol. 82, pp. 297 306, 1990.

[7] Zimmerman, B. J., "Becoming a self-regulated learner: An overview," *Theory into Practice*, vol. 41, pp. 64-70, 2002.

[8] Hudesman, J., and Zimmerman, B. J., *A Comprehensive Cognitive Skills Academy for Associate Degree Freshmen*. Fund for the Improvement of Post Secondary Education, United States Department of Education, #P116B010127, 2001-2003.

[9] Hudesman, J., and Zimmerman, B. J., *The Dissemination of the Self-Regulated Learning Model to Improve Student Performance in High Schools and Colleges.* Fund for the Improvement of Post Secondary Education, United States Department of Education, #R305H060018, 2006-2008.

[10] Schoenfeld, A., "Learning to think mathematically: Problem solving, metacognition, and Sense making in mathematics," In Grouws, D.A. (ed.), *Handbook of research on mathematics teaching and learning*, (Macmillan, New York), pp. 334-370, 1992.

[11] Black, P., and Wiliam, D., "Assessment and classroom learning," *Assessment in Education: Principles, Policy & Practice,* vol. 5, pp. 7-71, 1998.

[12] Bandura, A., Self-efficacy: *The exercise of control,* (Freeman, New York), 1997.

Rapid Assessment for Improved Student Learning and Satisfaction

John C. Chen*, Dexter C. Whittinghill, and Jennifer A. Kadlowec
Rowan University

ABSTRACT

In this project our goal is to improve student learning in foundation engineering courses. Our hypothesis is that learning is improved by providing rapid feedback to students on their understanding of key concepts and skills being taught. This hypothesis was tested through experiments in which student performance on quizzes was measured after classes in which some were and some were not provided rapid feedback. The feedback is enabled through wireless-networked handheld computers (PDAs). In each of the past two years, the feedback system was implemented in two sections of Statics. A crossover design of experiment was used. Student performance on a quiz at the end of each treatment period provided the data for comparison between the two groups. A general linear statistical model was used to analyze the treatment factor while controlling for the other "nuisance" or confounding factors. We found a significant and positive effect when students received feedback.

INTRODUCTION

Core engineering courses, such as Statics, teach key concepts and skills that students need to master in order to succeed in follow-on courses. Students must comprehend these concepts at sufficient depth (as opposed to rote memorization of procedure) and transfer this understanding to other courses and contexts. In this multiyear project, our hypothesis is that such learning is facilitated in an active, peer-assisted environment in which the students are provided frequent and rapid feedback on their state of learning.

BACKGROUND AND MOTIVATION

Providing feedback to students on their current level of understanding of concepts is critical for effective learning. It is also important for the professor. This feedback is typically realized through homework sets, quizzes, and tests. All of these techniques, however, suffer the faults of being too slow, too late, and too tedious to apply frequently. Freeman and McKenzie [1] discuss several issues that inhibit better student learning in higher education. For students, there is a lack of individual feedback on learning; few opportunities for dialogue to improve learning; and a feeling that the subject is impersonal. From the faculty members' perspective, the difficulties lie in knowing what students are really learning, providing individualized feedback, addressing students' specific misconceptions, attending to diverse learning styles, and engaging students in learning.

Bransford et al. [2] state: "Learners are most successful if they are mindful of themselves as learners and thinkers. In order for learners to gain insight into their learning and their understanding, frequent feedback is critical. Students need to monitor their learning and actively evaluate their strategies and their current levels of understanding." Freeman and McKenzie [1] support this idea, noting that "Feedback is fundamental to learning... Students may receive grades on tests and essays, but these are summative assessments… What are needed are formative assessments, which provide students with opportunities to revise and improve the quality of their thinking and understanding. If the goal is to enhance understanding and applicability of knowledge, it is not sufficient to provide assessments that focus primarily on memory for facts and formulas.

Our project addresses these issues by providing students with timely feedback and opportunities to improve learning. Our goal is to combine rapid feedback with conceptual learning and skills development and to evaluate our methods through rigorous experimental design and data analysis.

PROJECT DESIGN AND IMPLEMENTATION

Course Description

At Rowan University, Statics is a required course for sophomores in three of the four engineering disciplines (Civil & Environmental, Electrical & Computer, and Mechanical Engineering). The course content is similar to that of most engineering programs in the United States, although the pace and length of the course is unusual. Rowan students take Statics in a compressed, half-semester (7.5 weeks) format, with classes meeting for three 75-minute periods each week. Students receive two semester-hour credits upon passing the course. The format dictates a faster-than-usual pace of coverage of the material with little time spent in reviewing course material from previous lectures. Statics is delivered in the first half of the fall semester, followed in the second half-semester by Dynamics. In the first half of the spring semester, Civil & Environmental and Mechanical Engineering students continue in the engineering mechanics sequence by taking Solid Mechanics (also known as Mechanics of Materials).

In fall 2003, we began this study with one of the authors teaching two sections of this course. We collected some data to practice for what we might expect in the following years and focused on the details of implementing this project. Essentially, we treated the year as a "trial run." For example, we acquired all the personal digital assistants (PDAs) that were to be used for this study, set up, tested, and practiced with the software used to collect data, and developed most of the quizzes for which rapid feedback would be provided to students. In fall 2004 and fall 2005, we repeated what was implemented in 2003 except that data were taken for subsequent analysis. As will be explained later, what differed in the two latter years was the control group used to compare with the treatment group, which received rapid feedback with the PDAs. One of the authors always taught two sections of Statics as part of this study. This was done in order to minimize any differences in teaching style or content between the two sections. Having a single professor also ensured that the two sections maintained the same pace through the course from day to day.

The in-class portion of this study is conducted in a similar manner to that described by Mazur [3]. The professor presents a new topic or concept for no more than 10-15 minutes, using traditional lecture, demonstration, or sample problem solutions. Thereafter, he poses a "concept question" or a "skill quiz" to gauge the students' understanding. If the student responses when feedback is used show that a high percentage of students do not understand the concept or have not mastered the skill, the professor elaborates on or further explains the topic. If the responses show that a reasonable fraction of students understands (a distribution of answers, but a plurality with the correct answer), the professor directs the students to take time and explain the concept or skill to each other. Thereafter, the students are asked to either respond again to the same question, or a different question on the same topic may be posed. The final scenario occurs when the student response shows a high percentage of correct answers, indicating that students understand the topic. In this case, the professor simply continues to the next topic.

Traditional assessment methods were used to determine a student's course grade. In addition to assigned homework sets, which were completed by students in two-person teams, quizzes and tests were given. In the 7.5-week period of the course, nine homework sets were assigned, and eight quizzes and two non-cumulative examinations were given. Identical homework sets were assigned to the two sections. Whenever a homework set was submitted by the students, a brief quiz based on a concept covered in the homework was given. Quizzes were designed to be similar, but not identical, between the two sections. The scores on the quizzes were analyzed, as described later, to assess for any treatment effect due to the feedback provided.

A crossover design of experiment is used in this study [4]. The method is intended to eliminate potential confounding factors that cannot be controlled for using a standard analysis of variance model. For example, students may not be randomly assigned to each of the two Statics sections (i.e., one section may have mostly electrical engineering students, who have a different motivation level than the other section, which might be populated mainly with mechanical engineering students), or the time at which each section is held may affect student performance. Without the crossover, a potential treatment effect would have been indistinguishable from a section effect.

In a crossover design, one of two study groups (course sections in this case) will be a "treatment" group randomly chosen to receive instruction with the PDA-enabled feedback system. The second group will act as the "control" for a fixed period of time, or "treatment period." For the next treatment period, the two sections simply swap the roles of treatment and control, and this continues for the duration of the course. In this manner, each student acts as his or her own control to eliminate the non-correctible confounders. This design has the additional advantages of eliminating any bias that may be introduced by the professor in course delivery in the two sections, and minimizing any attitude bias that may be displayed by students of either section due to receiving a single method of feedback for the entire course if swapping did not occur. The treatment periods generally lasted from two to five class meetings, as was determined to be logical based on the skills or concepts being covered during the period.

In fall 2004, the control group used a flashcard system, similar to that described by Mehta [5], to provide rapid feedback. In fall 2005, the control group used no feedback as a comparison with the treatment group. Although the students could not respond to the concept question or skill quiz using a rapid feedback method, the problem was still presented to the control group and the instructor used cooperative-learning methods in these sessions. The students were instructed to work cooperatively on each problem and were encouraged to provide answers, which were recorded on the whiteboard for the class. We emphasize that regardless of the feedback method, or its absence, the instructor used identical teaching methods in both sections of the course, which included various active learning techniques, and identical materials were presented.

Rapid Feedback Methods

The flashcard method for providing feedback to students was developed by Mehta. In short, double-sided and color-coded cards are used by students to display their answer to a question posed by the professor. Each card can display one of six possible responses. The cards provide a quick means for the professor to scan the class's response and qualitatively determine the distribution of answers. The students are also able to see the class's response by a quick visual scan. Because of this, the professor asks the students to respond simultaneously to the posed question so that students who are uncertain of their answer cannot vote with the majority response.

A fleet of PDAs is used for the PDA-enabled feedback method. All of the PDAs have wireless networking capabilities (802.11b or WiFi) and communicate with the professor's Windows XP Tablet PC using a peer-to-peer networking mode. Custom-designed software is used to manage the intercomputer communications and to record and display student responses from the PDAs.

Regardless of the feedback method used each time, the concept question or skill quiz is posed by the professor through his Tablet PC and is projected to the front of the class along with the possible solutions. The correct solution is embedded with incorrect answers, also known as "distractors," which are derived from common student mistakes or misconceptions. Students are given time to reflect on the question posed, discuss it with their peers, and then must select from the possible solutions. The major differences between this and the flashcard method are that the PDA/software-based method allows for 1) quantitative and permanent recording of the student responses for future review, and 2) a display of the tallied student responses, which is projected up on the screen nearly instantaneously after the students respond. As mentioned previously, in fall 2005, when the control group received no rapid feedback through PDAs or flashcards, the concept question or skill quiz was still presented to the students, and the professor used active learning techniques to encourage students to solve the problem and to share the solution with the class.

Data Analysis

This project is composed of three major components: the development of a suite of concept questions and skills quizzes for the course, the use of rapid feedback and peer-assisted learning in the classroom, and a comparison between the two methods of providing rapid feedback (or not) to students. The third component required the bulk of the statistical analysis. The goal of this analysis was to see if the method of implementing the rapid-feedback – using PDAs or flashcards or nothing – had an effect on the students' learning. The response variable tested is the score on a quiz for the corresponding period of instruction where one section had the treatment and the other the control. This would be done while controlling for factors (or variables) other than the treatment factor that might affect the scores.

To analyze the treatment factor while controlling for the other "nuisance" factors that could affect scores but are not attributable to the treatment, we employed the following general linear model using the DataDesk statistical software package:

$$y_{ijklm} = \mu + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \alpha_1 + \gamma(\alpha)_{j(i)} + \delta_k + \tau_1 + \varepsilon_{ijklm} \quad (1)$$

where:

y = the score on the quiz
μ = the grand mean (average score with no factors taken into account)
x_1 = the student's Calculus I grade
x_2 = the student's Calculus II grade
x_3 = the student's Physics I grade
α = the Section in which the student is enrolled
γ = the student nested in section, or Student-in-section
δ = the Period (or quiz)
τ = the Treatment (PDA = "treatment" and flashcard/no feedback = "control")
ε = random error

Students' Calculus I, Calculus II and Physics I grades were treated as continuous covariates in the analysis. The Section factor was discrete, and the Student factor was discrete, and nested in section (student 1 in Section

01 is not the same as student 1 in Section 02). The Period (or quiz) factor was discrete and included because some quiz topics may be intrinsically more difficult than others. The Treatment factor was discrete as well. Although the quiz scores in both years were skewed towards zero (i.e., they were bunched toward the higher scores), the residuals were nearly normal, and no transformation of the data was needed.

RESULTS AND DISCUSSION

We have previously described in detail the results from fall 2004 [6], so only a summary will be provided here. During that semester, we conducted a crossover experiment in which two sections of students were provided rapid feedback and their performance on a series of quizzes was compared. The two rapid feedback methods were using the PDAs and using the flashcards, as described earlier. The most important finding was that there was no statistically significant difference between the student performance between these two groups. In other words, it does not matter how one provides rapid feedback. Although we had thought that the "coolness" of the PDA might affect a student's learning, it really would only affect their interest during the physical activity in class of reporting their answers. In the end their scores were not influenced by which of the two feedback methods was used.

Student survey results from 2004 indicate that students overwhelmingly felt that having rapid feedback of their state of learning was helpful to them, regardless of the means of providing feedback. Specifically, a great majority of students felt that either method of feedback was at least "somewhat helpful" to their learning, with a significant preference for the PDAs over the flashcards. Hence, although the use of PDAs versus flashcards did not affect the actual learning (as measured by the analyses of the quiz scores), the use of PDAs was *perceived* by students to be more helpful to their learning than the flashcards. Finally, 65 percent of the students believed that they would have performed worse in a course in which rapid feedback was not provided, while the remainder believed they would have performed at the same level.

The rapid feedback also had impacts on the authors as instructors. Regardless of the feedback method, we had to be more organized for each class and to plan ahead in preparing skill and concept questions and placing them appropriately in the lecture period. We also found that posing the feedback question was useful to get students to refocus or review, even if a question was created "on the spot" during class. We observed that the students took the feedback quizzes quite seriously and tried hard to answer them correctly even though no grade was involved. This was an additional benefit in that the students were forced to think about the concepts now rather than later (or perhaps much later) when they sat down to do homework. The results of the rapid feedback questions also allowed us to note what concepts were most difficult for students and thus improve future instruction. While technical difficulties and set-up time may be cumbersome for the PDA/software system, especially for the novice user, the authors believe that the benefits for the students and the faculty far outweigh these negatives.

Table 1 below presents results of our data analyses for the fall 2005 cohort of Statics students. Recall that this cohort was subjected to a crossover comparison between having rapid feedback with the PDAs versus having no feedback. This comparison would allow us to determine the effect of having feedback or not on student learning, as a complement to the fall 2004 comparative study. Each row within the table represents a different statistical model used to analyze the data (scores from eight quizzes administered at the end of each treatment period). The most noteworthy finding is that for all models examined *the treatment (of having rapid feedback) was statistically significant*. That is, student scores on the quizzes were positively influenced by the provision of rapid feedback. We will now elaborate more on the details of the statistical analyses.

Some general observations can be made for all models that we examined. First, none of the covariates included (students' grade in Calculus I, Calculus II, and Physics I) were significant. This is in contrast to fall 2004, when we found that the quiz scores were dependent on the students' performance in Calculus II and Physics I. We hypothesized then that the students' grades in Calculus II were a reflection of their general abilities in mathematics, rather than specific skills learned in that course, and that Physics I grades were significant because most of the concepts in Statics are derived directly from application of physics concepts. The fact that we find no significance in these covariates in fall 2005 is somewhat puzzling, but not worrisome to us since the more important finding – that the treatment was significant – is not dependent on this.

The second general finding from fall 2005 is that the Section in which the student belonged was not significant. This simply means that there was no difference in performance between the two sections of the course despite the fact that they were offered on different days and times. Third, the Student factor was always significant (at a=0.05), which is not surprising since each student is expected to perform differently and somewhat consistently. Finally, we found that the treatment Period was highly significant (at a=0.001), which implies that the quizzes were inherently different in their degree of difficulty. Again, this is not surprising because some topics in Statics are easier than others, so this finding simply reflects that fact.

Table 1. Results of statistical analyses in various models of the data

Response	Covariates	← Factors →				Interactions (if any)
Score	1, 2, 3	Section	Student-in-Section*	Period***	Treatment* (p=.0318)	
Score	1, 2, 3	Section	Student-in-Section*	Period***	Treatment** (p=.0062)	Section/Period
Score	1, 2, 3	Section	Student-in-Section*	Period***	Treatment* (p=.0144)	Student/Treatment
Score	1, 2, 3	Section	Student-in-Section*	Period***	Treatment* (p=.0204)	Period/Treatment
Score	1, 2, 3	Section	Student-in-Section*	Period***	Treatment** (p=.0033)	Section/Period Student/Treatment

For each model the factors marked with "*" were significant at a = 0.05 (5%), with "**" at a = 0.01 (1%), and with "***" at a = 0.001 (0.1%). Underlined factors were "significant" at a = 0.10 (or 10%). Note that covariate 1 = Calculus I, 2 = Calculus II, and 3 = Physics I.

The first row of results in Table 1 shows the basic model, which does not examine any interactions between the factors. The treatment effect was significant at p=.0318. Subsequent models examined the two-way interactions between Section by Period, Student by Treatment, Period by Treatment, and finally, a two-way interaction between Section by Period and Student by Treatment. In none of these models were any two-way interactions significant at a=0.05. The important result from these models is that the Treatment effect was at least as significant as in the model without them [Equation (1)]. This confirms the statistical significance of our finding that the rapid feedback positively influenced the students' performance on the quizzes.

CONCLUSIONS

Our most noteworthy finding to date is that rapid feedback has a significant and positive effect on student performance. This confirms the value of providing frequent and rapid feedback to students. We theorize that this provides the students with knowledge of their state of learning, allows them to make adjustments in their strategies for learning, and encourages immediate reflection on and practice in the concept or skill at hand. Although we did not utilize the currently popular feedback devices known as "personal response systems" or "clickers" that are offered by several commercial vendors, our method of using handheld wireless computers no doubt is analogous to these devices, which are gaining in popularity in higher education. Our findings provide strong evidence for the usefulness of these feedback devices.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of the National Science Foundation through grants DUE-0243227 and EIA-0312868.

REFERENCES

[1] Freeman, M. and McKenzie, J., "Aligning Peer Assessment with Peer Learning for Large Classes: the Case for an Online Self and Peer Assessment System," In *Peer Learning in Higher Education*, Boud, D., Cohen, R., and Sampson, J., (eds.), (Kogan Page Limited, London), pp.156-169, 2001.

[2] Bransford, J.D., Brown, A.L., and Cocking, R.R., (eds.), *How People Learn: Brain, Mind, Experience, and School*, (National Academy Press, Washington, D.C.), 1999.

[3] Mazur, E, *Peer Instruction: A User's Manual,* (Prentice Hall, Upper Saddle River, N.J.), 1997.

[4] Mason, R.L., Gunst, R.F., and Hess, J.L., *Statistical Design and Analysis of Experiments, with Applications to Engineering and Science,* Wiley Series in Probability and Mathematical Statistics, (Wiley, New York), 1989.

[5] Mehta, S.I., "A Method for Instant Assessment and Active Learning," *J. of Engr. Educ*., vol. 84, p. 295, 1995.

[6] Chen, J., Kadlowec, J., and Whittinghill, D., "Using Technology for Concepts Learning and Rapid Feedback in Statics," *Proceedings of the 2005 ASEE Annual Conference*, Session 1668, Portland, Ore., 2005. See also, Kadlowec, J., Chen, J., and Whittinghill, D.C., "Using Rapid Feedback to Enhance Student Learning in Mechanics," *Proceedings of the 2005 Frontiers in Education Conference*, Indianapolis, Ind., 2005.

Transferable Assessments for Capstone Engineering Design

D.C. Davis* - *Washington State University* | **S.W. Beyerlein** - *University of Idaho*
Olakunle Harrison - *Tuskegee University* | **Phillip Thompson** - *Seattle University*
Michael Trevisan - *Washington State University*

ABSTRACT

Capstone engineering design courses play vital instructional and assessment roles in preparing engineering graduates for professional practice. Their effectiveness in either role is highly dependent on the quality of the assessment and feedback provided in these courses. This paper describes elements of a capstone engineering design course assessment system intended to meet this need. Three assessments are described to address three areas of performance derived from a model for design learning and achievement for capstone design courses. Assessment exercises comprise multiple exercises, each with a scoring rubric, to measure student achievement. Results may be used for providing formative feedback, assigning grades, or documenting achievements aligned with Accreditation Board for Engineering and Technology (ABET) outcomes.

INTRODUCTION

National leaders call for reform of engineering education to better prepare engineering graduates for the competitive global marketplace [1,2]. Among capabilities cited as deficient in student preparation are professional skills and abilities to innovate technical products in the context of business conditions [3,4]. Much of the professional skill development is assigned to capstone engineering design courses that incorporate open-ended design projects with significant professional challenges. Instructors of capstone design courses, however, indicate that instructional focus and assessment in these courses vary greatly depending on instructor preferences and abilities.

Educational assessment experts state the importance of classroom assessment for guiding effective instruction [5,6]. Therefore, high-quality classroom assessment in capstone design courses is vital due to the pivotal roles that capstone design courses play in engineering curricula. Also, because these courses are a required part of accredited engineering programs in the United States, they provide a rich environment for assessing a variety of student learning outcomes and associated program achievements.

A number of researchers report strategies for assessing student performance in capstone engineering design courses. Some instructors assess student work in capstone design courses by focusing more on design steps than on the quality of design products, while others focus on design products with little attention given to design processes [7,8]. However, success after graduation requires engineers to produce high-quality products while also refining their processes and developing professionally to support continuous improvement of product quality. This argues the importance of capstone design course outcomes related to both student learning and product development.

The authors of this paper are leading a National Science Foundation project, "Transferable Assessments for Capstone Engineering Design," which is developing and testing research-based assessments for capstone design courses across disciplines and institutions. In this process, they found that few engineering faculty use

proven assessments for formative assessment in capstone design courses [9]. This has justified development of a learner and solution development model for capstone design [10], and a framework for guiding development of sound assessments for capstone design courses [11]. These provide the basis for creating assessments for capstone engineering design courses.

GOAL

The goal of the NSF project, "Transferable Assessments for Capstone Engineering Design," is to produce versatile, sound classroom assessment instruments for assessing student achievement in capstone engineering design courses. This paper describes the set of assessments produced by this project.

AREAS OF PERFORMANCE

Capstone design courses teach students how to complete a significant design project within the constraints of one or more classes. These projects provide the context for developing student professional skills and for assessing their abilities to create a valuable design product. Thus, outcomes of a capstone engineering design course include both learner development outcomes and solution development outcomes, as defined below.

Learner development outcomes are defined under two areas of performance:

1) ***Personal capacity:*** *Individuals performing and improving individual skills essential to engineering design*
2) ***Team processes:*** *Teams developing and implementing collective processes that support team productivity in design*

Solution development outcomes are defined under two additional areas of performance:

3) ***Solution requirements:*** *Definition of targeted design solution performance and features expected to satisfy stakeholder needs and constraints*
4) ***Solution assets:*** *Results from a design project that meet needs and deliver satisfaction and value to key project stakeholders*

Performance criteria set expectations in the following areas:

Personal Capacity Performance Criterion: "Individuals accomplish challenging goals related to design by employing goal-driven initiative, competence in problem solving, integrity and professionalism, and ongoing reflective development of their personal abilities."

Team Processes Performance Criterion: "The team achieves challenging goals in productivity and team function by strategic use of team resources, synergistic collaboration, decisions that add real value, and assessment-driven refinement of processes."

Solution Requirements Performance Criterion: "Stated requirements reflect an in-depth understanding of customer needs, business issues, state of the technology, and societal concerns about the solution, while providing clear targets for the development of a valuable solution."

Solution Assets Performance Criterion: "Design solutions meet or exceed expectations of stakeholders by delivering proven value in desired functionality, economic benefits, implementation feasibility, and favorable impacts on society."

ASSESSMENT INSTRUMENTS

To date, assessment instruments have been developed for three of the four performance areas. Within each area, assessment exercises address critical factors embodied in the performance criterion. For example, for the *team processes* performance area, one assessment exercise (team process review) addresses students' abilities to perform, assess, and improve *team processes* that affect team productivity, while also addressing the ABET outcome of teamwork. A second assessment exercise (team member citizenship) addresses individual team member contributions and a member's ability to make insightful recommendations. Available assessment exercises are summarized in Table 1.

Details of the assessment exercises are too extensive for inclusion in this paper, so condensations are presented below. For details, refer to the project web site, part of the WSU Engineering Education Research Center web site, at: www.eerc.wsu.edu/ASA .

Table 1. Summary of Capstone Design Course Assessments Available

Performance Area	Assessment Exercise	Description of Assessment	ABET Outcomes Addressed
Personal Capacity	Personal Growth	400-500 word essay: reflective analysis of growth during project	• 3g Communication (written) • 3i Lifelong learning
	Professional Practices	400-500 word essay: analysis of professional/ ethical issues in project	• 3f Professional & ethical • 3g Communication (written)
Team Processes	Team Member Citizenship	Short answer: rating of member contributions; analysis for improvement	• 3d Teamwork
	Team Process Development	Short answer: analysis of strong and weak team processes	• 3d Teamwork
Solution Requirements	Stakeholder Needs	Short answer: identification of key stakeholders and needs	• 3h Solution impact
	Project Outcomes	Short response: statement of problem, expected outcomes, and benefits	• 3h Solution impact
	Solution Specifications	Short response: list of high-priority requirements with targeted outcomes	• 3h Solution impact

*(***ABET designations. 3d***: An ability to function on multi-disciplinary teams;* **3f***: An understanding of professional and ethical responsibility;* **3g***: An ability to communicate effectively;* **3h***: The broad education necessary to understand the impact of engineering solutions in a global, economic, environmental, and societal context;* **3i***: A recognition of the need for, and an ability to engage in, life-long learning)*

PERSONAL CAPACITY

Personal Growth paper

Write a 400 to 500 word reflective essay around a significant personal growth opportunity you encountered in your project this term. Begin by identifying a situation that challenged your knowledge, skills, or attitudes. Explain how you recognized the need for personal growth and the actions you took to address the challenge: your thoughts, judgments, goal setting, search for help, and steps to overcome the challenge. Describe impacts of your actions: learning, insights, impacts on the project, etc. Also reflect on how your challenge-driven personal growth experience will affect your ability to address personal challenges in the future.

Professional Practices paper

Write a 400 to 500 word paper describing your analysis of an ethical or professional situation encountered in your project this term. (Note: These issues can often arise around decisions on product cost or functionality, user safety, intellectual property, codes or standards, manufacturing processes, product disposal, marketing practices, product testing, or client interactions). If you did not personally address an ethical/professional dilemma, identify one from your project that should be addressed, then analyze this situation.

TEAM PROCESSES

Team Member Citizenship

Rate members of your team (including yourself) on their actions contributing to an effective team. In each cell, assign a person a rating (1-5) for each of the 12 actions, based on definitions given. Next, identify members' percent contributions to project achievements this term.

For each of your team members (including yourself):

a. Assess a key strength: Describe what makes it strong, and identify impacts this strength has in making your team work together better.

b. Recommend an improvement: Define specific actions to produce a desired improvement, and describe expected benefits these actions will have on making a better team.

Team Process Development

Describe one of your most effective team processes for achieving high team productivity. Identify and describe the process, and explain what worked well and why. Your explanations will be strongest when they show understanding of what makes this team process effective in enhancing team productivity.

Explain how to improve one of your weaker team processes that is important to achieving your team's goals. Name the team process, explain what improvements are needed, and describe steps the team should take to achieve desired improvement. Your explanations will be strongest when they show understanding of how to improve the process's effectiveness.

SOLUTION REQUIREMENTS

Stakeholder Needs

In each box provided, identify a person or group that fits that particular stakeholder category: customer/user, business/financial, technical, or society. Then define that person or group's most important needs that should be satisfied by your design solution. Demonstrate your understanding of key people in each of these four categories and needs they perceive to be most crucial to the success of your project.

Project Outcomes

Write a brief paragraph that describes the problem or opportunity being addressed, what will be delivered by your design team, and the benefits that should come from your solution. Show your understanding of the core problem, the solution, and benefits offered by the solution.

Solution Specifications

Identify the most important requirements to be met by your solution. Be specific enough so these can be used to determine your solution's success (e.g., cite standards, codes, or norms to be met). For each:

(a) Give a brief definition or description of the specification

(b) Give a value or other evidence indicating when the specification has been achieved

ASSESSMENT SCORING

Student performance on each assessment is scored based on the performance criterion for the respective performance area and the scope of the specific assessment exercise. A five-point scoring rubric has been established for each assessment exercise to define performances levels ranging from novice to expert.

Table 2 presents an example scoring rubric for the Professional Growth paper. It is used to score 1) personal growth and 2) written communication. Space is also provided for writing comments and suggestions for formative feedback.

Table 2. Scoring Rubric for Personal Growth Paper

		Scoring Scale				
Personal Growth		***1***	***2***	***3***	***4***	***5***
GOALS	**Understanding of need**	Clueless	Vague idea	General grasp	Comprehends	Insightful
	Growth goals	No goals	Phony	Vague	Clear, specific	Motivating
ACHIEVEMENT	**Plan to achieve goals**	No plan	Vague	Usable	Valuable	Strategic
	Achievement of growth	No growth	Little evidence	Moderate	Valued, proven	Strong, insightful
Written Communication		***1***	***2***	***3***	***4***	***5***
MECHANICS	**Word choice/ use**	Many incorrect	Distracting	Acceptable	Effective	Powerful
	Spelling/ punctuation	Many errors	Minor errors	No serious errors	Effective	Flawless
IMPACT	**Sentences/ paragraphs**	Misleading	Detracting	Acceptable	Effective	Eloquent
	Overall effectiveness	Damaging	Useless	Weak case	Effective	Persuasive

Personal Growth Score (1 – 5) : _____ | **Written Communication** Score (1 – 5) : _____ | Comments:

A five-point scoring rubric has been established for seven assessment exercises associated with the personal capacity, team processes, and solution requirements areas of performance. To conserve space in this paper, a list of factors used in scoring each exercise is presented in Table 3. Details of rubrics may be found at the project website at: www.eerc.wsu.edu/ASA.

Table 3. Performance Factors Used in Scoring Assessment Exercises

Assessment Exercise	Performance Factors Used in the Rubric
Personal Growth Paper	• Personal Growth: understanding of need, growth goals, plan to achieve goals, achievement of growth • Written Communication: word choices/use, spelling, punctuation, sentences/paragraphs, overall effectiveness
Professional Practices Paper	• Professional Practices: understanding issues, ethics/codes/policies, reasoning, outcome achieved • Written Communication: word choices/use, spelling, punctuation, sentences/paragraphs, overall effectiveness
Team Member Citizenship	• Rating Member Contributions: own contribution level, all member contribution levels, own value added, own vs. others' views • Improving Member Contributions: description of strengths, impacts of strengths, actions for improving, expected benefits
Team Process Development	• Effective Process: evidence of strength, understanding of strength • Proposed Improvements: understanding of need, plan for improvement
Stakeholder Needs	• Customer/User: need fit to group, understanding of needs • Business/Financial: need fit to group, understanding of needs • Technical: need fit to group, understanding of needs • Society: need fit to group, understanding of needs
Project Outcomes	• Problem Definition: scope of project, understanding of problem • Solution Envisioned: technical elements, non-technical elements • Solution Benefits: technical benefits, non-technical benefits
Solution Specifications	• Functional Performance: extent addressed, specificity of target • Financial or Business: extent addressed, specificity of target • Technical Feasibility: extent addressed, specificity of target • Social, Ethical, Professional: extent addressed, specificity of target

ASSESSMENT UTILIZATION

The capstone design course assessments may be used for formative feedback or for summative evaluation of performances. For effective use as formative assessments, the assessment exercises should focus on providing students feedback to guide their improvement. For summative assessments, the assessment exercises should focus on the assignment of scores that best represent student performances. Use of the same assessment exercises for both formative and summative purposes may be useful to document changes in student understanding and development over a given time period.

CONCLUSIONS

Improvement of capstone engineering design courses hinges on effective methods for measuring student performance. The credibility of assessment exercises discussed in this paper stems from the student and product development model desired in capstone engineering design courses. Performance criteria for four performance areas establish targets of importance to design education. Scoring rubrics aligned with the performance criteria direct feedback that relates to the desired performances. Using these assessments to guide instruction and feedback given to students will improve engineering design learning. Students will understand desired performance, and faculty will give feedback that guides improvement. These assessments offer the potential to elevate student achievement in line with both ABET outcomes and society's desires for engineering graduates.

ACKNOWLEDGEMENTS

Research defining outcomes, developing assessments, and testing assessments was funded by the Division of Undergraduate Education of the National Science Foundation under grant NSF/DUE 0404924.

REFERENCES

[1] The National Academies, *Rising above the gathering storm*, (National Academies Press, Washington, D.C.), 2006.

[2] The Steering Committee of the National Engineering Education Research Colloquies, "Special Report: The National Engineering Education Research Colloquies," *Journal of Engineering Education*, vol. 95, no. 4, p. 257, 2006.

[3] National Academy of Engineering, *The engineer of 2020: Visions of engineering in the new century,* (The National Academies Press, Washington, D.C.), 2004.

[4] Davis, D., Beyerlein, S., and Davis, I., "Deriving design course learning outcomes from a professional profile," *International Journal of Engineering Education*, vol. 22, no. 3, pp. 439-446, 2006.

[5] Stiggins, R., "Revitalizing classroom assessment: The highest instructional priority," *Phi Delta Kappan*, vol. 69, pp. 363-368, 1988.

[6] Black, P., and Wiliam, D., "Assessment and classroom learning," *Assessment in Education,* vol. 5, no. 1, pp. 7-74, 1998.

[7] McKenzie, L.M., "End-of-program assessment: An investigation of senior capstone design assessment practices", doctoral dissertation, College of Education, Washington State University, Pullman, Wash., 2002.

[8] Sobek, D.K., and Jain, J.K., "Two instruments for assessing design outcomes of capstone projects," *Proceedings of the American Society for Engineering Education Annual Conference & Exposition*, 2004.

[9] Trevisan, M., Davis, D., Beyerlein, S., Thompson, P., and Harrison, O., "A review of literature on assessment practices in capstone engineering design courses: Implications for formative assessment," *Proceedings of the American Society for Engineering Education Annual Conference & Exposition*, 2006.

[10] Davis, D., Beyerlein, S., Harrison, K., Thompson, P., Trevisan, M., and Mount, B., "A conceptual model for capstone engineering design performance and assessment," *Proceedings of the American Society for Engineering Education Annual Conference & Exposition*, 2006.

[11] Beyerlein, S., Davis, D., Trevisan, M., Harrison, K., and Thompson, P., "Assessment framework for capstone design courses," *Proceedings of the American Society for Engineering Education Annual Conference & Exposition*, 2006.

The Calculus Concept Inventory

Jerome Epstein* - *Polytechnic University*

INTRODUCTION

The Calculus Concept Inventory (CCI) is a test of conceptual understanding (and only that — there is no computation) of the most basic principles of differential calculus. The idea of such a test follows the *Force Concept Inventory* (FCI) in physics [1,2,3,4], a test which has spawned a dramatic movement of reform in physics education and a large quantity of high-quality research. The FCI showed immediately that a high fraction of students in basic physics emerged with little or no understanding of concepts that all faculty assumed their students knew at exit. More dramatic, the FCI in Hake's analysis showed a very dramatic correlation with teaching methodology, where the normalized gain in Interactive-Engagement (IE) sections exceeded that in Traditional Lecture (TL)-based sections by two standard deviations.

Mathematics education is largely mired in the "math wars" between "back-to-basics" advocates and "guided-discovery" believers. There is no possibility of any resolution to this contest between competing faiths without real scientific evidence of what works and what doesn't. Such evidence requires widespread agreement on a set of very basic concepts that students should be expected to master in, for example, first semester calculus. The CCI is a first element in such a development and is an attempt to define such a basic understanding.

The CCI has undergone extensive development and evaluation, funded by the National Science Foundation. It was developed by a panel of respected calculus educators and a consultant who is nationally known for development and validation of standardized tests. The test shows good performance characteristics (see below) and exposes exactly what the physics test showed. In the fall semester of 2006 the test was given at two new test sites with programs using alternative teaching methodologies. This made comparison of gain from two widely divergent methodologies possible, which will be discussed in this paper.

The paper will also discuss the development and validation process in detail. Statistical parameters (p-values, discrimination, and the reliability number) are discussed for the test as well. Those interested in using the test should contact the author. Maintaining the security of the test is a crucial issue.

The CCI is the first in an anticipated series of "Basic Conceptual Understanding" instruments for various levels in mathematics (including high school and earlier) that can hopefully serve to provide a scientific basis for discussions about teaching methodology and curricula. We are currently seeking funding for an Elementary Algebra Concept Inventory.

CONCEPT INVENTORIES

The production of "concept inventories" has become a small cottage industry. These are tests of the *most basic* comprehension of foundations of a subject, rather than computation. They are quite different from final exams and make no pretense of testing everything in a course. All of them trace their roots to the FCI in physics [1,2, 5, 6], and there is general agreement that physics education is well ahead of other disciplines in the use of concept tests as measures of teaching effectiveness. The FCI consists of multiple-choice items that

test understanding of the basic foundations of Newtonian mechanics. The questions are carefully designed to test ability to use fundamental physical laws and principles in simple, qualitative, yet profound situations, where calculations are neither needed nor helpful. The FCI is designed to measure conceptual understanding that is considered to be absolutely fundamental for any useful understanding of physics. Halloun and Hestenes say in their abstract [1]:

> *"An instrument to assess the basic knowledge state of students taking a first course in physics has been designed and validated. Measurements with the instrument show that the student's initial qualitative, common sense beliefs . . . have a large effect on performance in physics. But conventional instruction induces only a small change in those beliefs."*

Both the FCI in physics and the CCI in calculus show that traditional instruction has remarkably little effect on basic conceptual understanding, and this has been the greatest shock to faculty. Research dating back at least 30 years has shown that most students emerge from standard introductory courses without a solid grasp of the basic concepts. This was clearly documented in physics by Arons [7,8]. But prior to the development of the FCI, there was no generally accepted measure of how well students understood the basic concepts. It was thus difficult, if not impossible, to convince faculty of a need to consider changing the way they taught.

Results from research using the FCI have caused a dramatic transformation in a modest, but rapidly increasing, number of physics programs in the last ten years. There are two main reasons why the FCI has been so effective in changing views, and these are instructive for mathematics also. First, faculty recognize in the FCI questions that arise in any practical use of basic principles, *including* those requiring standard computations. All acknowledge that the concepts measured are absolutely necessary (but not sufficient) for any useful understanding. Second, Hake [4, 9] has shown that the FCI provides a *reproducible* and *objective* measure of how a course improves *comprehension* of principles, not merely how bright or prepared the students are, nor what they have memorized. In a study of some 20 institutions, 60 classes, and 6000 students Hake compared FCI scores at entry with scores at exit. Patterns found in the data led to a performance measure that Hake calls the normalized gain. The FCI is administered once at the start and once at the end. The class performance is measured by the *normalized gain*, defined to be

$$g = \frac{\mu_f - \mu_0}{100 - \mu_0},$$

where μ_0 is the mean score of the class at the start and μ_f is the mean score at the end (in percent correct). This measures the gain in the class's performance on the FCI as a fraction of the maximum possible gain. Few of the groups studied had a normalized gain much less than 0.15. On the other hand, the best performing classes in recent studies in physics have a normalized gain of about 0.70.

Hake's findings are striking. They show that g is independent of the level μ_0 of the students at entrance, and largely independent of instructor and text. It is, however, strongly dependent on the teaching methodology used. Classes that used a *Traditional-Lecture* (TL) approach had an average normalized gain of 0.23 (standard deviation 0.04). In contrast, classes that used an *Interactive Engagement* (IE) approach had an average normalized gain of 0.48 (SD = 0.14), roughly two standard deviations above that of the TL classes. The consistency and

predictability, and the strong correlation with teaching methodology, make this difficult to ignore. They provide strong evidence that IE methods are more effective than TL methods. An increasing number of departments use FCI results to measure the effectiveness of physics courses, and this movement, while still small, is growing rapidly. The data and analysis have provided *objective* evidence, which convinced many to change the way they teach. The growth in this movement in physics has been impressive, and there are now concept tests in more advanced parts of physics, and new concept inventories in biology, astronomy, mathematics (the CCI), chemistry, and others. The results on the CCI (nearly all data so far from TL classes) match those from the FCI. The gains all cluster around 0.15 – 0.23, with one exception (described later).

Many, particularly in mathematics, are skeptical, believing that students taught with IE are less able to do standard computational problems. There is, however, much physics research that shows otherwise. Studies by Mazur [10], Redish [11], Redish & Steinberg [12], and Saul [13] have found IE students solving standard problems are no worse than those in TL courses. When he introduced *Peer Instruction*, Mazur expected — and looked for — a decline on standard problems. In *Peer Instruction* the instructor spends much less time lecturing and working examples. Still, Mazur found no difference between TL students and those using *Peer Instruction* on standard "end-of-chapter" problems. He *did* find the latter performed significantly better on tests of concept understanding. *Workshop Physics*, developed by Priscilla Laws [14] at Dickinson College, has no lectures at all and has produced similar outcomes (and the highest g values of any).

The studies in more basic mathematics (often back to elementary school) seem to show the same thing. Schoenfeld says (p. 16 of [15]):

> *"Now, more than a decade after the publication of the Standards, hard data on large-scale implementations of these curricula are beginning to come in. To briefly summarize:*
>
> *1. On tests of basic skills, there are no significant performance differences between students who learn from traditional or reform curricula.*
>
> *2. On tests of conceptual understanding and problem solving, students who learn from reform curricula consistently outperform students who learn from traditional curricula by a wide margin.*
>
> *3. There is some encouraging evidence that reform curricula can narrow the performance gap between whites and under-represented minorities."*

COGNITIVE LABORATORIES

Cognitive Laboratories [16,17] are of great help in knowing what test items are really measuring, and they were used in the validation of the CCI. Scores on items and on tests can tell a lot when properly analyzed, but it is surely true that students get right answers for wrong reasons and can get wrong answers that are at least in part the fault of the item. Cognitive labs (sometimes called "analytic interviews") are a marvelous technique to discover this. They are a highly structured interview technique where individual students are asked to think out loud as they work on a problem. Probing questions are then used to access the student mental process (not to tutor the student!). These probing questions for each item are contained in a carefully designed protocol. It is subtle to design this protocol. We utilized consultant services to do this for the CCI. Cognitive Labs are helpful on an item

with poor discrimination (good students got it wrong and/or poor students got it right), but also on a few items that perform well, to be sure that students are not getting right answers for the wrong reasons or getting wrong answers due to a problem in wording the item.

DEVELOPMENT OF THE CCI

Some more history on the CCI is needed and relevant.

It is rare that developers of tests in mathematics submit their product to scientific validation, with the exception of national tests such as the SAT. There is a huge literature and many trained professionals who validate tests. It is a subject known as "psychometrics". College faculty rarely consult such people. We think this is a mistake and have incorporated what is known about the validation of tests from the outset. A highly recommended resource is the National Research Council's *Knowing What Students Know* [18]. The process begins with the identification of the *fundamental constructs* that the test is designed to measure. One then puts together a panel of item writers with expertise in the subject matter and, hopefully, some knowledge of the principles of good assessment.

An outline of steps for developing and validating the CCI follows:

- **Test specifications:** *Decide through panel review what principles will be represented on the test.*

- **Item specifications:** *Identify precisely how to measure each concept.*

- **Item development:** *Develop a large pool of test items according to the item specifications.*

- **Item review:** *The test items are reviewed through panel discussion at a meeting. Distractors (incorrect multiple choice answers) are developed from an initial pilot test of constructed response items. Cognitive Labs are a big help here.*

- **Pilot testing and analysis:** *An item must display good 'performance characteristics' – statistical measures including the p-value, measuring the difficulty and the discrimination, measuring how a student's score for the item correlates with the score on the full test, and thus whether the item really does distinguish students who understand a concept from those who do not. Poorly performing items are discarded and, if necessary, new items are developed.*

- **Field testing and analysis:** *Preliminary versions of the full operational test are assembled. These tests are administered, and results analyzed using the same measures as in the pilot testing. Cognitive labs are very useful here to analyze items that are not performing so well, but also to check a few items that seem to be fine.*

- **Post–examination analysis:** *Once field testing is completed, an extensive statistical analysis of all of the data is performed, in particular to assess gain scores, to measure the reliability coefficient, and to (in this case) examine whether there is a correlation of gain with teaching methodology.*

We gave the first pilot test of the CCI to about 250 students at 6 schools in the spring of 2005. There was *no gain anywhere*, and scores were near the random guess level of 20 percent (even at post-test). This shocked even us. Extensive discussion among the panel led to a significant modification of the test, and to making it considerably easier. The conclusion was that if most faculty believe the test is trivial, we are probably about right. The panel developed the first field test of the CCI for the fall of 2005 for about 1,100 students in 12 U.S. universities and one in Finland. Performance was much improved. The average scores (36 percent) were now well above random guess, and there was some gain everywhere. The normalized gain g clustered very strongly between 0.15 and 0.23, essentially the same as for the FCI in TL courses. There was one small section with $g = 0.41$. We could discover no reason for this section to be different after discussion with the instructor. We suspected that either there had been teaching to the test, or that it was just a statistical outlier.

We attempted to survey instructors asking them to report on the degree of 'interactive engagement' in their classes. This showed – not surprisingly – no correlation with gain score. Instructors' own views of their interactivity are just not a satisfactory measure, and it was clear to us that all sections were predominantly lecture in any case. A set of Cognitive Labs was done with students from the fall semester early in the spring semester. These confirmed that all of the test items except one were indeed hitting the misconceptions they were designed to hit. Students were not being tripped up by confusing wording, or on some other unanticipated issue. The panel stripped out this item and one other. The latter was a question where students were being confused by the issue of whether a function that is negative and getting more negative is increasing. This left a 'final' test of 22 items. Consultant Howard Everson presented the detailed psychometric analysis, which looked pretty good. Discrimination numbers were all acceptable. There seemed to be two 'dimensions' to the exam, which correlate well internally, but not as well with each other. These were roughly: 1) 'Functions', and 2) 'Derivatives', and a smaller third dimension on limits, ratios, and the continuum. Of most interest from the psychometric point of view was the reliability coefficient, which came in at 0.7 – considered respectable, given the wide variety of testing circumstances. Professional test developers like to see 0.8, and the SAT consistently comes in around 0.85. But Dr. Everson assured us that our results were quite respectable.

The second field test has just finished, and at the time of this writing we are just getting the data analyzed. Of most interest were obviously any sections that could be viewed as clearly alternative teaching methods. We got data from Uri Treisman's group at the University of Texas, and from the *Calculus with Mathematica* group at the University of Illinois [19]. The group from Illinois came in at the same g as the lecture based groups (quite low). We suppose changing to *Mathematica* by itself is insufficient to see any change without some thought about how the computer exercises work with the student mind — but actually we do not yet have any firm handle on this. The most optimistic results were from Texas. Uri Treisman has received a MacArthur award for his extraordinary history of success in mathematics instruction. He did not expect much, he said, because he was stuck with a large class of some 85 students. Nevertheless, he came in with $g = 0.30$, which is well outside the range of all the standard lecture based sections (0.15 to 0.23), though significantly lower than what was seen in physics. Obviously, the amount of data from good alternative instruction is far too small for any final conclusions, and the foundational question of whether teaching methodology strongly affects gain (on the CCI) as it does for physics (on the FCI) will have to await further data. Significant new data will come from several IE programs this coming spring semester. In the next few months, we will publish the results and make the test available to qualified faculty via the web. Others will use it and accumulate meaningful data far faster than we could. This is what happened with the FCI.

The explosion of physics education reform arose *after* the publication of the FCI, and use of the test did in fact feed back into improved education. The dramatically improved gain scores (up to 0.70) arose over a period of 13 years between Halloun and Hestenes' publication of the original test and Hake's analysis. We expect something quite similar to happen with the CCI.

PRINCIPLES OF EVALUATION

In 1998 the National Research Council (NRC), with support from NSF, convened a Committee on the Foundations of Assessment, which produced *Knowing What Students Know* [18]. This volume assembled into useable form for both researchers and educators what was known about the science of assessment in education. The field has exploded in the past 20 years, and resources of knowledge vastly exceed what was known in 1985 when the FCI was developed. We cite central points.

1. The NRC report stresses throughout that assessment rests on three pillars: *cognition, observation*, and *interpretation*. Very briefly: 'Cognition' is a *model* of how the student represents knowledge and develops deeper understanding of concepts. 'Observation' means a method of taking data (not necessarily just pencil and paper tests). 'Interpretation' means how the researcher or educator evaluates data from observation to learn what is actually happening in the student's mind. We add a fourth leg, that quality assessment must be aligned with curriculum and teaching to have any effect.

2. Evaluation must make visible the processes happening in the student mind when faced with a task. This was done explicitly in the design of the CCI.

3. Assessment should distinguish between short-term recall and 'expert' use of concepts stored in long-term memory. Assessments claimed to test for higher level functioning should set tasks that require higher level cognition, not just recall of facts or formulas.

4. Design of assessments should consider the types of mental processes that the student is expected to use in order to demonstrate competence. The NRC says (page 62):

> *"This information is difficult to capture in traditional tests, which typically focus on how many items examinees answer correctly or incorrectly, with no information being provided about how they derive those answers or how well they understood the underlying concepts. Assessment of cognitive structures and reasoning requires more complex tasks that reveal information about thinking patterns, reasoning strategies, and growth in understanding over time."*

5. Assessment should provide information that can be used to improve instruction (our fourth leg).

> *"... most current large-scale tests provide very limited information that teachers and educational administrators can use to identify why students do not perform well.... Such tests do not reveal whether students are using misguided strategies to solve problems or fail to understand key concepts... Indeed, it is entirely possible that a student could answer certain types of test questions correctly and still lack the most basic understanding" (p. 27 [18]).*

CONCLUSION

The hardest part of making progress will be to reach traditional mathematics faculty who usually do not go to meetings, or sessions at meetings, involving mathematics education. There is no easy fix for this, since, at national meetings and most regional meetings, it is abundantly clear which are the education sessions. They are clearly delineated in conference programs, and are often actually under different sponsorship (A.A.P.T. instead of A.P.S. or M.A.A. instead of A.M.S.). There are many who will avoid like the plague the "education" type, though this is improving. There is now one advantage that some sort of improvements will actually take hold, an advantage which did not exist a generation ago. There is a large research base on what 'learning' actually means and how students actually learn. From that research base flows the current consensus among most of the relevant national organizations on the broad shape of the path that should be pursued. Thus it seems that the national organizations themselves will likely lead in trying to acquaint the greater mass of college faculty and pre-college school system personnel at least that there is a serious field of real research into education in mathematics and science. Many do not believe it (or, at the pre-college level, are frightened of it), and it is a great impediment to progress. It is our hope and expectation that a validated CCI will be a tool to aid the national organizations in this effort.

Arons [7] documented the utter irrelevance of most 'science methods' courses for teachers, and 'physical science survey' courses for liberal arts majors. It is abundantly clear as soon as you ask students to demonstrate even the most rudimentary comprehension beyond memorized words. But these courses continue to exist, revived every few years by a new generation of junior faculty who seem convinced that the form has never been tried before, or not with quite their particular style and enthusiasm. Mathematics instructors tend to believe "Oh, we would never do something as silly as that." But basic skills testing makes clear that basic mathematics courses do exactly the same thing all the time. The CCI is not a magical cure for this. But it is our hope that the results will be sufficiently difficult to ignore, and will so clearly distinguish between programs that work and those that don't, that at least some, an increasing number, will have no choice but to sit up and take notice.

REFERENCES

[1] Halloun, I. and Hestenes, D., "Common Sense Concepts About Motion," *American Journal of Physics,* vol. 53, pp. 1056–1065, 1985.

[2] Halloun, I. , "The Initial Knowledge State of College Physics Students," *American Journal of Physics,* vol. 53, pp. 1043–1055, 1985.

[3] Hake, R. R., "Interactive Engagement Methods in Introductory Mechanics Courses," available at http://www.physics.indiana.edu/~sdi, 1998.

[4] Hake, R. R., "Interactive Engagement Versus Traditional Methods: a Six–thousand Student Survey of Mechanics Test Data for Physics Courses," *Am. J. Physics*, vol. 66, pp. 64-74, available at http://www.physics.indiana.edu/~sdi, 1998.

[5] Hestenes, D., Wells, M., and Swackhammer, G., "Force Concept Inventory," *Physics Teacher*, vol. 30, pp. 141–158, 1992.

[6] Halloun, I., Hake, R.R., Mosca, E.P., and Hestenes, D., "Force Concept Inventory (Revised)," 1995, http://modeling.asu.edu/R&E/Research.html. Note: a password is needed to access the test. Follow directions on website under "Force Concept Inventory."

[7] Arons, A., "Toward a Wider Public Understanding of Science," *Am. J. Physics*, vol. 41, no. 6, pp. 769-782, 1973.

[8] Arons, A., "Addendum to Toward a Wider Public Understanding of Science," *Am. J. Physics*, vol. 42, no. 2, pp. 157-158, 1974.

[9] Hake, R. R., "Lessons from the Physics Education Reform Effort," Conservation Ecology, vol. 5, no. 2, p. 28, available at http://www.ecologyandsociety.org/vol5/iss2/art28/, 2002.

[10] Mazur, E., *Peer Instruction: A User's Manual,* (Prentice–Hall), 1997.

[11] Redish, E. F., "*Millikan Lecture*: Building a science of Teaching Physics," *Am. J. Physics*, vol. 67, pp. 562–573, 1999.

[12] Redish, E.F. Steinberg, R.N., "Teaching Physics: Figuring Out What Works," *Physics Today*, vol. 52, pp. 24–30, 1999.

[13] Saul, J. M., *Beyond Problem Solving: Evaluating Introductory Physics Courses Through the Hidden Curriculum*, doctoral dissertation, University of Maryland, 1998.

[14] Laws, P., Workshop Physics, Dickinson College, http://physics.dickinson.edu, and "Workshop Physics: Learning Introductory Physics by Doing It," *Change*, vol. 23, pp. 20-27, 1991.

[15] Schoenfeld, A.H., "Making Mathematics Work for All Children," *Educational Researcher*, vol. 31, no. 1, pp. 13-25, 2002.

[16] Garavaglia, D. R., *The Relationship Between Item Format and Cognitive Processes in Wide-Scale Assessment of Mathematics*, unpublished doctoral dissertation, 2001.

[17] Ericcson, K.A., and Simon, H., *Protocol Analysis*, (MIT Press), 496 pp., 1993.

[18] Pellegrino, J., Chudnowsky, N., and Glaser, R., (eds.), *Knowing What Students Know: The Science and Design of Educational Assessment,* (National Academy Press, Washington, D.C.), 2001.

[19] Davis, B., Porta, H., and Uhl, J., *Calculus and Mathematica,* (Addison–Wesley), 1994.

Developing And Assessing Student Scientific Abilities

Eugenia Etkina*, David T. Brookes, Sahana Murthy,
Anna Karelina, Maria Ruibal Villasenhor, and Alan Van Heuvelen
Rutgers University

OVERVIEW

In this manuscript we describe formative assessment tasks and rubrics that we developed to help students acquire and self-assess scientific process abilities. We show results of a research study of students' development of experimentation-based scientific abilities.

INTRODUCTION

Several hundred thousand science and engineering students take introductory physics courses each year. What are the goals of these courses? Instructors want to help students acquire a conceptual and quantitative understanding of major physics principles and the ability to use this understanding to solve problems. In this paper we argue that, in addition to these goals, our introductory courses and, in fact, all courses in physics, should help students develop other abilities that will be useful in their future work. According to many studies, after leaving academia, our students will be asked to solve complex problems, design experiments, and work with other people [1,2,3,4]. Several documents guiding the K–16 program design and evaluation incorporate as primary goals the development of such abilities as 1) identifying questions and concepts that guide scientific investigations; 2) designing and conducting scientific investigations; 3) using technology to improve investigations and communications; 4) formulating and revising scientific explanations and models using logic and evidence; 5) recognizing and analyzing alternative explanations and models; and 6) communicating and defending a scientific argument [5]. New accreditation requirements for engineering colleges demand that students acquire various abilities that are important in the practice of science and engineering and that are similar to the abilities listed above [6].

Today, even the most reformed introductory physics curricula do not focus explicitly on developing these abilities, and, more importantly, on assessing them. The physics education research (PER) community uses summative assessment instruments that tell us whether students mastered the concepts of Newton's laws, thermodynamics, electricity and magnetism, and so on. Physics by Inquiry, Workshop Physics, Interactive Lecture Demonstrations, and the Washington tutorials [7,8,9,10,11,12], use a formative assessment of student learning in the process of learning, but their focus is also mostly on conceptual understanding. Certain reformed curricula, such as SCALE-UP [13], have recognized some of these goals and implemented strategies to achieve them. However, in the PER community, there are no instruments that assess whether students can design and conduct investigations, communicate, defend scientific arguments, etc. In this paper we describe a National Science Foundation (NSF) - American Statistical Association (ASA) project whose goal was to develop tasks and formative assessment instruments that can be used to help achieve the "science-process" goals outlined above. We also describe some of the results of using these tasks and instruments in introductory physics courses with a curriculum designed specifically to achieve these goals, in addition to the traditional goals of a physics course. A more complete description of this work can be found in [14].

DEFINING SCIENTIFIC ABILITIES

We use the term "scientific abilities" to describe some of the most important procedures, processes, and methods that scientists use when constructing knowledge and when solving experimental problems. Using the term scientific abilities instead of science-process skills underscores that these are not automatic skills, but are instead reflectively and critically used processes [15]. The list of scientific abilities that our physics education research group developed is based on analyses of the history of the practice of physics [16,17,18], taxonomies of cognitive skills [19,20], recommendations of science educators [21], and an analysis of science-process test items [22].

This list includes the abilities:

1) To represent physical processes in multiple ways
2) To devise and test a qualitative explanation or quantitative relationship
3) To modify a qualitative explanation or quantitative relationship in light of new data
4) To design an experimental investigation
5) To collect and analyze data
6) To evaluate experimental predictions and outcomes, conceptual claims, problem solutions, and models
7) To communicate.

To help students develop these abilities, one needs to engage students in appropriate activities, to find ways to assess students' performance on these tasks, and to provide timely feedback. Activities that incorporate feedback for the students and that are used by the instructors to modify their work are called formative assessment activities. Black and Wiliam found that learning gains produced by effective use of formative assessment are larger than those found for any other educational intervention (effect sizes of 0.4 – 0.7) [23]. They also found that self-assessment during formative assessment is more powerful than instructor-provided feedback. This means that the individual, small-group, and large-group feedback systems enhance learning more than instructor-guided feedback.

For formative assessment to be effective, Sadler [24] suggested that students need to understand the target concept or ability that they are expected to acquire. They also need to understand the criteria for good work relative to that concept or ability and assess their own efforts in light of the criteria. Finally, students need to share responsibility for taking action based on the feedback. The quality of the feedback, rather than its existence or absence, is a central point. Black and Wiliam found that feedback should be descriptive and criterion-based, as opposed to numerical scoring or letter grades without clear criteria.

With all the constraints of modern teaching, including large-enrollment classes and untrained teaching assistants, how can one make formative assessment and self-assessment possible? One way to implement formative assessment and self-assessment is to use assessment rubrics containing descriptions of different levels of performance, including the target level, to help students see the learning and performance goals, self-assess their work, and modify it to achieve the goals. Instructors can use the rubric to evaluate students' responses, to provide feedback, and to modify instruction based on student work.

FINE TUNING SCIENTIFIC ABILITIES AND DEVISING RUBRICS TO ASSESS THEM

After making the list of scientific abilities that we wanted our students to develop, we started devising assessment rubrics (descriptive scoring schemes [25]) to guide their work. This activity led to a fine-tuning of the

abilities where we broke each ability into smaller sub-abilities that could be assessed. For example, for the ability to collect and analyze data we identified the following sub-abilities: 1) the ability to identify sources of experimental uncertainty, 2) the ability to evaluate how experimental uncertainties might affect the data, 3) the ability to minimize experimental uncertainty, 4) the ability to record and represent data in a meaningful way, and 5) the ability to analyze data appropriately.

Each item in the rubrics that we developed corresponded to one of the sub-abilities. We agreed on a scale of 0–3 in the scoring rubrics to describe student work (0, missing; 1, inadequate; 2, needs some improvement; and 3, adequate) and devised descriptions of student work that could merit a particular score. For example, for the sub-ability "to record and represent data in a meaningful way," a score of 0 means that the data are either missing or incomprehensible, a score of 1 means that some important data are missing, a score of 2 means that all important data are present but recorded in a way that requires some effort to comprehend, and a score of 3 means that all important data are present, organized, and recorded clearly.

While refining the list of abilities, we started devising activities that students could perform in recitations and laboratories. Defining sub-abilities and developing rubrics to assess them informed the writing of these activities. After we developed the rubrics, we started using them to score samples of student work. Our nine-person group scored students' work and revised the wording of the rubrics until we achieved 80 percent or higher agreement among our scores.

Overall we created 9 rubrics to self-assess scientific abilities, with the ability to design an experimental investigation broken into three big rubrics according to the classification described below. Some of the rubrics contained only a few items, while some contained as many as 10. The total number of described items in all rubrics is now 38. Although we validated the wording of the rubrics three years ago, as we started using them to assess student work, we continued to revise and repeatedly validate them as we scored more and more examples of student work. Presently the rubrics are posted at http://paer.rutgers.edu/scientificabilities/.

In the section below we give an example of how, after we identified one ability, we proceeded to break it down into sub-abilities and devise items in the scoring rubric to help students assess their work. We also provide an example of tasks, many of which targeted several abilities, that we gave to the students. Then we report on how we used the rubrics to study students' acquisition of this particular ability.

ABILITY TO DESIGN AN EXPERIMENTAL INVESTIGATION

One of the central abilities identified in all documents described in the introduction is that of designing an experimental investigation. For pedagogical purposes we have classified experimental investigations that students perform in introductory courses into three broad categories [26]: 1) Observational experiments, 2) Testing experiments, and 3) Application experiments.

When conducting an observational experiment, a student focuses on investigating a physical phenomenon without having expectations of its outcomes. When conducting a testing experiment, a student has an expectation of its outcome based on concepts constructed from prior experiences. In an application experiment, a student uses established concepts or relationships to address practical problems. In the practice of physics there is a great deal of overlap among the experiments. However, we find it very useful to make distinctions

for educational purposes. The distinctions between the experiments relates to the procedures. When students perform an observational experiment they do not make an explicit prediction of its outcome, while for a testing experiment they are required to make a prediction and state explicitly which model or hypothesis they used to make this prediction. Often students test their own hypotheses, but sometimes they test hypotheses proposed in the lab write up. These usually are based on research on student learning difficulties. In the process of scientific research the same experiment can fall into more than one of these categories.

What abilities do students need when designing these investigations? We have identified the following steps that students need to take to design, execute, and make sense out of a particular experimental investigation. We assigned a sub-ability for each step and wrote corresponding descriptors in the rubrics. Students do not follow the steps listed in the table in the chronological order. Instead, we used them as a guidance to write the lab tasks and the rubrics. The results of these discussions are presented in Table 1.

Table 1. Sub-abilities involved in designing three different types of experimental investigation.

Sub-abilities involved in the ability to design and conduct an observational experiment	Sub-abilities involved in the ability to design and conduct a testing experiment	Sub-abilities involved in the ability to design and conduct an application experiment
Identifying the phenomenon to be investigated	Identifying the hypothesis or a relation to be tested	Identifying the problem to be solved
Designing a reliable experiment that investigates the phenomenon	Designing a reliable experiment that tests the hypothesis or the relation using the available equipment	Designing an experiment to solve the problem
	Making a prediction of the outcome of the experiment based on the hypothesis or relation under test	Devising a mathematical procedure to solve the problem
	Identifying additional assumptions used to make the prediction	Identifying assumptions in the mathematical procedure
Deciding what is to be measured, and identifying independent and dependent variables	Deciding what is to be measured, and identifying independent and dependent variables	Deciding what is to be measured
Using available equipment to make measurements	Using available equipment to make measurements	Using available equipment to make measurements
Describing what is observed	Deciding whether the outcome of the experiment matched	Making a judgment about the results of the experiment
Describing a pattern or devising an explanation	Making a reasonable judgment about the hypothesis or relation	Evaluating the results by means of an independent method
Identifying shortcomings in an experimental design and suggesting specific improvements	Identifying shortcomings in an experimental design and suggesting specific improvements	Identifying shortcomings in an experimental design and suggesting specific improvements

For each of the identified sub-abilities, we devised a rubric item that describes different levels of proficiency. Examples of selected rubric items for the ability to design an application experiment are given in Table 2.

Students use these rubrics in the laboratories. Each lab focuses on a limited number of rubrics that students need to use for self-assessment (see the example below). The rubrics guide them as to what experimental aspects they should specifically pay attention to. After they perform the experiment, they write a lab report (in

the lab). After the report is written, they use the rubrics to assign scores for particular sub-abilities chosen for this lab. Then they see what scores need to be improved, and revise the report. At the end of the lab they hand in their report to the instructor who assesses them based on the same rubrics. At the beginning of the next lab the instructor discusses with the students the discrepancies in assessment and ways to improve. These can be whole section discussions or individual group discussions depending on the nature of student difficulties. Below is an example of a lab whose write up guides students through the development and use of sub-abilities involved in the application experiments and relevant rubric items that students use for self assessment and instructors use for grading. More laboratory investigations with the observational and testing experiments are posted at http://paer.rutgers.edu/scientificabilities/ .

Application Experiment Example: Specific heat of unknown object

Design two independent experiments to determine the specific heat of the given object. The material of which the object is made is not known.

Self-assessment - After you finish your lab report use rubrics D2, D4, D7, D8, AND D9 to self-assess your work, then revise the report if necessary.

Experiment - Equipment: You have access to the following equipment: water, ice, container for water, hot plate, calorimeter, balance, and digital thermometer.

1. First, recall (you don't have to write anything!) why it is important to design two experiments to determine a quantity.
2. Play with the equipment to find how you can use it to achieve the goal of the experiment. Devise as many designs as possible. Working with your lab partners, choose the best two designs. Indicate the criteria that you used to decide which designs were the "best". Show it to your lab instructor.

For each method in your lab report:

1. Write a verbal description and draw a labeled sketch of the design you chose. Include the quantities you will measure.
2. Write the mathematical procedure you will use.
3. List all assumptions you make in your design and procedure. Decide which assumption affects your results most. Explain how the outcome of the experiment depends on this assumption, i.e. if the assumption is not correct, your result is larger or smaller than the "real" value.
4. Describe an additional experiment to determine whether the main assumption is valid in your experiment. Estimate quantitatively the effect of this assumption on the value of your measurement and compare it with the instrumental uncertainty (what will happen to your measurements if the assumption is not valid).
5. List sources of experimental uncertainty. Decide which is the largest source of uncertainty. Use the weakest link rule to estimate the uncertainty in your result. How would you minimize uncertainties?
6. Perform the experiment and record the data. Do not forget to minimize experimental uncertainties and the effect of the assumptions.
7. Calculate the specific heat including experimental uncertainty.
8. Compare the outcomes of two experiments after you have done both. Discuss if they are within your experimental uncertainty of each other. If not, specifically explain what might have gone wrong – perhaps one of your assumptions was not valid. If your experiments are not close to each other within experimental uncertainty, perform the experiment again taking steps to improve your design.

Table 2. Rubric to assess students' ability to design and conduct an application experiment

Ability to design and conduct an application experiment					
Scientific Ability		**Missing (0)**	**Inadequate (1)**	**Needs some improvement (2)**	**Adequate (3)**
1	Is able to identify the problem to be solved	No mention is made of the problem to be solved.	An attempt is made to identify the problem to be solved but it is described in a confusing manner.	The problem to be solved is described, but there are minor omissions or vague details.	The problem to be solved is clearly stated.
2	Is able to design a reliable experiment that solves the problem	The experiment does not solve the problem.	The experiment attempts to solve the problem, but due to the nature of the design the data will not lead to a reliable solution.	The experiment attempts to solve the problem but due to the nature of the design there is a moderate chance the data will not lead to a reliable solution.	The experiment solves the problem and has a high likelihood of producing data that will lead to a reliable solution.
3	Is able to use available equipment to make measurements	At least one of the chosen measurements cannot be made with the available equipment.	All of the chosen measurements can be made, but no details are given about how it is done.	All of the chosen measurements can be made, but the details about how they are done are vague or incomplete.	All of the chosen measurements can be made and all details about how they are done are provided and clear.
4	Is able to make a judgment about the results of the experiment	No discussion is presented about the results of the experiment	A judgment is made about the results, but it is not reasonable or coherent.	An acceptable judgment is made about the result, but the reasoning is flawed or incomplete.	An acceptable judgment is made about the result, with clear reasoning. The effects of assumptions and experimental uncertainties are considered.

5	Is able to evaluate the results by means of an independent method	No attempt is made to evaluate the consistency of the result using an independent method.	A second independent method is used to evaluate the results.	A second independent method is used to evaluate the results. The results of the two methods are compared using experimental uncertainties.	A second independent method is used to evaluate the results and the evaluation is done with the experimental uncertainties. The discrepancy between the results of the two methods, and possible reasons, are discussed. A percentage difference is calculated in quantitative problems.
6	Is able to identify the shortcomings in an experimental design and suggest specific improvements	No attempt is made to identify any shortcomings of the experimental design.	An attempt is made to identify shortcomings, but they are described vaguely and no specific suggestions for improvements are made.	Some shortcomings are identified and some improvements are suggested, but not all aspects of the design are considered.	All major shortcomings of the experiment are identified and specific suggestions for improvement are made.
7	Is able to choose a productive mathematical procedure for solving the experimental problem	Mathematical procedure is either missing, or the equations written down are irrelevant to the design.	A mathematical procedure is described, but is incorrect or incomplete, due to which the final answer cannot be calculated.	Correct and complete mathematical procedure is described, but an error is made in the calculations.	Mathematical procedure is fully consistent with the design. All quantities are calculated correctly. Final answer is meaningful.
8	Is able to identify the assumptions made in using the mathematical procedure	No attempt is made to identify any assumptions.	An attempt is made to identify assumptions, but the assumptions are irrelevant.	Relevant assumptions are identified but are not significant for solving the problem.	All assumptions are correctly identified.
9	Is able to determine specifically the way in which assumptions might affect the results	No attempt is made to determine the effects of assumptions.	The effects of assumptions are mentioned but are described vaguely.	The effects of assumptions are determined, but no attempt is made to validate them.	The effects of the assumptions are determined and the assumptions are validated.

DO STUDENTS DEVELOP SCIENTIFIC ABILITIES?

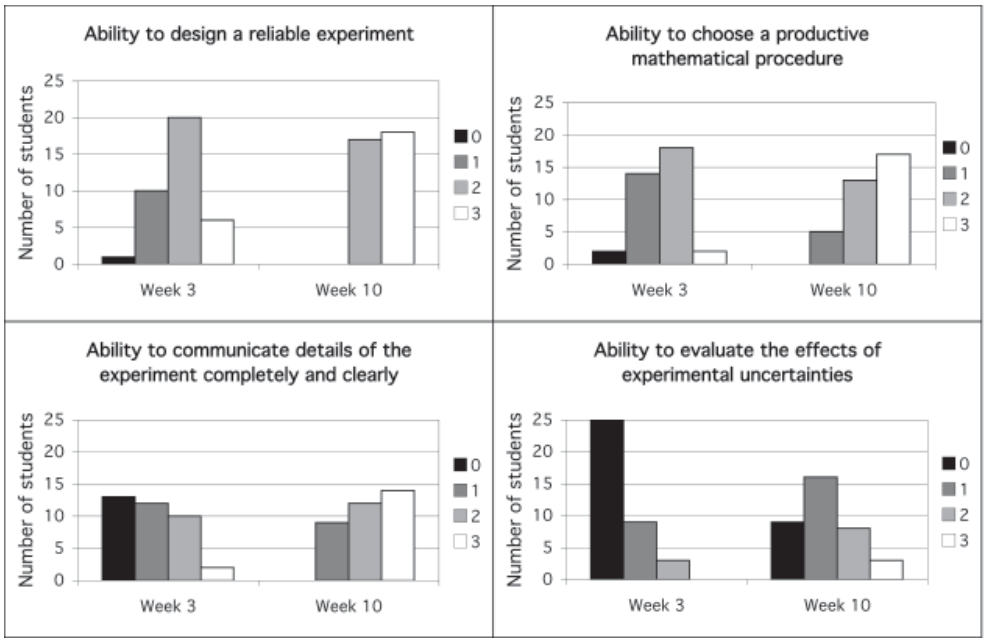
In this section, we briefly describe one research project that investigated whether students acquire experimental abilities. Other projects conducted as a part of the grant are described in [14,27,28,29,30,31,32,33,34,35].

THE STUDY OF EXPERIMENTATION ABILITIES

We conducted two studies of experimentation abilities, both in large-enrollment introductory physics courses (a 500-student course in 2003 and a 180-student course in 2005) for science majors (premed, prevet, biology, chemistry, exercise science, environmental science, etc.). Both courses had an integrated lab component, and focused on the development of experimentation-based scientific abilities. The courses used the same curriculum materials and had the same labs–eleven 3-hour sessions per semester. In each lab, students had to design one experiment from scratch, similar to the example above, and devise a procedure for the other experiment with the given equipment. The experiments had guidelines that focused on different scientific abilities that we had identified.

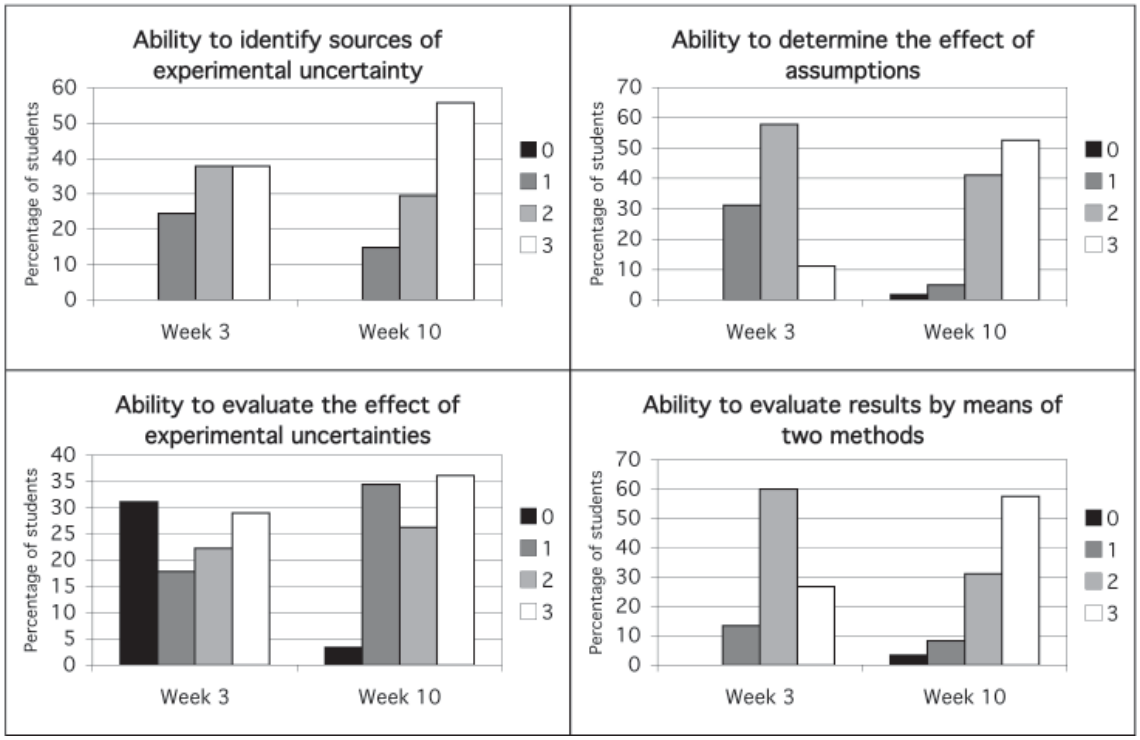
To measure the development of students' scientific abilities, we scored their lab reports each week based on the scientific abilities rubrics. In the first study conducted in 2003, we focused on the following abilities: 1) designing a reliable experiment to solve a problem, 2) choosing a productive mathematical procedure, 3) communicating details of the experiment, and 4) evaluating the effects of experimental uncertainties and theoretical assumptions. Our sample consisted of 35 randomly chosen students who were distributed among four lab sections. We found that the students' abilities to design an experiment, to devise a mathematical procedure to solve an experimental problem, and to communicate the details of the procedure performance improved significantly [27,28]. We used chi square statistics to determine that the improvement was statistically significant ($p < 0.01$) for these abilities. However, the changes in students' abilities to evaluate experimental uncertainties and assumptions were not significant [36]. Our results are shown in Figure 1. In prior studies [37,38], the ability to evaluate the effects of experimental uncertainties were found to be difficult for students.

Figure 1. Comparison of students' scientific abilities between week 3 and week 10 in 2003. Bars indicate the number of students whose report was scored as 0 (missing), 1 (inadequate), 2 (needs improvement), or 3 (adequate) on a particular sub-rubric item.



Based on the results from 2003 we then revised the lab write-ups to emphasize these abilities. We added specific exercises related to experimental uncertainties and theoretical assumptions, and made the guidelines more explicit (as in the example above). After these revisions we repeated the study in 2005 with a sample of 61 students from 3 lab sections. We noticed significant improvements ($p<0.005$ using the chi square test) in these "difficult" abilities (see Figure 2).

Figure 2. Comparison of students' scientific abilities between week 3 and week 10 in 2005. Bars indicate the percentage of students in the sample whose report was scored as 0 (missing), 1 (inadequate), 2 (needs improvement), or 3 (adequate) on a particular sub-rubric item.



SUMMARY

This paper has described the development of tasks and assessment rubrics that help students acquire some abilities useful in science and engineering. Our approach to learning introductory college physics emphasizes the acquisition of various scientific abilities as one of the main course goals. Development of these abilities is considered important in science courses at K–16 levels [22,39,40,41,42,43]. There are summative assessment questions developed for national and international evaluations of student learning of some scientific abilities, but we are unaware of any systematic efforts to build a library of formative assessment tasks and rubrics to help students develop these abilities.

In order to help students acquire these abilities, we have developed a large number of activities used formatively during instruction. In addition, we have developed rubrics, which indicate what is needed for proficiency relative to the different sub-abilities. The rubrics can also be used by instructors to provide formative feedback to the students or for summative evaluation of the students and the learning system. Students can also use the rubrics for self-evaluation as they perform the activities. Students need to be able to revise their work after they have evaluated it using the rubrics. Here, the instructor's help is essential, as the rubrics themselves do not

provide content-related feedback. The rubrics also can be used for research purposes to monitor students' progress and to compare students from different courses. They can add to our library of assessment instruments that allow the PER community to evaluate learning. So far, most of the PER-developed instruments assess conceptual understanding and graphing skills.

When we used the rubrics as a research tool to study whether students' scientific abilities in introductory physics courses improved, the results were positive. In this paper we have described significant changes in students' experimentation abilities. Other studies, described elsewhere, show that students improve their ability to represent knowledge in multiple ways, and to evaluate problem-solving approaches [31,32,35]. Thus, we have evidence supporting the argument that if a course explicitly focuses on helping students develop scientific abilities, in addition to the content knowledge, students improve!

Finally, we have evaluated student traditional problem-solving performance in this learning system that emphasizes the development of science process abilities. For this evaluation, we used student work in the course where the project was carried out (2003/2004) and in the same course when taught traditionally (2000/2003). For the final exam, the professor who taught the course in 2000/2003 selected 12 difficult multiple-choice problems that were used on the final exams in prior years and for which he had the data on student performance. The professor for the reformed course used 8 of these problems on his final exam. Prior to the project the students' average score on these problems was 58 percent; the average score in the course emphasizing the scientific abilities was 73 percent. Using the standard deviations for both groups we found the effect size to be above 2, which makes the difference statistically significant [44]. Assuming that the population in the course did not change, we can say that the reformed curriculum does not harm students' problem solving abilities.

In summary, it seems that it is possible to help students in introductory physics courses start acquiring some of the science process abilities that are needed for work in the 21st century workplace. The learning system also enhances student performance in terms of traditional measures. Developed rubrics are not physics specific and can be used in all college science courses to help students acquire abilities that they will need in their future professional lives. The rubrics can be also used for research purposes to collect data about student progress and for comparison of student learning in different courses.

ACKNOWLEDGEMENTS

The authors thank Suzanne Brahmia, David Rosengrant, Aaron Warren, Michael Gentile, Michael Lawrence, Marina Milner-Bolotin, Hector Lopez, and Juliana Timofeeva for their help in the development of rubrics and formative assessment tasks. We thank Richard Fiorillo and Gabriel Alba for their help in implementing the laboratories. We also thank the National Science Foundation (DUE-0241078, DUE-0088906, and REC 0529065) for providing us with the funding for the project.

REFERENCES

[1] Czujko, R., "The physics bachelors as a passport to the workplace: Recent research results," in Redish, E.F., and Rigden, J.S., (eds.), *The Changing Role of Physics Departments in Modern Universities*, AIP Conference Proceedings,(Woodbury, NY), vol. 399, pp. 213 - 224, 1997.

[2] "What Work Requires of Schools: A SCANS Report for America 2000," (U.S. Department of Labor, Secretary's Commission on Achieving Necessary Skills, Washington D. C.), 1991.

[3] Bransford, J.D., Brown, A.L., and Cocking, R. R., *How people Learn: Brain, mind, experience, and school*, (National Academy Press, Washington, D.C.), 1999.

[4] Carneval, A., Gainer, L., and Meltzer, A., *Workplace Basics: The Essential Skills Employers Want*, (Jossey-Bass, San Francisco), 1990.

[5] National Research Council, *National Science Education Standards,* (National Academy Press, Washington, D.C.), 1996.

[6] See the ABET home page at URL http://www.abet.org/forms.shtml

[7] McDermott, L.C., and the Physics Education Group at the University of Washington, *Physics by Inquiry* (Wiley, New York), 1996.

[8] Redish, E.F., *Teaching Physics with the Physics Suite*, (Wiley, New York), 2003.

[9] Sokoloff, D.R., Thornton, R.K., and Laws, P.W., *RealTime Physics,* (Wiley, New York), 2004.

[10] Sokoloff, D.R., and Thornton, R.K., *Interactive Lecture Demonstrations*, (Wiley, New York), 2004.

[11] Laws, P.W., *Workshop Physics Activity Guide*, 2nd ed., (Wiley, New York), 2004.

[12] McDermott, L.C., and Shaffer, P.S., *Tutorials in Introductory Physics*, (Prentice Hall, New York), 1998.

[13] See the SCALE-UP project webpage at URL http://www.ncsu.edu/per/scaleup.html

[14] Etkina, E., Van Heuvelen, A., White-Brahmia, S., Brookes, D.T., Gentile, M., Murthy, S., Rosengrant, D., and Warren, A., "Scientific abilities and their assessment," Phys. Rev. ST Phys. Educ. Res., vol. 2, no. 2, article ID# 020103, 2006.

[15] Salomon, G., and Perkins, D.N., "Rocky road to transfer: Rethinking mechanisms of a neglected phenomenon," *Ed. Psychologist*, vol. 24, no. 2, pp. 113 – 142, 1989.

[16] Holton, S., and Brush, S., *Physics, The Human Adventure*, (Rutgers University Press, New Brunswick, N.J.), 2001.

[17] Lawson, A.E., "The generality of hypothetico-deductive reasoning: Making scientific thinking explicit," *Am. Bio. Teach*., vol. 62, no. 7, pp. 482 – 495, 2000.

[18] Lawson, A.E., "The nature and development of hypothetico-predictive argumentation with implications for science teaching," *Int. J. Sci. Ed*., vol. 25, no. 11, pp. 1387 – 1408, 2003.

[19] Bloom, B.S., *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I Cognitive Domain,* (David McKay Co., Inc., New York), 1956.

[20] Dressel, P., and Mayer, L., *General education: Exploration in evaluation, Final Report of the Cooperative Study of Evaluation in General Education*, (Council on Education, Washington, D.C.), 1954.

[21] Schunn, C.D., and Anderson, J., "Acquiring expertise in science: Exploration of what, when and how," in Crowley, K., Schunn, C.D., and Okada, T., (eds.), *Designing for science: Implications from everyday, classroom and professional settings*, (Erlbaum, Mahwah, N.J.), pp. 351 – 392, 2001.

[22] Gonzales, P., Calsyn, C., Jocelyn, L., Mak, K., Kastberg, D., Arafeh, S., Williams, T., and Tsen, W., "Pursuing excellence: Comparisons of international eighth-grade mathematics and science achievement from a U.S. perspective, 1995 and 1999," *Statistical Analysis Report NCES 2001-028*, U. S. Department of Education, National Center for Education Statistics, May 2001.

[23] Black, P., and Wiliam, D., "Inside the black box: Raising standards through classroom assessment," *Phi Delta Kappan,* vol. 80, no. 2, pp. 139 – 148, 1998.

[24] Sadler, D.R., "Formative assessment and the design of instructional systems," *Instr. Sci*., vol. 18, pp. 119 – 144, 1989.

[25] Brookhart, S.M., "The Art and Science of Classroom Assessment: The Missing Part of Pedagogy," *ASHE-ERIC Higher Education Report*, (The George Washington University, Graduate School of Education and Human Development, Washington, D.C.), vol. 27, no.1, 1999.

[26] Etkina, E., Van Heuvelen, A., Brookes, D.T., and Mills, D., "Role of experiments in physics instruction — a process approach," *Phys. Teach*., vol. 40, no. 6, pp. 351 – 355, 2002.

[27] Murthy, S., and Etkina, E., "Developing scientific abilities in a large class," in Marx, J., Heron, P., and Franklin, S., (eds.), *Proceedings of the 2004 Physics Education Research Conference,* (American Institute of Physics, Melville, N.Y.), vol. 790, pp. 133 – 136, 2005.

[28] Etkina, E., Murthy, S., and Zou, X., "Using introductory labs to engage students in experimental design," *Am. J. Phys*., vol. 74, no. 11, pp. 979 – 986, 2006.

[29] Etkina, E., Warren, A.A., and Gentile, M., "The role of models in physics instruction," *Phys. Teach*., vol. 44, no. 1, pp. 34 – 39, 2006.

[30] Etkina, E., and Murthy, S., "Design labs: Students' expectations and reality," in Heron, P., McCullough, L., and Marx, J., (eds.), *Proceedings of the 2005 Physics Education Research Conference* (American Institute of Physics, Melville, N.Y.), vol. 818, pp. 97 – 100, 2006

[31] Rosengrant, D., Etkina, E., and Van Heuvelen, A., "Free-body diagrams: Necessary or sufficient?" in Marx, J., Heron, P., and Franklin, S., (eds.), *Proceedings of the 2004 Physics Education Research Conference*, (American Institute of Physics, Melville, N.Y.), vol. 790, pp. 177 – 181, 2005.

[32] Rosengrant, D., Etkina, E., and Van Heuvelen, A., "Case study: Students use of multiple representations in problem solving," in Heron, P., McCullough, L., and Marx, J., (eds.), *Proceedings of the 2005 Physics Education Research Conference,* (American Institute of Physics, Melville, N.Y.), vol. 818, pp. 49 – 53, 2006.

[33] Etkina, E., Karelina, A., and Ruibal Villasenhor, M., "Studying transfer of scientific reasoning abilities," in *Proceedings of the 2006 Physics Education Research Conference,* (in press).

[34] Karelina, A., and Etkina, E., "When and how do students engage in sense-making in a physics lab," in *Proceedings of the 2006 Physics Education Research Conference,* (in press).

[35] Warren, A., "Evaluation as means for learning physics," *Phys. Rev. ST Phys. Educ. Res.*, (in press).

[36] Murthy, S., and Etkina, E., "Using rubrics as self-assessment instruments for students in introductory laboratories," AAPT National Meeting, Albuquerque, N.M., January, 2005.

[37] Buffler, A., Allie, S., Lubben, F., and Campbell, B., "The development of first-year physics students' ideas about measurement in terms of point and set paradigms," *Int. J. Sci. Ed.*, vol. 23, no. 11, pp. 1137 – 1156, 2001.

[38] Lippmann, R., "Students' understanding of measurement and uncertainty in the physics laboratory: social construction, underlying concepts, and quantitative analysis," Ph.D. thesis, University of Maryland, 2003.

[39] Kruglak, H.,"Experimental outcomes of laboratory instruction in elementary college physics," *Am. J. Phys.*, vol. 19, no. 3, pp. 136 – 141, 1952.

[40] Kruglak, H.,"The measurement of laboratory achievement," *Am. J. Phys.*, vol. 22, no. 7, pp. 442 – 451, 1954.

[41] Kruglak, H., "The measurement of laboratory achievement, part ii, paper-pencil laboratory achievement tests," *Am. J. Phys.*, vol. 22, no. 7, pp. 452 – 462, 1954.

[42] Kruglak, H., "The measurement of laboratory achievement, part iii, paper-pencil analogs of laboratory performance tests," *Am. J. Phys.*, vol. 23, no. 2, pp. 82 – 87, 1955.

[43] National Assessment Governing Board, "Science Framework for the 2005 National Assessment of Education Progress," (U.S. Department of Education, Washington, D.C.), 2004.

[44] Study reported in: Van Heuvelen, A.,"Implementing a science process approach in large introductory physics courses," talk given at Rutgers University, November, 2004.

The Design and Validation of the E-NSSE and E-FSSE Surveys of Student Engagement in Engineering

Malcolm Drewery, Stefani Bjorklund and Norman Fortenberry*
National Academy of Engineering

OVERVIEW

The Engineering National Survey of Student Engagement (E-NSSE), and its faculty version, the Engineering Faculty Survey of Student Engagement (E-FSSE), are two new instruments designed to measure adherence to "best instructional practices" in engineering education, and engineering graduates' achievement of certain desired learning outcomes. The questions within the two surveys have been designed to focus strongly on student and faculty engagement in the learning process and to assess their views of the success of the learning process at producing specific engineering education learning outcomes. These instruments should not be confused with others that seek to delve into student understanding of the scientific approach or key concepts within a discipline.

This paper introduces the E-NSSE and E-FSSE and their design phases, and provides preliminary analysis in the validation process of the E-FSSE survey, which includes analyzing results from three engineering departments, and validity and reliability tests. Methodology used to determine categories and how to analyze the strength of categories for probing various facets of student learning outcome and teacher instructional principles are described. This paper serves as the foundation for the results, conclusions, and next steps from the analysis of our survey data.

INTRODUCTION

In the last several years, we have seen an influx of articles, dialogue, and meetings of engineering educators looking for ways to improve engineering education by introducing and strengthening their commitment to assessing specific approaches to teaching, learning, and student learning outcomes. Engineering instruction is framed within the fundamental nature of engineering. Unlike science, which seeks fundamental understanding of natural processes, engineering seeks to apply scientific/mathematical/technological principles, within multiple constraints, to the creation of new products, processes, and services. Engineering is also concerned with understanding and optimizing the performance of new and existing products, processes, and services.

Thus, a key focus of engineering education is developing students with the knowledge, skills and abilities to perform these tasks. The report, *The Engineer of 2020: Visions of Engineering in the New Century* [1], offers one exposition of the attributes and abilities engineers will need to perform well in a world driven by rapid technological advancements, national security needs, aging infrastructure in developed countries, environmental challenges brought about by population growth and diminishing resources, and the creation of new disciplines at the interfaces between engineering and science. To ensure that future engineers have these capabilities, they must be educated to be not only technically proficient, but also ethically grounded global citizens who can become leaders in business and public service.

IMPORTANCE OF STUDY

More recently, educators have been trying to improve engineering education by introducing and strengthening their commitment to assessing specific approaches to teaching, learning, and student learning outcomes. In their article, "Assessment in Engineering Education: Evolution, Approaches, and Future Collaboration", Olds, Moskal, and Miller describe the current movement toward the assessment of student learning outcomes within the engineering community, and assert that, as recently as 1997, the engineering community had relatively little experience in conducting outcomes assessment [2]. Further, Bjorklund and Fortenberry assert that while researchers and educators have developed a number of classroom and college-wide assessments – oftentimes in preparation for an Accreditation Board for Engineering and Technology (ABET) visit – no national assessments exist to measure engineering student learning outcomes and the instructional practices that support those outcomes [3].

In response, the Center for the Advancement of Scholarship on Engineering Education (CASEE) has developed two surveys to assess the extent to which engineering faculty are engaging in identified "best instructional practices" and engineering students are achieving certain desired learning outcomes. The questions within the two surveys have been designed to focus strongly on student and faculty engagement in the learning process and to assess their views of the success of the learning process with respect to attainment of specific learning outcomes of engineering education. CASEE is in the initial stages of validating the faculty version of the instrument (E-FSSE) at nine universities across the country. This paper briefly describes the survey design in Phases I through V and provides some preliminary analysis of the validation process and the next steps in the project in Phase VI.

MEASURING STUDENT AND FACULTY ENGAGEMENT IN ENGINEERING EDUCATION

Phase I: Identifying Desired Student Learning Outcomes

ABET's "3a through k" criteria identified 11 learning outcomes expected of engineering graduates. Based on a rigorous review of the literature, the first phase of our work revealed four additional student outcomes desired by the engineering education community. Although many more outcomes were mentioned in the literature, each of the four learning outcomes was cited at least 16 times, which was also the number of times the least-cited ABET criterion was referenced in the same body of literature [4]. The following list illustrates the 15 foundational technical and nontechnical learning outcomes, ABET's 3a through k and CASEE's "plus four," on which we focus our work [5].

Engineering graduates must have:

ABET Criteria 3a through k

a) An ability to apply knowledge of mathematics, science, and engineering
b) An ability to design and conduct experiments, as well as to analyze and interpret data
c) An ability to design a system, component, or process, to meet desired needs
d) An ability to function on multi-disciplinary teams
e) An ability to identify, formulate, and solve engineering problems
f) An understanding of professional and ethical responsibility
g) An ability to communicate effectively
h) The broad education necessary to understand the impact of engineering solutions in a global and societal context
i) A recognition of the need for, and an ability to engage in, life-long learning

j) A knowledge of contemporary issues
k) An ability to use the techniques, skills, and modern engineering tools necessary for engineering practice

CASEE's "plus four"

l) An ability to manage a project (including a familiarity with business, market-related, and financial matters)
m) A multidisciplinary systems perspective
n) An understanding of, and appreciation for, the diversity of students, faculty, staff, colleagues, and customers
o) A strong work ethic

Phase II: Identifying Principles of Effective Teaching and Learning

During a literature search for the best instructional practices thought to contribute to the 15 student learning outcomes identified in Phase I, ten principles of effective instruction repeatedly emerged. Although a number of authors used different wording for similar concepts, we used three sources as the primary tools for organizing the ten principles and use their "language" in listing them. The three sources: Chickering and Gamson [6], Bransford et al. [7], and the American Psychological Association's "learner-centered psychological principles" [8], were chosen based on their prominence in engineering education circles, engineering educators' existing familiarity with these concepts, and, of course, their appropriateness. The ten principles of effective instruction include:

1) Encouraging student-faculty interaction
2) Developing reciprocity and cooperation among students
3) Communicating high expectations
4) Providing prompt feedback
5) Using active learning techniques
6) Emphasizing time on task
7) Respecting diverse talents and ways of thinking
8) Building on correct pre-existing understandings; dispelling false preconceptions
9) Providing factual knowledge, facilitating understanding of the facts and ideas in the context of a conceptual framework, and organizing knowledge that facilitates retrieval and application
10) Encouraging students' motivation to learn

Phase III: Linking Instructional Principles to Learning Outcomes

While in Phase I we discovered desired student learning outcomes, and in Phase II we identified instructional principles and practices believed to promote those learning outcomes. The third phase of the project endeavored to illustrate which of the instructional principles, based on a review of the literature, are related to specific student outcomes (see Table 1).

The first column of Table 1 lists the identified instructional principles. Columns two and three list student outcomes (ABET 3a through k and CASEE's "plus 4") associated with each instructional practice. More specifically, the second column lists outcomes, shown in the literature, that result from students engaging in teaching and

learning strategies associated with each instructional principle. Column three lists outcomes we *expected* to result from students engaging in teaching and learning strategies associated with each instructional principle; however, no evidence was found in our prior research to substantiate these expectations.

Evident in Table 1, many opportunities exist for scholars to conduct novel and needed research examining the extent to which these principles and outcomes are related, and exactly which instructional best practices contribute to students achieving each of the identified outcomes.

Phase IV: Developing Survey Items

Using the information gathered in Phases I through III, and by adapting some of the items used in the National Survey of Student Engagement (NSSE) [9], the Faculty Survey of Student Engagement (FSSE) [10], and the EC2000 Study [11] instruments, we developed two survey instruments (faculty and student versions). Because single-item measures often fail to capture the nuances and complexities of such concepts, we developed constructs – each consisting of two to six items – for each of the 15 student learning outcomes and 10 instructional practices. CASEE's surveys also include items regarding demographic information.

Table 1. Instructional Principles' Relationships to Student Outcomes

Instructional Principles		Related Student Outcomes (ABET 3a-k) plus four	
		Evidence found in engineering education research	Expected relationship
1.	Encouraging student-faculty interaction	3b, 3c, 3f	All
2.	Developing reciprocity and cooperation among students	3g, L	3d, 3g, L, N
3.	Communicating high expectations		All
4.	Providing prompt feedback	3b, 3c, 3f	All
5.	Using active learning techniques	3b, 3c, 3d, 3e, 3g, 3k	All (except 3j)
6.	Emphasizing time on task		All
7.	Respecting diverse talents and ways of thinking	3a, 3b, N	All (especially N)
8.	Building on correct pre-existing understandings; dispelling false preconceptions	3a	All (especially those related to design)
9.	Providing factual knowledge, facilitating understanding of the facts and ideas in the context of a conceptual framework, and organizing knowledge that facilitates retrieval and application		All (especially those related to design)
10.	Encouraging students' motivation to learn		All (especially 3i)

The CASEE instruments differ from the NSSE and FSSE instruments in that their content focuses specifically on engineering education. They are distinguished from the EC2000 instruments as the EC2000 instruments concentrate only on the 11 EC2000 3a-k student outcomes criteria and a substantially smaller set of instructional practices.

Phase V: Focus Groups and Item Refinement

In the preliminary work for this project (Phases I through IV), we took great care in crafting clear, well-defined survey items with high face and content validity. We based the items on the scholarship of survey

design and engineering education and, with the authors' permission, items used in the NSSE, FSSE, and EC2000 instruments. The NSSE instrument development team established, for example, through the use of focus groups, that each survey item has the same meaning for each respondent [13].

Similarly, we conducted two 90-minute focus groups (one with faculty and one with students) at each of five engineering colleges in February and March 2005 to establish that each of our survey items meant the same thing to each reader. Focus group participants received a set of survey items to review prior to participating in the focus groups. During the focus groups, participants 1) discussed the meaning of each item to ensure that every reader interpreted the item in the same way, and 2) suggested additional items and alternative ways to word certain items. Refining the items was an iterative process. CASEE staff refined the items as suggested by focus group participants between visits at each campus. There was a great deal of discussion in the first few focus groups and, as the items were refined, subsequent focus group participants believed the items were clear and relevant to the instruments' intent. Final drafts of the instruments were completed at the end of June 2005. The current iteration of the survey drafts are approximately seven pages each and include items regarding demographic information and the student outcomes and teaching practices. Table 2 lists the nine colleges of engineering that have agreed to participate in the pilot administrations of the CASEE faculty and student questionnaires.

Phase VI: Large-scale Validation

Sampling

In October 2006, email messages were sent to contact persons in the dean's office in each of the nine engineering colleges that initially agreed to participate in the pilot administration of the survey. These institutions were selected because of their leadership and interest in the field of engineering education, their geographic diversity, and their willingness to administer the surveys college wide. Five large, doctoral-granting research universities were selected to maximize the number of potential responses for this pilot of the questionnaire. Two primarily masters degree-granting universities were invited to participate to ensure the participation of a variety of institution types. For the same reason, a specialized and primarily baccalaureate degree-granting institution was included.

Table 2. Institutions Participating in this Study

Institution	Public/Private	Region	Focus	MSI*
Cal State LA	Public	West	Comprehensive	Yes
FAMU/FSU	Public	South	Comprehensive/Research	Yes
Georgia Tech	Public	South	Research	No
Montana State	Public	West	Research	No
Penn State	Public	East	Research	No
Purdue	Public	Mid-west	Research	No
University of Wisconsin	Public	Mid-west	Research	No
Rose-Hulman	Private	Mid-west	Baccalaureate	No
Rowan	Private	East	Comprehensive	No

*MSI = minority serving institution

As each participating engineering college receives its Institutional Review Board (IRB) approval, email messages are sent to faculty that include a letter from the office of the dean of the college describing the study, ask

that individuals complete the anonymous survey, and provide a link to the online survey. The dean's office in each college sends out a follow-up email to all engineering faculty one to two weeks after the first solicitation in order to increase the number of survey respondents. Of the nine participating colleges, three have received IRB approval thus far and have started filling out the E-FSSE. The preliminary data collected from those three schools are being used in this report.

The faculty sample includes all engineering faculty at participating institutions. "Engineering faculty" includes all adjunct, assistant, associate, and full professors who teach at least one undergraduate course during the academic year. We are focusing on faculty with teaching responsibilities, as many survey items ask about teaching practices. Faculty members with "research only" appointments are not well suited to answer many of the questions posed in the survey. Table 3 illustrates the total number of faculty at each institution that will be asked to complete the survey.

Table 3. Sample population by Location

Engineering College	Faculty*
Cal State LA	42
FAMU/FSU	100
Georgia Tech	418
Montana State	70
Penn State	384
Purdue	321
University of Wisconsin	200
Rose-Hulman	82
Rowan	34
Total	1,651

*Preliminary estimate of total faculty in each college

EVALUATION OF THE QUESTIONNAIRE

The instruments are being translated into online questionnaires using the commercial provider FormSite (www.formsite.com), which offers various tools for survey development and results presentations, as well as secure data input. Data from the online surveys were initially stored on the FormSite server and then downloaded to our server for analysis using SPSS software. Data were examined in SPSS to identify any discrepancies, problems, or outliers. In analyzing the results of the pilot, survey response rates are reported and descriptive analyses of variables conducted. After the pilot, questions with highly skewed responses or high non-response rates will be removed or rewritten.

VALIDITY MEASURES

Validity of measures was tested using a principle components extraction method with varimax rotation to statistically test whether scale items fit together. An exploratory factor analysis that analyzes the results from grouped questions that were answered similarly by faculty into independent factors was also performed. For more detail on factor analysis see references [14], [15], and [16].

For the principal components analysis, the Eigenvalue limit was set at one. We calculated component

scores by scoring questions from one to five (five always representing maximum learning outcome or teacher instruction), summing them, and expressing the total as a percentage of the maximum possible score for the component. If a respondent omitted half or more of the questions in a component we excluded these data from analysis. Questions that best described student outcomes desired by the engineering community were retained, thus maximizing our chances of achieving content validity. Evidence of construct validity was sought by calculating a Pearson's correlation coefficients matrix containing components for the overall student learning outcomes and teaching instructional practices scales.

RELIABILITY MEASURES

To make sure the reliability of measures was stable, repeatable, and consistent across respondents, Cronbach's alpha, an index of reliability associated with the variation, was used. By calculating Cronbach's alpha coefficient we estimated the internal consistency or reliability of each component. A Cronbach's alpha index above 0.7 for each component is generally accepted by experts.

RESULTS

Of the 102 respondents in the current sample of responses from three engineering colleges, the F-NSSE questionnaire was completed by 80 (78.4 percent) faculty members. The median (interquartile range) completion rate for questions was 96.5 percent (95.7 to 97.1 percent). Scale scores were calculated for a median (interquartile range) of 97.7 percent (94.5 to 98.1 percent) of responses.

Survey participants responded very well to the questions from the survey. At the end of the survey, participants were asked to make suggestions or comments about the survey and no new questions were identified by faculty members.

VALIDITY

Validity was measured in three steps. First, as highlighted in "Linking Student Learning Outcomes to Instructional Practices—Phase II," this project is grounded in the ABET engineering outcomes criteria and builds upon that grounding with the demonstrated knowledge base of the National Survey of Student Engagement as well as contemporary engineering education literature and the EC2000 study instruments [17]. Second, as highlighted in Bjorkland and Fortenberry [12], we conducted focus groups using a set of our survey items with both faculty and students at five engineering colleges. During the focus group, participants discussed the survey items to confirm the clarity and meaning of questions. Third, we did a principle component analysis and exploratory factor analysis, which analyzes the results from all questions and then groups questions that were answered similarly by faculty members into independent factors. The results from this analysis were used to indicate potentially bad questions (gave inconsistent results or seemed to be independent of the rest of the questions) and provided a set of independent categories. We also rechecked construct validity by calculating the intercomponent correlations.

The principal components analysis of the 80 completed questionnaires showed that the 15 learning components and 10 instructional practices were each judged to be coherent and to represent two separate scales related to student learning outcomes and teacher instructional practices. Tables 4 and 5 list each component with its Cronbach alpha coefficients, the means and standard deviations of the scale scores, and the variance explained by each scale. Tables 6 and 7 list each scale and the corresponding number of questions with each.

Table 4. Student Outcomes

Questionnaire Components	Cronbach's coefficient	Mean scale score	Standard Deviation	% Variance
1. An ability to apply knowledge of mathematics, science, and engineering	0.832	10.09	2.11	4.43
2. An ability to design and conduct experiments, as well as to analyze and interpret data	0.856	12.76	2.80	7.85
3. An ability to design a system, component, or process to meet desired needs	0.783	10.20	2.36	5.58
4. An ability to function on multi-disciplinary teams	0.909	21.39	5.30	28.11
5. An ability to identify, formulate, and solve engineering problems	0.894	13.24	3.24	10.49
6. An understanding of professional and ethical responsibility	0.919	14.44	4.53	20.48
7. An ability to communicate effectively	0.913	14.08	3.31	10.96
8. The broad education necessary to understand the impact of engineering solutions in a global and societal context	0.889	4.90	1.78	3.15
9. A recognition of the need for, and an ability to engage in, life-long learning	0.771	9.94	2.80	7.85
10. A knowledge of contemporary issues	0.924	10.49	3.28	10.73
11. An ability to use the techniques, skills, and modern engineering tools necessary for engineering practice	0.867	14.31	3.31	10.98
12. An ability to manage a project (including a familiarity with business, market-related, and financial matters)	0.893	13.95	4.08	16.63
13. A multidisciplinary systems perspective	0.789	9.10	2.51	6.29
14. An understanding of, and appreciation for, the diversity of students, faculty, staff, colleagues, and customers	0.777	14.27	3.59	12.94
15. A strong work ethic	0.346	17.62	2.86	8.16

Table 5. Instructional Practices

Questionnaire Components	Cronbach's coefficient	Mean scale score	Standard Deviation	% Variance
1. Encourage student-faculty interaction	0.594	23.35	2.64	6.983
2. Develop reciprocity and cooperation among students	0.753	16.17	2.64	6.99
3. Communicate high expectations	0.300	10.59	1.20	1.45
4. Give students feedback	0.649	10.51	1.37	1.87
5. Use active learning techniques	0.804	11.87	2.41	5.80
6. Emphasize time on task	0.555	9.70	2.71	7.37
7. Respect diverse talents and ways of thinking	0.463	16.68	2.74	7.51
8. Build on correct preexisting understandings, dispel false preconceptions	0.632	14.07	2.43	5.90
9. Provide factual knowledge, facilitate understanding of facts and ideas in context of a conceptual framework and organize knowledge that facilitates retrieval of application	0.518	5.97	1.37	1.87
10. Encourage students' motivation to learn	0.709	16.99	2.26	5.12

Table 6. Student Learning Outcome Scale

Student Outcomes	Number of Corresponding Questions
1. An ability to apply knowledge of mathematics, science, and engineering	3
2. An ability to design and conduct experiments, as well as to analyze and interpret data	4
3. An ability to design a system, component, or process to meet desired needs	3
4. An ability to function on multi-disciplinary teams	6
5. An ability to identify, formulate, and solve engineering problems	4
6. An understanding of professional and ethical responsibility	5
7. An ability to communicate effectively	4
8. The broad education necessary to understand the impact of engineering solutions in a global and societal context	2
9. A recognition of the need for, and an ability to engage in, life-long learning	5
10. A knowledge of contemporary issues	4
11. An ability to use the techniques, skills, and modern engineering tools necessary for engineering practice	4
12. An ability to manage a project (including a familiarity with business, market-related, and financial matters)	5
13. A multidisciplinary systems perspective	3
14. An understanding of and appreciation for the diversity of students, faculty, staff, colleagues, and customers	4
15. A strong work ethic	6

Table 7. Instructional Practices Scales

Instructional Practices	Number of Corresponding Questions
1. Encouraging student-faculty interaction	7
2. Developing reciprocity and cooperation among students	5
3. Communicating high expectations	3
4. Providing prompt feedback	3
5. Using active learning techniques	4
6. Emphasizing time on task	15
7. Respecting diverse talents and ways of thinking	9
8. Building on correct pre-existing understandings; dispelling false preconceptions	5
9. Providing factual knowledge, facilitating understanding of the facts and ideas in the context of a conceptual framework, and organizing knowledge that facilitates retrieval and application	2
10. Encouraging students' motivation to learn	5

Reliability

Cronbach's alpha was used to check the reliability of the intercorrelations of the constructs resulting from the factor analysis. Tables 4 and 5 list the questions in each scale and their constructs with their Cronbach alpha coefficients, the means and standard deviations of the scale scores, and the variance explained by each scale. The learning component scale has satisfactory internal reliability with Cronbach's alpha coefficients greater than 0.70 for all learning components and greater than 0.80 for 10 of the 15 learning components. Only 3 of the 10 instructional practices components had Cronbach's alpha coefficients greater than 0.70, indicating that intercorrelation was not very good.

DISCUSSION

The acceptability of the questionnaire to engineering faculty is shown by the high response rates for each question (median 96.5 percent) and the high proportion of responses for which we could calculate scale scores. We achieved response rates of over 50 percent from the three engineering programs who have thus far participated in the piloting of this survey after being online for three weeks. This shows that the instrument can successfully be administered via the web and probably by postal mail to a broad range of engineering faculty.

As indicated previously, content validity was ensured during questionnaire development through an extensive review of the relevant literature and by adapting items used in the National Survey of Student Engagement (NSSE), Faculty Survey of Student Engagement (FSSE), and EC2000 Study instruments. Content validity was initially shown by the outcome of the 10 focus groups conducted at five engineering colleges. Further evidence of content validity came from the outcome of the principal components analysis. The interscale correlations show that each scale is correlated with, and hence related to, student learning outcomes and teacher instructional practices. The scales assess different aspects of student learning outcomes and teacher instructional practices, a finding which argues in favor of construct validity. Future evaluations of this questionnaire should further examine construct and criterion validity.

The survey appears to have satisfactory internal reliability for at least one of its scales, the Learning Component scale, with the Cronbach's alpha index for each component above 0.7, which is generally accepted by experts. The Instructional Practices scale did not yield such good results, indicating that it may be necessary to add questions to improve the reliability of each component. Also, conflicting styles of teaching applications by professors may have lead to differences in scores on the scale. To further determine the reliability of the instruments, we anticipate a second administration of the surveys in 2007-08 to make sure faculty with similar characteristics respond in approximately the same way from year to year.

CONCLUSIONS AND FURTHER WORK

The current analysis indicates that this questionnaire has satisfactory reliability and validity, but more data and tests are needed to complete the validation process. This paper serves as the foundation for the results and conclusions from the analysis of our survey data and future applications of the survey. As indicated, the results presented here are preliminary. Further development of the questionnaire with a larger population is desirable. Data collection has been completed as of May 2007 for both faculty and students.

Table 8 provides a project timeline for the next phases of this validation process. Additionally, a subsequent pilot administration of the surveys to further investigate the psychometric properties of the instruments will be undertaken. The specific analyses we anticipate conducting include:

Reliability

Test-retest reliability – We will ask a group of respondents to complete the survey twice in a relatively short time period (e.g., one week) in order to determine test-retest reliability.

Stability – We will run two sample t-tests comparing data from the first year pilot and the second year pilot for respondents with similar characteristics, to check for differences in responses to each item from one year to the next. Items that have coefficients less than .6 would indicate those differences. They would not be considered stable and would be removed from the instruments.

Validity

Construct validity – In the second year of the grant, we will compare student survey responses and their end of the semester and cumulative GPAs.

Convergent validity – We will compare sample data to E-NSSE data from at least one participating institution.

Data collection, analysis, and refinement of the E-NSSE and E-FSSE surveys are still in progress. Over the next few months we plan to perform a factor analysis of the results for the E-NSSE, a final revision of the current questions, and creation of questions to target other categories that were not adequately addressed by the current version of the survey.

Table 8. Project Timeline

Activity		Period of Performance
Task 1a: Pilot Faculty Instruments		Nov – Dec 2006
Task 2a: Analyze Results of Faculty Pilot		Dec 2006 -Feb 2007
	2.1 Descriptive analysis of the variables	
	2.2 Principle components factor analysis	
	2.3 Check reliability of scales derived from factor analysis (using Cronbach's alpha)	
	2.4 Multiple regression analysis to examine the relationships between independent and dependent variables	
Task 1b: Pilot Student Instruments		Mar – Apr 2007
Task 2b: Analyze Results of Student Pilot		May – Jul 2007
Task 3: Refine the Survey Instruments		Jul – Aug 2007
	3.1 Delete redundant or extraneous items	
	3.2 Item analysis	
	3.3 Factor analyses using reduced model (including reliability tests)	
	3.4 Reformat instruments	
	3.5 Prepare for large-scale testing of the survey instruments	

REFERENCES

[1] The National Academies, *The Engineer of 2020: Visions of Engineering in the New Century,* (National Academy Press, Washington, D.C.), 2004.

[2] Olds, B., Moskal, B., and Miller, R., "Assessment in Engineering Education: Evolution, Approaches, and Future Collaborations," *Journal of Engineering Education*, vol. 94, no.1, pp.13-25, 2005.

[3] Bjorklund, S.A., and Fortenberry, N.L., *Final Report: Measuring Student and Faculty Engagement in Engineering Education*, CASEE Report 5902001-20050705, 2005.

[4] Cupp, S.M., Moore, P. D., and Fortenberry, N.L. "Linking Student Learning Outcomes to Instructional Practices – Phase I," *2004 ASEE Annual Conference and Exposition, Salt Lake City, UT*, Session 1630, (American Society for Engineering Education, Washington, D.C.) American Society for Engineering Education, 2004.

[5] *2003-2004 Criteria for Accrediting Engineering Programs*, (Accreditation Board for Engineering and Technology, Baltimore, Md.), 2002.

[6] Chickering, A.W., and Gamson, Z.F., (eds.). "Applying the Seven Principles for Good Practice in Undergraduate Education," *New Directions for Teaching and Learning,* vol. 47, 1991.

[7] Bransford, J.D., Brown, A.L., and Cocking, R.R., (eds.), *How People Learn: Brain, Mind, Experience, and School,* (National Academy Press, Washington, D.C.), 2000.

[8] "Learner-Centered Psychological Principles: A Framework for School Redesign and Reform," (American Psychological Association), http://www.apa.org/ed/lcp2/lcp14.html, 2004.

[9] National Survey of Student Engagement, NSSE 2006 Standard Version, available at http://nsse.iub.edu/html/sample.cfm.

[10] Faculty Survey of Student Engagement, FSSE 2006, available at http://fsse.iub.edu/html/fsse_survey_options.cfm.

[11] *Engineering Change: A study of the impact of the EC2000, 2006,* available at http://www.ed.psu.edu/cshe/abet/ec2000.html.

[12] Bjorklund, S. A., and Fortenberry, N. L.. "Linking Student Learning Outcomes to Instructional Practices -- Phase III," *2005 ASEE Annual Meeting and Exposition, Portland, Ore.,* Session 3630, (American Society for Engineering Education, Washington, D.C.) 2005.

[13] National Survey of Student Engagement, "Frequently Asked Questions About NSSE's Psychometric Properties." Available at http://nsse.iub.edu/nsse_2002/faqpsych.htm.

[14] Crocker, L. and Algina, J., *Introduction to Classical and Modern Test Theory,* (Hold, Rinehart and Winston, Inc, Fort Worth, Texas), 1986.

[15] Kachigan, S.K., *Statistical Analysis,* (Radius Press, New York), 1986.

[16] Kim, J. and Mueller, C.W., *Factor Analysis—Statistical Methods and Practical Issues*, (SAGE Publications, Beverly Hills), 1978.

[17] Moore, P. D., Cupp, S. M., and Fortenberry, N.L., "Linking Student Learning Outcomes to Instructional Practices – Phase II," *Annual Frontiers in Education Conference,* (ASEE/IEEE,Boulder, Colo.), 2003.

Assessment Resource Tools for Assessing Students' Statistical Literacy, Reasoning, and Thinking

Joan Garfield* and Robert delMas - *University of Minnesota*
Ann Ooms - *Kingston University (UK)*
Beth Chance - *California Polytechnic State University*

OVERVIEW

This paper describes the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project, which aimed to assist faculty who teach statistics across many disciplines in assessing student learning of statistics, and to enable them to: better evaluate individual student achievement; evaluate and improve their courses; and allow them to assess the impact of reform-based instructional methods on the attainment of statistical literacy, reasoning, and thinking. ARTIST consists of a website that provides resources, including a large, searchable assessment item database, several online tests assessing knowledge of a particular statistical topic, and a comprehensive test of statistical literacy and reasoning called the Comprehensive Assessment of Outcomes in a first Statistics course (CAOS). Details of the development of the ARTIST resources, results from an extensive evaluation of the project, and plans for development of future ARTIST resources are presented.

THE ARTIST PROJECT

The National Science Foundation (NSF) funded the ARTIST project (DUE-0206571) to address many current assessment challenges in statistics education. These were presented by Garfield and Gal [1], who outlined the need to develop reliable, valid, practical, and accessible assessment instruments. The ARTIST website (https://app.gen.umn.edu/artist/) now provides resources for evaluating students' statistical literacy (e.g., understanding words and symbols, being able to read and interpret graphs and terms), reasoning (e.g., reasoning with statistical information), and thinking (e.g., asking questions and making decisions involving statistical information). These resources were designed to assist faculty who teach statistics across various disciplines (e.g., mathematics, statistics, and psychology) in assessing student learning of statistics, in evaluating individual student achievement, in evaluating and improving their courses, and in assessing the impact of reform-based instructional methods on important learning outcomes. The project was run by a team of three co-investigators (delMas, Garfield, and Chance), each with unique areas of expertise in statistics education, and a graduate research assistant (Ooms, later a postdoctoral fellow) with expertise in technology and evaluation. The ARTIST project was fortunate to have a strong and diverse advisory group: Julie Clark (Hollins University), George W. Cobb (Mount Holyoke College), John Holcomb (Cleveland State University), Frances Lawrenz (University of Minnesota), Carl Lee (Central Michigan University), Marsha Lovett (Carnegie Mellon University), Anthony Onwuegbuzie (University of South Florida), Roxy Peck (California Polytechnic State University), Michael Rodriguez (University of Minnesota), Allan Rossman (California Polytechnic State University), Deborah J. Rumsey (Ohio State University), and Candace Schau (CS Consultants). The advisors provided expertise in development, evaluation, and implementation of assessment items, resources, and instruments.

The project, which is currently being funded by a supplement grant, has so far produced the following products:

- *A collection of over one thousand high-quality assessment items and tasks, coded according to content (e.g., normal distribution, measures of center, bivariate data), type of cognitive outcome (e.g., statistical literacy, reasoning, or thinking), and type of item. Visitors can use a set of linked pages (called the Assessment Builder) to search, review, select, and download items into rich text format (rtf) files that may be saved and modified on their own computers with a word processing program.*

- *A website that provides access to the assessment item database, as well as many other resources, including references and links to articles on assessment, information on alternative assessment methods including samples of project guidelines and student work, grading rubrics, research instruments, materials from professional development offerings, ARTIST advisory board responses to questions on assessment implementation issues, and web links.*

- *Separate online tests that measure conceptual understanding in 11 important areas of a first course in statistics that have high validity and reliability.*

- *CAOS, a test that measures statistical literacy and reasoning.*

- *Professional development opportunities for teachers that include mini-courses, workshops, and conference presentations to encourage and assist statistics instructors in using assessment resources to improve student learning, improve their courses, and evaluate course outcomes.*

In addition, an extensive evaluation of the ARTIST resources was used to develop a model for evaluating online educational resources. A project in progress is development of a Statistical Teaching Inventory to assess teachers' instructional methods and beliefs that affect their teaching. This inventory may be used as part of future research studies in statistics education.

DEVELOPMENT OF ONLINE TEST MATERIALS

Item Database

The assessment item database was one of the first products developed. Initially, items were collected from exams of the project staff (co-investigators and advisory board members) and also solicited from the statistics community through a posting on the ARTIST website. The co-investigators and some of the advisors reviewed items to organize them by topic and learning outcome. Items that were purely computational were eliminated from the database, unless they could be modified into a literacy, reasoning, or thinking type of item. A context was developed and added to items that were not set in a context. Most true/false items were changed into three-option forced-choice items, and forced-choice versions were created for numerous open-ended items, although there are also many open-ended items in the database. All items were edited for statistical content and typographical errors. Knowing that some errors may have been missed, a mechanism was developed so that users of the assessment item database can report concerns with individual items. The database currently consists of more than 1,100 items, with new items being added periodically as they are submitted by ARTIST users and reviewed.

Online Unit Tests

The eleven online unit tests were developed over two years. During this process, the ARTIST team developed and revised items and the ARTIST advisory board provided valuable feedback and ratings of the validity of items, which were used to determine and improve scale validity. The topics for the 11 online unit tests are: Data Collection, Data Representation, Measures of Center, Measures of Spread, Normal Distribution, Probability, Bivariate Quantitative Data, Bivariate Categorical Data, Sampling Distributions, Confidence Intervals, and Tests of Significance. These tests cover an intersection of topics included in most introductory statistics courses. Each test consists of seven to twelve multiple-choice items that assess literacy and reasoning for that topic. Online versions of each topic scale were created and evaluated in two rounds of class testing and test revision as described below.

The CAOS Test

The Comprehensive Assessment of Outcomes in a first Statistics course (CAOS) was created through a similar process of development, revisions, feedback from advisors and class testers, and a large validity assessment using 30 experienced Advanced Placement Statistics readers. The current version of CAOS consists of 40 multiple choice items and can be administered online or in a print copy using a machine-scannable bubble sheet. Topics covered on the CAOS test include basic literacy and reasoning about descriptive statistics, probability, bivariate data, and basic types of statistical inference. Again, the intent was to develop a set of items that students completing any introductory statistics course would be expected to understand.

In order to access the online tests, an instructor requests an access code, which is then used by students when they are ready to take the test. As soon as the students have completed the test, either in class or out of class, the instructor may download two reports of students' data. One is a copy of the test, with percentages filled in for each response given by students, and with the correct answers highlighted. The other report is a spreadsheet with percentage correct scores for each student.

CLASS TESTING OF ONLINE TESTS

A large-scale class testing of the online instruments (11 unit tests and CAOS test) was conducted during spring 2005. Students for the study were obtained via invitations sent to high school Advanced Placement (AP) and college statistics instructors through email lists of major United States organizations likely to have a membership that represents this population (e.g., AP listserv, Statistics Education Section of the American Statistics Association), ads placed in magazines and newsletters (e.g., AMSTAT News), and information posted on the ARTIST website. Instructors registered their students to take the ARTIST topic scales at points in their courses when students had covered the material assessed by a scale. Nearly 100 secondary-level students and 800 college-level students participated. The spring 2005 results were used to make minor revisions and produce final versions of each scale during summer 2005.

During the class testing phases, data were gathered from student responses and used to review each item so that answers could be revised or deleted as needed. Input from class testers (course instructors) was also utilized during this period to revise and improve the tests. Class testers have found these tests to be useful for a variety of purposes, including testing (for a grade), review, self-testing, and extra credit.

The eleven unit tests and the CAOS test were administered in a second large-scale testing during fall 2005 and spring 2006. The numbers of students who have taken each test across the class testing periods are listed in Table 1.

Table 1. Number of students who have taken each ARTIST test*

TEST	NUMBER OF STUDENTS
CAOS	5155
Data Collection	1064
Data Representation	1477
Measures of Center	1796
Measures of Spread	1352
Normal Distribution and Measures of Position	1640
Probability	1127
Sampling Variability	957
Confidence Intervals, One-Sample	1479
Tests of Significance	1577
Bivariate Data, Categorical	1286
Bivariate Data, Quantitative	1879

*Counts represent all students who took any version of each test.

PSYCHOMETRIC ANALYSIS OF CAOS TEST

A large sample of students enrolled in higher education institutions within the United States was used to assess the reliability of the CAOS test. The sample was composed of a total of 1,470 introductory statistics students, taught by 35 instructors from 33 higher education institutions in 21 states across the United States (see Table 2). The majority of the students whose data were used for the reliability analysis were enrolled at a university or a four-year college, with about one-fourth of the students enrolled in two-year or technical colleges. A little more than half of the students (57percent) were females, and 74percent of the students were Caucasian. The CAOS test was found to have high internal consistency reliability (Cronbach alpha = .82).

Table 2. Number of higher education institutions, instructors, and students per institution type for students who completed the CAOS posttest

Institution Type	Number of institutions	Number of instructors	Number of students	Percent of students
2-year/technical	6	6	341	23.1
4-year college	13	14	548	37.3
University	14	15	581	39.5
Total	**33**	**35**	**1470**	

In March 2006, a final analysis of the content validity of CAOS was conducted. A group of 18 members of the advisory and editorial boards of the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) were used as expert raters. These individuals are statisticians who are involved in teaching statistics at the college level, and who are considered experts and leaders in the national statistics education community. They were given copies of the CAOS test, which had been annotated to show what each item was designed to

measure. After reviewing the annotated test, they were asked to respond to a set of questions about the validity of the items and instrument for use as an outcome measure of student learning after a first course in statistics. There was unanimous agreement by the expert raters with the statement "CAOS measures basic outcomes in statistical literacy and reasoning that are appropriate for a first course in statistics," and 94 percent agreement with the statement "CAOS measures important outcomes that are common to most first courses in statistics." In addition, all raters agreed with the statement "CAOS measures outcomes for which I would be disappointed if they were not achieved by students who succeed in my statistics courses." Although some raters indicated topics that they felt were missing from the scale, there was no agreement among these raters about the topics that were missing. Based on this evidence, the assumption was made that CAOS is a valid and reliable measure of important learning outcomes in a first course in statistics.

EVALUATION OF THE ARTIST WEBSITE

After developing the ARTIST website, item database, and online assessments, an evaluation was designed and administered to determine how instructors were using the materials and how the materials were helping to achieve the original project goals related to improving the teaching and learning of statistics.

As part of this evaluation, Ann Ooms created an evaluation model called The Iterative Evaluation Model for Improving Online Educational Resources as part of her doctoral thesis [2]. The model consists of four components: evaluation planning, educational value (including web design and web content), the use of the educational resource, and educational impact (including impact on instructors' perceptions, impact on educational environment, impact on student outcomes, and sustainability of impact). The model was used as a guide to select data collection methods and to create the data collection instruments. The evaluation of the ARTIST website included:

- *A 50-item online survey of ARTIST users to learn how they were using the materials and how they were changing their assessment practices. Ninety-eight ARTIST users responded to the online survey.*

- *A 52-item online survey of non-users to find out why they were not using the resources and to determine whether they differed from the users in important ways, such as wanting more computational items. The survey was identical to the one used for ARTIST users except for the addition of two questions. Eighty-nine non-users responded.*

- *Observations of two new and three experienced users of the ARTIST assessment item database as they interacted with web pages designed to access the assessment item database. Participants were asked to think aloud while performing certain tasks, such as creating a test, adding a question to a test, and removing a question.*

- *In person interviews with seven non-users of ARTIST resources to find out why they were not using these materials.*

- *In person interviews with seven frequent users of ARTIST resources to learn about how their use of ARTIST materials and resources had impacted their assessment practices and/or teaching.*

Results were summarized and examined for the four components of the evaluation model. In terms of web content, all the evaluation findings were positive and did not indicate the need for major changes to the website. The user survey respondents were fairly regular users of the ARTIST website, with 85 percent having visited the site more than twice, and 54 percent having visited the website more than 5 times. The website seems to be reliable, with less than 14 percent of the people who browsed the website indicating they encountered at least one error message. With respect to navigating the ARTIST website, 94 percent reported that the links were labeled in a way that was descriptive of their content. The organizational scheme of the website received high ratings: 74 percent rated the quality of the organizational scheme as good, and 19 percent rated it as excellent. Ninety percent of respondents found it easy to locate a particular topic within the website. The majority of respondents (more than 90 percent) reported finding the website "somewhat attractive" to "attractive."

Results from the non-user survey indicated that many had not used ARTIST resources because they were not aware of the website. Forty-five percent of the non-users had not heard of the ARTIST website, and 69 percent were not aware of the ARTIST assessment item database resource. Interviews with non-users identified some misconceptions about the ARTIST website (e.g., it was for the assessment of teaching, or that it required students to have online access in the classroom). As a result, we have publicized the website and its resources more broadly with detailed information about what it provides.

In terms of web design, survey results showed that some steps in the process of using the web resource were "somewhat difficult to use." A summary of the results from questions on ease of use for the Assessment Builder indicated that users found it fairly easy to log on, to view items, select and remove items, and to download a set of selected items. More difficulty was indicated for conducting and refining a search. Respondents also rated the quality of several features of the Assessment Builder (e.g., instructions, information buttons, navigation, response time). All features were rated to have high to very high quality by a majority of the respondents. Observations of new and experienced ARTIST users identified problem areas that suggested many changes and improvements in the web design and content that have been made or are currently being made (e.g., ways to improve searching the assessment item database; and the addition of large data sets for use in take home finals or student projects).

Evaluation data were also gathered on estimated frequency of use of various ARTIST web pages and ratings of their usefulness. Some web pages were clearly being used more than others. Of the web pages that were rarely used, the vast majority of those who accessed these pages still found them useful. One question that the evaluation was not able to address was why people who visited the website once did not return to use it again.

The section on the Assessment Builder in the ARTIST user survey contained a set of questions on the quality and use of the assessment items. Respondents were asked to compare the quality of the ARTIST assessment items to items from other sources such as textbook item databanks. In general, ARTIST items were found to be of the same or higher quality. Notably, ARTIST items were judged to focus more on the assessment of conceptual understanding and to incorporate a context more often when compared to items from other sources. Ninety-one percent of the respondents used the ARTIST items in quizzes and for exams. The second most frequent use of the items was for review (44 percent), and the third most frequent use was for class examples (37 percent). Twenty-five percent of respondents used the items for class group activities and 19 percent for homework.

Evaluation of educational impact of ARTIST products was based on interviews with ARTIST users. At the time of the interviews, ARTIST users had not yet implemented any changes to their instruction in response to the ARTIST products. However, there was evidence that they had started to think about changes to their teaching, and that the ARTIST products were having a positive impact on instructors' perceptions. The survey and interview results indicate that instructors who use many of the ARTIST scales and tests have been rethinking some of their teaching approaches in order to have a greater impact on students' statistical reasoning and thinking. Still, it was not possible at this stage to evaluate the educational impact of ARTIST in terms of the educational environment.

CURRENT STATUS OF ARTIST RESOURCES

There are currently 1,115 registered users of the ARTIST item data base, and 326 registered users of the online tests. An additional instrument was developed called the Statistical Thinking and Reasoning Test (START), by identifying a subset of items from CAOS that had high internal consistency and covered the main topics of an introductory statistics course. This 14-item test can be given multiple times over a course more easily than the 40-item CAOS test, allowing researchers to study longitudinal growth of students' statistical reasoning. Data are currently being gathered at two institutions on the START test.

A second current development is the plan to move the ARTIST website to causeweb.org, where it will be maintained and supported in the years to come. CAUSE [3] is funding this transition as part of their National Science Digital Library (NSDL) grant from NSF.

ARTIST PROJECTS IN DEVELOPMENT

The Statistics Teaching Inventory (STI) is being developed to gather information on teachers' beliefs about teaching, the alignment of their instructional practices with current reform recommendations, and the constraints under which they teach (school and student variables). This instrument will be given to instructors whose students take tests on ARTIST, and the data will be stored so that they may be analyzed along with the student test data.

FINAL REMARKS

The ARTIST project has developed many high-quality, valuable, and useful assessment resources for the statistics community. Based on the number of the users, the positive feedback from evaluations, and the frequent recommendations in the statistics education community to access and use the ARTIST website, the project seems to be successful in meeting many of its goals. However, it still remains to be seen how use of these resources is impacting students' learning of statistics. That can be the focus of a subsequent grant!

REFERENCES

[1] Garfield, J. and Gal, I., "Assessment and Statistics Education: Current Challenges and Directions," *International Statistical Review,* vol. 67, pp. 1-12, 1999.

[2] Ooms, A., *The Iterative Evaluation Model for Improving Online Educational Resources*, doctoral dissertation, University of Minnesota, 2005.

[3] Consortium for Advancing Undergraduate Statistics Education (CAUSE), http://www.causeweb.org.

Using Self-assessment Within an Electronic Portfolio as a Framework for Development of Student Problem-solving and Analysis Abilities Across Courses and Majors Within Science and Mathematics

Lauralee Guilbault*, Sherry Dollhopf, Susan Pustejovsky, Tracy Thompson, Leona Truchan, and Chris Young - *Alverno College*

This paper describes the ongoing work of a group of Alverno faculty who have been designing and modifying a series of learning experiences to help students develop analysis and problem-solving abilities in science, mathematics, and technology (SMT) courses. The results of students' work in these learning experiences, along with student self-assessment and instructor feedback, are stored in Alverno's electronic portfolio called the diagnostic digital portfolio (DDP). Faculty working on this project also seek to determine how best to use the DDP to help build coherence in student development of analysis and problem-solving abilities across majors, disciplines, and courses.

Entries in the DDP are organized in a variety of searchable categories, including by student, by course, by discipline, and by the abilities that are demonstrated in the student's work [1, 2]. These abilities, which define Alverno's ability-based curriculum, provide descriptions of what students should master in order to graduate [3,4,5]. This project focuses on two key abilities central to disciplines within SMT: Analysis and Problem Solving.

Analysis has been defined by the Alverno faculty in developmental levels based on articulated criteria as the ability to appropriately: observe individual parts of phenomena and their relationship to one another (beginning level); use disciplinary concepts and frameworks with growing understanding (intermediate level); and consciously and purposefully apply disciplinary frameworks to analyze complex phenomena (advanced level).

Problem-Solving has been defined by the Alverno faculty as the ability to appropriately: articulate the problem-solving process and understand how a discipline framework is used to solve a problem (beginning level); take thoughtful responsibility for processes and proposed solutions to problems (intermediate level); and use problem-solving strategies in a wide variety of professional situations (advanced level).

In the first year of the grant the three PIs, representing biology, mathematics, and chemistry, surveyed DDP portfolios of their majors to see what types of analysis and problem-solving projects were being recorded in SMT courses. DDP entries were found for upper-level work in SMT, but very little for intermediate-level work, and almost nothing for beginning-level work. In fact, it was found that most early documentation of SMT students' problem-solving and analysis work came from other disciplines. In order to help students and faculty document the students' development of problem solving and analysis in SMT courses many more learning experiences and assessments needed to be developed or modified for DDP inclusion.

Guidelines for creating and modifying assessment projects were created and presented at a workshop for a group of faculty who had been recruited to work, over the summer, on the next phase of the project. The workshop provided participants with a forum for brainstorming and idea sharing, and each participant was

provided with a list of resources put together by the PIs. One of the main ideas emerging from this workshop was that learning should be developmental, within, across, and beyond courses.

As follow-up to the summer work, faculty met regularly throughout the 2005-06 school year to share observations and experiences about the implementation of their projects. From this discussion, common themes emerged. Most projects stretched across the semester, and relied heavily on student self-assessment to document student analysis, and help students focus on their development. Students were often directed to self-assess their analysis and problem-solving skills throughout a semester as they repeated similar projects.

The Alverno faculty defines self-assessment as the ability of a student to observe, analyze, and judge her performance on the basis of criteria and determine how she can improve it [6]. The steps of the self-assessment process are more clearly delineated in the Alverno College self-assessment framework (Appendix 1).

This definition of self-assessment emphasizes the analytical process students go through in order to evaluate their performance in relation to specific criteria. Students should make observations about their work then use these observations as evidence to draw conclusions about if and how well they met criteria for a particular assignment. Finally, they set goals for further development, and sometimes propose strategies to reach these goals.

Through their discussions, faculty found similar challenges getting students to do the types of detailed self-assessments that both showed good accurate analysis, and were appropriate tools for student development. In self-assessments students often stated conclusions about their performance without giving evidence to support these conclusions, or they might tell what they did, but not connect their actions to evidence or stated criteria. Students also used self-assessment as a place to reflect on an issue, such as *green chemistry is good*, instead of reflecting on how they did their work related to that issue. To overcome these challenges, the group developed a number of guide questions and models for students to use in more clearly articulating the analytical steps involved in self-assessment. Some examples are shown in Appendix 2.

DEVELOPMENTAL PROBLEM SOLVING – BIOLOGY

Some examples from biology exemplify Alverno's developmental stages of problem solving defined previously. At the beginning level, the students are learning to articulate their problem-solving process in the context of a biological system. Introductory biology students use a computer-based problem-solving simulation in which they manipulate variables, such as oxygen levels, in an effort to make the greatest profit raising fish. Students initially use the simulation in a simple experiment, each student adjusting the variables affecting one fish in a well-defined environment. In later classes, students reflect on the effectiveness of their procedures in working with their first fish, and build on this experience when they independently repeat the experiment with a mystery fish. The students' work and self-assessment, which focuses on their views of themselves as problem solvers, is recorded on the DDP and serves as evidence for beginning students' level of performance in problem solving and analysis.

At the intermediate level, students take more ownership of design and implementation of their problem-solving process. In microbiology laboratory, students learn a variety of manipulative techniques. They apply these techniques to a final project in the second half of the semester where they use their analysis and problem-solving abilities to determine the identity of an unknown organism. To prepare students for this independent problem-

solving project, biology faculty designed an ongoing simulated research project for the first eight weeks of the laboratory. Students select one organism, and research and analyze how the laboratory procedures introduced each week help in their understanding of their chosen organism. Through this entire process they create an analytic concept map of their research that demonstrates how data gained from each lab procedure are pieced together to create an overall picture of their organism. At the end of the semester, students use their self-assessment to reflect on both projects and how the virtual experience prepared them for the actual experiment.

At the advanced level, in Mechanisms of Disease students evaluate more complex, real-life situations. In this course students use case studies to analyze the underlying causes of physiological phenomena. In the past, students often approached case studies as if they were simply being asked to remember detailed information, not to solve a problem, thus many students performed poorly and ultimately failed in the course. To help students develop problem-solving skills, faculty introduced a series of self-assessment prompts, such as the ones in Appendix 2, focusing on the student's process in working through case studies. Students gain confidence and awareness of their own problem-solving abilities by doing these self-assessment activities. This supports their growth in the way they view themselves as professionals in their field.

COHERENCE WITHIN A MAJOR: MATHEMATICS

Mathematics faculty used a series of learning experiences and self-assessments related to reading and writing proofs as a common focus for building coherence across the mathematics major. Throughout their coursework students develop progressively more sophisticated understandings of proofs, and skill in this central mathematical process. Using self-assessment, students observe how their ability to work with proofs continues to develop. The collection of a portfolio of work, gathered over the course of a semester and periodically reflected on, became an assessment model that resonated with faculty. Some examples from projects that include this focus are:

- *Students need to recognize that more work is required beyond just solving a mathematical proof. There is a common formal style that is adopted, moving from what is given to what is to be proved, employing accepted rules of deduction. One mathematician found that the simple act of having students load their written proofs to the DDP intensified awareness of this formal aspect of their work, and seemed to result in better student effort to write with clarity and completeness.*

- *In geometry, students reviewed a portfolio of their work and categorized proofs according to similarities. Students often inappropriately grouped certain proofs together based on superficial similarities, such as if a problem involved a circle or a triangle. Faculty refined self-assessment activities to help students focus on the more salient aspects of their approach rather than surface similarities. This clarified for students that the focus of self-assessment is their process.*

- *In Axiomatic Systems students used the same set of criteria throughout the semester to evaluate all their problem solutions and proofs (Appendix 2). Through repeated practice, rewrites, and self-assessment reflection on how their work fulfilled criteria, students became more proficient at clearly articulating the parts of their proofs (hypothesis and conclusion) and the logical progression of the proof.*

The DDP extends this portfolio process through a student's entire program in mathematics. Faculty create, for example, opportunities within courses for students to look back at their proof work in previous courses. Students are asked to reexamine mathematical induction proofs that they completed in Discrete Mathematics, and compare them to proofs they complete in later courses. Using the DDP, students experience their coursework as part of a coherent whole with mathematical ways of thinking, reasoning, and communicating rather than merely separate distinct courses, each focused on different topic areas.

THE SELF-ASSESSMENT – FEEDBACK LOOP: CHEMISTRY

Chemistry students use iterative loops of self-assessment and instructor feedback to help develop themselves as better problem solvers andl as better self-assessors.

In analytical chemistry, students use goal setting, self-assessment, and feedback to develop their problem-solving skills throughout the semester. On the first day of class, students review basic chemistry problems involving conversions, and they explain, out loud to a classmate, the process they use to solve one problem. Their self-assessment for this project is based on Criterion-based Self-assessment on Problem Solving found in Appendix 2, in which they observe and describe the strengths and weaknesses in their process and then use these observations to set a problem-solving goal. Faculty give feedback on how realistic the goal is and on whether a student's strategies are specific enough to be useful in pursuing her goal. For instance, if a student observed that she was sloppy and made copying errors in her calculations, she might set a goal to carefully pay attention to details in the future work. Instructor feedback would include that some errors can be caught by carefully checking the numerical answer for its appropriateness.

At midterm students again evaluated their problem-solving strategies and progress on goals using self-assessment. Students compare their work on an in-class test to the work they did for a particular laboratory experiment to look for patterns in their problem-solving process. In preparation for the presentation of their final lab project, students review their past self-assessments so they can comment on the progress they made on their goal. Because they have been cognizant of their problem-solving goals throughout the semester, they typically describe multiple pieces of evidence from both laboratory and classroom work to show what they have been able to do.

In organic chemistry laboratory, besides chemical analysis and problem solving, emphasis is placed on the analytical nature of self-assessment. In evaluating their work, students create concept maps showing relationships between their observations, conclusions, and criteria. They use visual models to create awareness of the strong connections between the process of self-assessment and the analytical and problem-solving skills they are developing in their lab work. The instructor gives feedback on the self-assessment that may validate a student's ability to accurately assess her performance. Feedback provides guidance by helping the student to focus on using more descriptive and varied evidence, more clearly stated conclusions, or more concrete goals and strategies.

Early in the semester, students may be asked to revise self-assessments. The iterative nature of self-assessment and feedback improve students' ability to self-assess as evidenced by faculty validation in feedback. The electronic portfolio provides a common deposit for student assessment and instructor feedback to support development in self-assessment and accountability in following through with student-generated goals.

SUMMARY

Because the group was multidisciplinary, collaboration proved to be a direct faculty development experience. Sharing findings, questions, and successes, as well as challenges, provided rich extrapolation across disciplines. The group created and modified projects to help students become better self-assessors, as they also developed their skills in problem solving and analysis. Students applied these skills to document their progress at the beginning, intermediate, and advanced levels. Faculty built coherence within majors by focusing on common themes. Self-assessment played a significant role in all projects.

Our project addresses broader issues in higher education by highlighting the dramatic impact that self-assessment and the use of the digital portfolio have on student learning. Assessment of student achievement is traditionally focused on instructor assessment without giving students a chance to reflect or comment on their own work. This results in a dead end for the student, especially if the assessment comes at the end of a course where there is no opportunity for the student to continue to grow and improve based on feedback.

Well-defined criteria and self-assessment prompts provide needed guidance to students, as they reflect on their own progress and learning goals, evidenced by material in their portfolio. Using the DDP, an instructor can provide feedback on the original student performance and on the student's self-assessment, helping the student to improve more than one ability at a time. Self-assessment is missing in most assessments of student achievement at the college level. Our work raises awareness of the value of self-assessment evidenced in student work over time.

ACKNOWLEDGEMENTS

The authors thank Linda Ehley, James Wend, Lois Kailhofer, Marie Elizabeth Pink, Jennifer Johanson, and Patricia Walsh for their insight and collaboration on this project. We also thank the National Science Foundation (grant award # DUE-0404986) for funding this work.

REFERENCES

[1] Ehley, L., *Digital Portfolios: A study of undergraduate student and faculty use and perceptions of Alverno College's Diagnostic Digital Portfolio*, unpublished doctoral dissertation, Cardinal Stritch University, 2006.

[2] Rickards, W., Diez, M., Ehley, L., Guilbault, L.F., Loacker, G., Reisetter Hart, J., and Smith, P. C. *Learning, Reflection, and Electronic Portfolios: Stepping toward an assessment practice*, (under review).

[3] Alverno College Faculty, *Student assessment-as-learning at Alverno College*, (Milwaukee, Wis.), Alverno College Institute, original work published 1979, revised 1985 and 1994.

[4] Mentkowski, M., and Doherty, A., "Abilities that last a lifetime: Outcomes of the Alverno experience," *AAHE Bulletin,* vol. 36, no. 6, pp. 5-6 and11–14, 1984.

[5] Mentkowski, M., and Associates, *Learning that lasts: Integrating learning, development, and performance in college and beyond*, (Jossey-Bass, San Francisco), pp. 63-64, 2000.

[6] Alverno College Faculty, *Self-assessment at Alverno College,* (Alverno College Institute, Milwaukee, Wis.), 3 pp., 2000.

APPENDIX 1

Alverno College Self-assessment Framework

	Beginning	**Intermediate**	**Advanced**
Observing	I*n OBSERVING her performance a BEGINNING student uses specific behaviors and descriptive strategies to focus primarily on the development of her abilities. She uses concrete detail, articulates questions to show processes of thought, and systematically reviews parts to explain aspects of her processes or products. She develops ways to communicate what she intended to do, how she worked toward that goal, and what she achieved. She recognizes how her expectations, prior learning, thoughts, and emotions influenced her ability to focus on her performance.* DIMENSIONS OF COMPONENT • Reports own behavior (actions, thoughts, and emotions) in performance and/or in the process of producing a performance • Distinguishes actions from emotions	*In OBSERVING her performance an INTERMEDIATE student reflects on her explicit use of disciplinary, interdisciplinary, and/or ability frameworks in her performance. She shows, by meaningfully relating concepts to her performance, what aspects of the ability or discipline have influenced her.* DIMENSIONS OF COMPONENT • Identifies examples of her use of disciplinary, interdisciplinary, and/or ability frameworks • Communicates observations using language appropriate to the selected framework(s). • Shows she was aware of her performance at appropriate times	*In OBSERVING her performance an ADVANCED student applies disciplinary, interdisciplinary, and/or ability frameworks and shows creative judgment in their individual or combined use. She both attends to her current engagement in her learning and imagines her future practice as a professional.* DIMENSIONS OF COMPONENT • Applies disciplinary, interdisciplinary, and/or ability frameworks to the observation of performance • Maintains balance between personal distance and personal engagement • Interprets/Analyzes

	Beginning	Intermediate	Advanced
Interpreting/ Analyzing	*In INTERPRETING/ANALYZING her performance a BEGINNING student goes beyond treating behaviors discretely. She articulates connections among them in relation to a criterion, an ability, or a series of performances.* DIMENSIONS OF COMPONENT • Identifies patterns of strengths and weaknesses in behavior • Organizes details in relation to an identified focus • Relates self-assessment and feedback • Articulates impact of emotions on her ability to plan for a performance and to perform	*In INTERPRETING/ANALYZING her performance an INTERMEDIATE student identifies patterns in her performance, expressing her awareness of how these patterns affect the whole.* DIMENSIONS OF COMPONENT • Explains the significance of patterns in performance • Makes sense out of her performance in relation to disciplinary, interdisciplinary, and/or ability frameworks • Uses feedback to develop a larger picture of performance	*In INTERPRETING/ANALYZING her performance an ADVANCED student articulates how it is uniquely her own in relationship to disciplinary, interdisciplinary, and/or ability frameworks. She uses her imagination to project how she might extend and refine it.* DIMENSIONS OF COMPONENT • Explains components of performance that make it unique and distinctive and are part of a student's style or voice • Uses disciplinary, interdisciplinary, and/or ability frameworks in a way that reflects, extends, or recreates them • Synthesizes patterns of behaviors and processes over time and in varied contexts

	Beginning	Intermediate	Advanced
Judging	I*n JUDGING her performance a BEGINNING student uses her knowledge of the criteria to explain how her performance gives evidence of the behaviors inherent in the criteria. She explores meaning implied in the criteria.* DIMENSIONS OF COMPONENT • Makes connections between criteria and behaviors • Relates judgment of current performance to past experiences	*In JUDGING her performance an INTERMEDIATE student understands that the set of criteria as a whole interact to create a picture of the ability(ies)in performance and that she needs to assess it in terms of the ability(ies) rather than each criterion.* DIMENSIONS OF COMPONENT • Makes sense of a set of criteria as a whole in relation to her judgment of performance	*In JUDGING her own performance an ADVANCED student uses a picture of an ideal performance that illustrates the abilities integrated with disciplines. Based on her judgments during performance, she may modify her ideal expectations and maintain or change behaviors. She evaluates her use of judgment and her modification of behavior during performance and over time.* DIMENSIONS OF COMPONENT • Articulates her use of criteria and knowledge of the integration of her actions, thoughts, and emotions to self monitor and to adjust ongoing actions or plans accordingly • Shows where intervention or modification has or should have taken place

	Beginning	Intermediate	Advanced
Planning	*In PLANNING for further development a BEGINNING student shows awareness that she is changing in her approach to learning and can use various strategies to learn. She sees implications for her future learning.* DIMENSIONS OF COMPONENT • Identifies aspects to maintain for performance and/or the process of producing a performance • Identifies aspects to further develop her learning and suggests approaches for future performance	*In PLANNING for further development an INTERMEDIATE student is aware of herself as a learner. In considering her future performances she effectively uses feedback, uses specific goals to set direction, and recognizes her prior strengths and weaknesses.* DIMENSIONS OF COMPONENT • Relates goals for improvement to progress thus far and to possibilities for future development • Uses feedback in planning for future performance • Uses awareness of her emotional responses to plan for continuing development	*In PLANNING for future development an ADVANCED student is aware of herself as a lifelong learner. With attention to her style and creativity, she shapes her future performances by considering her past work, her intellectual processes, and a variety of discipline models and frameworks* DIMENSIONS OF COMPONENT • Uses multiple models of performance to set and continue to refine goals for continuous development • Considers her style and creativity in planning for future performances

APPENDIX 2:

Example of Criterion-based Self-assessment on Problem Solving

Criteria for this self-assessment:

a. Accurately describes own effectiveness as problem solver and analytical thinker
b. Reasonably describes patterns in own work as a problem solver and analytical thinker
c. Uses evidence from past performances to support own descriptions
d. Sets reasonable goal for development based on past performance

1. Choose one of the analysis skills listed below that you think you did well on in this key performance. Describe how you demonstrated that skill using specific examples from this project.

2. Choose one of the problem-solving skills listed below that you think you did well on in this key performance. Describe how you demonstrated that skill using specific examples from the work you did for this project.

3. Choose a problem or question that you had difficulty with in this project. Explain why you think it was difficult for you. Identify which analysis or problem-solving skill would apply to this particular problem.

4. Using the list of skills associated with problem solving and analysis below set a goal related to at least two skills that you will work toward in future coursework.

Skills associated with analysis:

- *clearly defining and understanding the problem (seeing the big picture)*
- *thoroughly collecting and organizing appropriate information*

- *making logical cause and effect connections*
- *sorting through and paying attention to the important details*

Skills associated with problem solving:
- *clearly defining and understanding the problem (seeing the big picture)*
- *thoroughly collecting and organizing appropriate information*
- *using knowledge and understanding of the system to deduce a solution*
- *using creative and novel ways to solve the problem*
- *continually evaluating how realistic your work and answer are*

(Used in chemistry and biology courses)

Example of Criterion-Based Self-assessment in Proof Writing

Criteria for this self-assessment:

Your solutions:

a. Are written in coherent English sentences
b. Integrate appropriate constructions and diagrams
c. Exhibit clear understanding of the problem, including distinguishing what is given from what is to be proved
d. Draw explicitly on definitions, axioms, theorems, and explicitly state dependence on assumptions, observations, or that which has already been proved
e. Exhibit clear and valid logical structure:
    - describe how constructions or diagrams incorporate hypotheses
    - describe how a statement follows from previous ones

1. Did you make improvements in your proof in response to the "refined criteria" and/or feedback based on these criteria?

2. Describe your goals for future proof work. Are you now consistently and easily meeting all criteria in your proofs? Are there particular areas (criteria or topics) you need to focus on?

Assessing Student Learning and Evaluating Faculty Capacity Development in the NSF-funded Regional Workshops Project NSF/DUE/ASA 0127725

Sumitra Himangshu* - *Macon State College*
Richard J. Iuli - *Empire State College*

OVERVIEW

This Assessment of Student Achievement (ASA) project (DUE 0127725) is a supplemental project, under the auspices of the NSF-funded Regional Workshops Project (RWP), designed to foster and assess student learning. The approach was to study both the faculty change process and the attendant student learning that resulted from an exposure to the environmental problem-solving-based program (EPS) conducted by the RWP. The study was a collaborative effort between a group at the Learning through Evaluation Adaptation and Dissemination Center (LEAD), University of Wisconsin (PI: Susan Millar; co-PI: Mark Connolly) and the University of Rochester (PI: Richard Iuli and Graduate Assistants: Sumitra Himangshu and Amy Cassata). This report focuses on the assessment of student learning conducted by the University of Rochester group. The main aim of this research was to assess change in student conceptual understanding as a result of attending a selected course taught by RWP faculty-participants.

PROJECT GENESIS

In the summer of 2001, an NSF-funded national dissemination project (DUE 008817) – the Regional Workshops Project (RWP) – began delivering a planned 15 regional workshops to 400 undergraduate Science, Technology, Engineering, and Mathematics (STEM) faculty over a five-year period. This ASA project was funded to examine the effects of that RWP on faculty capacity development and to assess student learning. The project's goals were to study the extent to which students in STEM courses: 1) learn STEM concepts in a meaningful way (as defined by Novak [1]); 2) construct a view of STEM disciplines that is consistent with views held by experts in those disciplines; and 3) construct integrative conceptual frameworks to facilitate understanding of STEM disciplines.

PROJECT IMPLEMENTATION

The research reported herein is part of the student assessment phase of the NSF-funded RWP. Assessing change in student conceptual understanding involved a two-step process: 1) using faculty and student interviews to generate concept maps which were used to assess change in student conceptual understanding, and 2) generating faculty expert concept maps to evaluate the impact of the EPS model on faculty-teaching methods that eventually impact student learning. The University of Rochester group, using the summer 2002 and summer 2003 faculty cohort participants for assessing the effectiveness of the Regional Workshops, began data collection on student learning. The following sections address the research design and results. Specifically, the results section addresses the following: 1) Assessment of faculty approaches to instruction and assessment of student learning, fall 2003 through spring 2005; 2) Conclusion of Student Assessment for the period fall 2003 through summer 2006; and 3) Report regarding the ongoing Faculty Assessment piece of the study between fall 2005 and summer 2007.

STUDY DESIGN

Our research design employed a mixed methods approach to collect and cross check assessment data. Both quantitative methods, such as course grades and results of the Learning and Studying Questionnaire (LSQ) [2] and qualitative methods, such as compiling classroom observational data using the Reformed teaching and Observation Protocol (RTOP) [3], and concept mapping based on faculty and student interviews, were used.

LEARNING AND STUDYING QUESTIONNAIRE (LSQ)

The Learning and Studying Questionnaire (LSQ) examines students' approaches to learning and studying. The LSQ consists of three sections, the first two of which contain items covering reasons for taking the degree program (learning orientations) and reasons for taking a particular course unit or module. The third section is an inventory which produces five scale scores (composites of several items) describing differences in students' approaches to learning and studying. These consist of information on deep/surface approaches to learning, study organization, effort, and concentration. For most of the items in the questionnaires, students respond on a 1-5 Likert scale (5=high).

The LSQ was administered prior to the first student interviews at the beginning of the semester. The rationale was to use the score differential between deep and surface learning in order to choose student participants from each class. This method minimized teacher and researcher biases and allowed the choice of mixed learners (those who reported using both deep and surface learning strategies equally) hypothesized to be most malleable to change within a short time (the duration of a semester).

STUDENT AND FACULTY CONCEPT MAPS

One year following faculty participation in the NSF-funded regional workshops, approximately ten faculty participants were selected from the pool of attendees for in-depth analysis of student learning in a selected course they taught. Case study faculty selection was based on institutional capacity for change, as measured by end-of-year interviews, class size, and institutional demographic diversity. Site visits were conducted at the academic institution of each case study faculty member at the beginning and at the end of a semester. During the first site visit, interviews were conducted with each faculty member for the purposes of identifying and making explicit key terms and essential concepts for students to comprehend upon completing the course. The faculty interviews were transcribed verbatim. The transcripts were used to develop an expert-level concept map.

Student interviews provided an opportunity for students to express their own ways of structuring the concepts they acquired and to give them the opportunity to choose the terms with which to relate and interpret their conceptual understanding. The questions asked provided an open-ended exploration of ***what*** and ***how*** the students think. All interviews were subsequently transcribed and the transcripts were used to generate concept maps. Training faculty and students to construct concept maps was not possible due to time constraints, and therefore warranted construction of concept maps by the researchers.

Between two and ten students from each course were selected for in-depth analysis of change in conceptual understanding over a semester. Because change in conceptual understanding can be influenced by individual differences in approaches to learning, as well as interaction with an environmental problem-solving curriculum, LSQ scores were obtained to provide a measure of individual learning approaches. Student selection was based on capacity for change, operationally defined as a lack of strong preference for either a surface-learning

or deep-learning approach indicated by self-reported LSQ scores on the use of both strategies to a similar degree. Each student selected participated in 30-minute interviews at the beginning and end of the semester. The interviews, based on concepts elicited in the faculty map, probed understanding of specific concepts and concept relationships. Student interviews were audio taped, transcribed verbatim, and concept maps were generated from these transcripts. After the completion of the final interview, students received compensation for their participation.

Data analysis of student concept maps addressed both ipsative comparisons (examining the change in quality of concept maps from the beginning to the end of semester for each student) as well as criterion-referenced comparisons (examining the similarity of the student maps to a faculty "expert" map at the beginning and end of semester). Maps were closely examined for accuracy by calculating the percentage of correct propositions in the faculty map that were also present in the student map. The percentages of correct propositions in the student map were compared at the beginning and end of semester to indicate a measure of growth in conceptual understanding. In addition, concept map scores were compared to student course grades as evaluated by the faculty.

CLASSROOM OBSERVATIONS

Classroom observational data were collected during one or more visits to classroom lecture and/or lab sessions. Tape-recorded lessons and field notes were analyzed to examine student behavior, teaching practices, content emphasized, amount of group and individual work observed, references to real-world applications, and level of communication both between students and between faculty and students. In the fall of 2004, an additional measure, the RTOP, was incorporated into classroom observations for purposes of triangulation. The RTOP is a criterion-referenced observational instrument that can be used to assess lesson design and implementation, propositional and procedural content knowledge, and classroom culture (including student-teacher interactions and communicative practices).

RESULTS

Assessment of faculty approaches to instruction and assessment of student learning fall 2003 through spring 2005

A total of eighteen sites (consisting of 13 faculty participants) were covered by the Rochester group between fall 2003 and spring 2005. Five of the eighteen sites were covered for two semesters each. Data from twelve of the eighteen sites were analyzed by constructing concept maps from 158 student interviews, consisting of a sample population of 79 students, out of a total of 129 students interviewed as part of the study. Overall, analysis of student concept maps indicated a strong correlation between student conceptual understanding (denoted by accuracy of organization of concepts in individual student maps) and students' approaches to learning as measured by the LSQ. However, within the different classrooms, the improvement in the quality and understanding represented in student concept maps was greater for faculty who largely used a hands-on, problem- solving approach to instruction.

Figure 1. Example of a student pre-map from cohort F showing a misconception regarding pH and acid rain. The different colors correspond to commonality of concepts with respect to cohort F's expert map (not included).

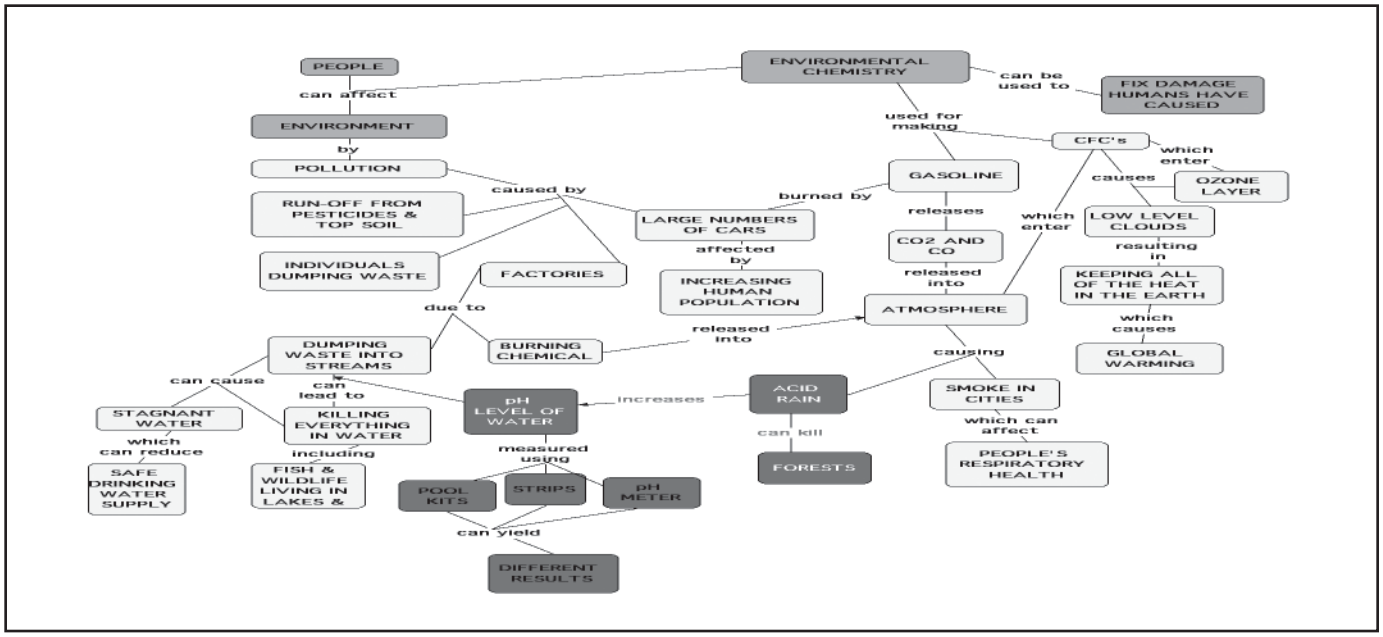


Table 1. Distribution of Student Learning Styles by Faculty

FACULTY	STUDENT	DISTRIBUTION	BY LEARNING STYLES
(#Number of Cohorts)	Deep Approach (%)	Mixed Approach (%)	Surface Approach (%)
Faculty A (2)	6	59	35
Faculty B (2)	23	77	0
Faculty C (2)	0	100	0
Faculty D (2)	0	70	30
Faculty E (1)	0	57	43
Faculty F (1)	0	50	50

Table 1 summarizes the distribution of student-learning strategies for different cohorts, as reported on the LSQ. Six out of ten cohorts had only mixed and surface learners (C, D, E & F), whereas both B cohorts had no surface learners and both A cohorts had all three groups of learners. Students in cohorts A, B, and C tended to use significantly fewer surface-level learning strategies since these cohorts had no surface-learners but consisted of deep and/or mixed learners (Table 1). Students in cohorts D, E, and F on the other hand, were evenly distributed by mixed and surface learning approaches and tended to use surface-level strategies to a greater degree. This correlated to the analysis of student post-maps for each cohort (Table 3).

Student map quality correlations, with individual approaches to learning denoted by the differential between the deep and surface LSQ scores, was significant ($n = 79$; $r = .865$, $p<0.001$), and provided an important basis and context for conceptual changes seen in students' concept maps between the beginning and the end of the semester. Faculty interaction and use of the EPS method to teach key STEM concepts, measured by the RTOP and classroom observations (Table 2), correlated strongly with the quality of student conceptual understanding as denoted by the number of concepts accurately represented between faculty expert concept maps (main ideas attained) ($n=79$; $r=.685$, $p<0.001$) and number of concepts accurately linked (quality) ($n=79$;

r=.865, p<0.001) of student concept maps for the total population (Figure 1). In addition, analysis of the quality of student understanding between highly interactive faculty (n=22; rh(4)=.92, p<0.001) and moderately interactive faculty (n=57; rm(8)=.75, p<0.001) was significant and indicated a robust correlation between student conceptual understanding and the interactive quality of the learning environment (Table 3).

Our data strongly suggest that construction of the faculty map had a positive correlation to faculty thinking regarding their teaching practices. Ten of the thirteen faculty observed from these two workshop cohorts have found the faculty map to be a positive influence on their teaching. They were excited about using it as a teaching tool since it helped them clarify what they needed to get across to their class. The faculty maps were based on broad areas targeted by each faculty and helped them develop a unified approach to the concepts they taught. Each class was examined on dimensions of instructional strategies and activities, integration of new concepts into familiar contexts, level of student-student and student-teacher communication, encouragement of active student participation, and level of support provided by teacher to enhance student investigations. As shown in Table 2, comparison of the classrooms on the variables of interest revealed that all faculty incorporated aspects of the EPS model to differing degrees in their teaching practice, by 1) connecting concepts to "real/local" issues, 2) providing a highly or moderately interactive setting where student input was encouraged, and 3) by emphasizing peer collaboration as an integral part of the course.

Our initial observations suggested a participant-effect between the pre- and post-site visits. We observed that five of the eight faculty followed in the first year of the study, clearly showed a marked increase in the interactive level of their instruction as denoted by the RTOP (data not shown). This resulted in a follow-up with five of the critical faculty between the two cohorts. All those visited twice, in-sequence, appeared to be more enthusiastic and willing to adapt the techniques proposed by the EPS model in their classroom. In addition, several of the critical faculty had expressed a desire to begin using concept maps for assessment purposes and were interested in collaborating with our team at the University of Rochester. *Thus, following the faculty beyond the first year was necessary and extremely important in making more meaningful interpretations of this observed participant effect.*

Table 2. Level of Faculty Interaction as Measured by Classroom Observations and RTOP

	Instructional Strategies	**New Concepts**	**Student to Student Communication**	**Teacher to Student Communication**	**Teacher Encouragement**	**Teacher Support**
Faculty A (Environmental Science)	Use of website	Related current events (Clean Air Act, EPA rules) to course content	Occasionally, teacher-facilitated	Prompted students to express ideas in class, worked with students on projects outside of class	Helped students continue and expand on thoughts, asking students to express personal views based on actual environmental data	Used several examples of actual data to clarify points (from EPA website)

	Instructional Strategies	New Concepts	Student to Student Communication	Teacher to Student Communication	Teacher Encouragement	Teacher Support
Faculty B (Environmental Science)	Use of website, whiteboard, video, and PowerPoint	Related current events (relationship between community and environment, emphasized merits of education vs. enforcement) to course content	Regularly, through discussions in class and in-class group projects	Relaxed but focused atmosphere, stimulated and facilitated discussion in and out of class	Encouraged students to express different viewpoints on environmental issues, guided students toward logical conclusions	Used several examples of actual data to clarify points (from USDA and Canadian agricultural department)
Faculty C (Biology)	Use of PowerPoint and blackboard	Related content with a variety of relevant examples	Regularly, through discussions in class and in-class group projects	Relaxed but focused atmosphere, stimulated and facilitated discussion in and out of class	Encouraged students to come up with their own answers, guided group discussion	Used examples relevant to everyday life (leaky mitochondria and runners' wall)
Faculty D (Biology)	Use of PowerPoint and blackboard	Used an inter-related framework to explain organ functions	Regularly through group lab projects	Relaxed atmosphere, encouraged questions	Encouraged students to use the web to find information and share during group work	Used examples (draws figures on board)
Faculty E (Analytical Chemistry)	Use of PowerPoint and blackboard	Related to standards for Chemical Society	Occasionally, teacher-facilitated	Focused atmosphere, faculty encouraged alternate modes of thinking	Participated as a resource person to guide student questions	Used several examples of actual data to clarify points (from ACA website)
Faculty F (Environmental Chemistry)	Use of website and blackboard	Related to local events (Paper Mill; Hurricane in NC)	Occasionally, teacher-facilitated and through group lab projects	Limited discussion during project work	Prompted students to express ideas regarding local environmental issues	Used several examples of actual data to clarify points (regarding local Paper Mill)
Faculty G (Environmental Science)	Use of PowerPoint and white board	Used examples and connected concepts	Faculty encouraged interpretation of evidence, alternate solutions	Relaxed atmosphere, faculty was patient and clarified student questions	Encouraged students to think aloud through their answers	Used local examples of pollution (ASARCO)

Student concept maps were compared with faculty concept maps to measure change in depth of understanding over a semester. Concept maps were analyzed for attainment of main concepts (as represented in faculty concept maps), presence or absence of misconceptions, tendency to make accurate connections between concepts, and level of conceptual organization. Table 3 gives a summary of preliminary results highlighting that the majority of students in each cohort attained the main ideas of the course as specified by individual faculty.

Table 3. Correlation of Student Post-maps with Respect to Faculty Teaching Style

Student Cohorts Belonging to Faculty:	Level of Faculty Interaction	Student Number (n)	% Students with Main Ideas	% Students with Misconceptions	% Students with Improved Organization	Average Gain in Terms (%)	Average Gain in Links (%)	% Surface Learners
B & C	High	22	86	34	62.5	29	22	0
A & D	Moderately High	27	96	24	70	16	13	33
E & F	Moderate	11	81.5	21.5	54	15	15.16	41.5

Analysis of student maps suggests that the quality of maps correlated to student approaches to learning and studying, as measured by the LSQ. Average gain from beginning to end of semester in number of terms and linkages between concepts was calculated. Student cohorts A and D (Table 3), largely using a mixed approach to studying, incorporating surface and deep strategies, showed the greatest improvement at 70percent in overall quality of conceptual understanding between pre- and post-test (Table 3). Consistently, maps from cohort E and F, using a greater degree of surface approaches to learning, showed only 54 percent of overall improvement in map quality (Table 3). Cohorts from the highly interactive faculty (B & C) that had no surface learners (Table 3) showed maximum gain in content and depth of understanding represented by the quality of their post-maps (Table 3). In addition, student concept maps were analyzed in comparison to faculty maps.

In short, the findings suggest that both for student learning and faculty development, our mixed method approach has resulted in an enriched assessment method to measure student conceptual change leading to meaningful learning. The importance of this as an assessment method to measure the effectiveness of adapting the RWP strategies in the classroom is enormous. In addition, the ease of transferability of this assessment method between disciplines makes it a powerful tool for undergraduate science.

SUMMARY OF ASA PROJECT AND FUTURE DIRECTIONS

This research concludes the three-year analysis of the Student Assessment piece of this project. Data from the Student Assessment piece indicates that student approaches to learning and faculty pedagogy are both important factors that influence student understanding. In addition, conceptual tools like concept maps, by providing a visual representation of knowledge organization, enable faculty to examine key processes that help students make meaning of STEM concepts and integration of the concepts into pre-existing knowledge bases. Faculty member checks also highlighted the differential between faculty expectations for their students and the reality of student understanding. The significant relation between student understanding and faculty interaction highlights the importance of designing assessment methods to match curricular changes and has broad implications for improving undergraduate science education.

3. Faculty Assessment: The analysis of faculty assessment by our collaborators in Wisconsin is ongoing and will be instrumental in triangulating this Student Assessment piece – this is a crucial piece that we expect will provide the basis for continued impact from the RWP. The NSF has generously considered our collaborator's request for an extension beyond the no-cost extension year (2005-2006) in order to enable a speedy completion of the Faculty Assessment piece in the current year.

It is important to note that as part of the dissemination aspect of this project, our findings have been well received at the following conferences:

1. Paper presented by Amy Cassata and Sumitra Himangshu at the First International Conference on Concept Mapping in September 2004.

2. Paper presented by Amy Cassata and Sumitra Himangshu at the National Association for Research in Science Teaching conference in April 2005.

3. Three papers were accepted at the European Science and Education Research Association 2005 conference in August 2005.

4. Paper presented by Sumitra Himangshu at the National Association for Research in Science Teaching Conference in April 2006.

5. Paper presented by Sumitra Himangshu at the Second International conference on concept mapping, September 2006.

6. Report of ASA findings by Sumitra Himangshu at the ASA-PI Dissemination Conference, October 2006.

7. Paper accepted at the European Science and Education Research Association 2007 conference in August 2007.

REFERENCES

[1] Novak, J.D., "Human constructivism: A unification of psychological and epistemological phenomena in meaning making", *The International Journal of Personal Construct Psychology*, vol. 6, no. 2, pp. 167-194, 1993.

[2] Entwhistle, N., McCune,V., and Hounsell, J., *Approaches to studying and the perceptions of university teaching-learning environments: Concepts, measures and preliminary findings*, ETL Project, Occasional Report 1, 2002. Available at: http://www.tlrp.org

[3] Swada, D., Piburn, M., Falconer, K., Turley, J., Benford, R., and Bloom, I., *Reformed Teaching Observation Protocol Technical Report* IN00-1, (Arizona State University), 2000.

Assessing Problem-solving Strategies in Chemistry Using the IMMEX System

Melanie Cooper - *Clemson University*
Ron Stevens - *University of California, Los Angeles*
Thomas Holme* - *University of Wisconsin, Milwaukee*

ABSTRACT

This paper outlines the development and utility of a problem-solving assessment system called IMMEX™ (Interactive MultiMedia Exercises). This system presents students with a complex problem to solve in an online environment. Student actions are tracked and data-mining strategies allow the clustering of many possible problem-solving pathways into specific strategy types. The system provides reliable and repeatable measures of student problem solving, which can be used to determine effective teaching strategies or to evaluate research studies in chemistry. Future developments that will allow the system to be part of a comprehensive assessment strategy within an undergraduate chemistry curriculum are also possible.

THE CHALLENGE

All stakeholders in science education efforts recognize the need for developing effective problem solvers; yet instructors continue to find it challenging to quantify students' strategic thinking in ways that can rapidly inform, and help modify instruction [1]. The current challenges for such assessments relate to the cognitive and non-cognitive differences among students, the difficulty of generalizing across problem-solving content domains, the design and development of appropriate tasks, and finally the speed, scale, and conceptual accessibility of assessment data.

Strategic problem solving is influenced by many variables, such as: students' prior knowledge and skill, cognitive and metacognitive abilities, task characteristics, gender, ethnicity, and classroom environment [2], as well as affective variables such as motivation and self efficacy [3]. An additional complication is that the acquisition of problem-solving skills is a dynamic process characterized by transitional changes over time as experience is gained and learning occurs [4].

While the challenges for developing assessments of problem solving are substantial, the real-time generation and reporting of metrics of problem-solving efficiency and effectiveness could fulfill many of the purposes for which educational assessments are used, such as grading and feedback for improving learning. [5,6]. In addition, such metrics could help target interventions for students, focus professional development activities for teachers, and influence training, implementation, and support decisions throughout school districts [7].

Effective and efficient problem solving in a domain requires both declarative (the 'what is') knowledge of the domain, as well as the ability to formulate and execute strategies (the procedural knowledge or the 'how to') to accomplish the goals posed by different problems. As expertise is developed there is a continual interplay between these two knowledge forms resulting in the development of refined representations or schema, which are defined as structures that allow problem solvers to recognize a problem state as belonging to a particular category of problem states that normally require particular moves. Such schema can serve as useful guides or roadmaps for future problem-solving situations.

In most undergraduate science courses the 'what is' is acquired through classroom instruction and outside readings, and with modern psychometric techniques, students' declarative abilities can be measured with some precision and reliability. Procedural knowledge is best gained through practice where elements of declarative knowledge can be combined and manipulated in various ways to solve common or not so common problems within the domain. In undergraduate courses these skills are often practiced in laboratory situations ranging from directed inquiry to complete inquiry. Procedural knowledge can be more difficult to assess than declarative knowledge, because determining if and how students failed at a complex task is equally as important as how they succeed.

THE IMMEX PROJECT

The IMMEX Project began in 1988 as a way to test the emerging diagnostic skills of medical students in ways other than Multiple Choice Testing, that is, as a way of promoting and assessing procedural skills. IMMEX problem solving follows the hypothetical-deductive learning model of scientific inquiry [8] where students frame a problem from a descriptive scenario, judge what information is relevant, plan a search strategy, gather information, and eventually reach a decision that demonstrates understanding.

The first IMMEX simulations presented patient diagnosis scenarios in cellular and molecular immunodeficiency [9] and were programmed in Turbo Pascal. These initial simulations had full database capabilities that could be queried by Structured Query Language (SQL) to provide a glimpse of not only if the student solved the problems, but more importantly, how they solved the problems [10].

The subsequent intellectual evolution of IMMEX reflects activities occurring along multiple parallel dimensions or strands relating to 1) grade and discipline expansion of the problem-solving concept, 2) the refinement of the visualization and data mining tools to better understand the idea of student problem-solving performance and progress, and 3) scale up of all aspects of IMMEX. The current ASA grant benefits from progress made by the IMMEX Project along all three dimensions.

Strand 1: Development and Implementation of IMMEX Simulations

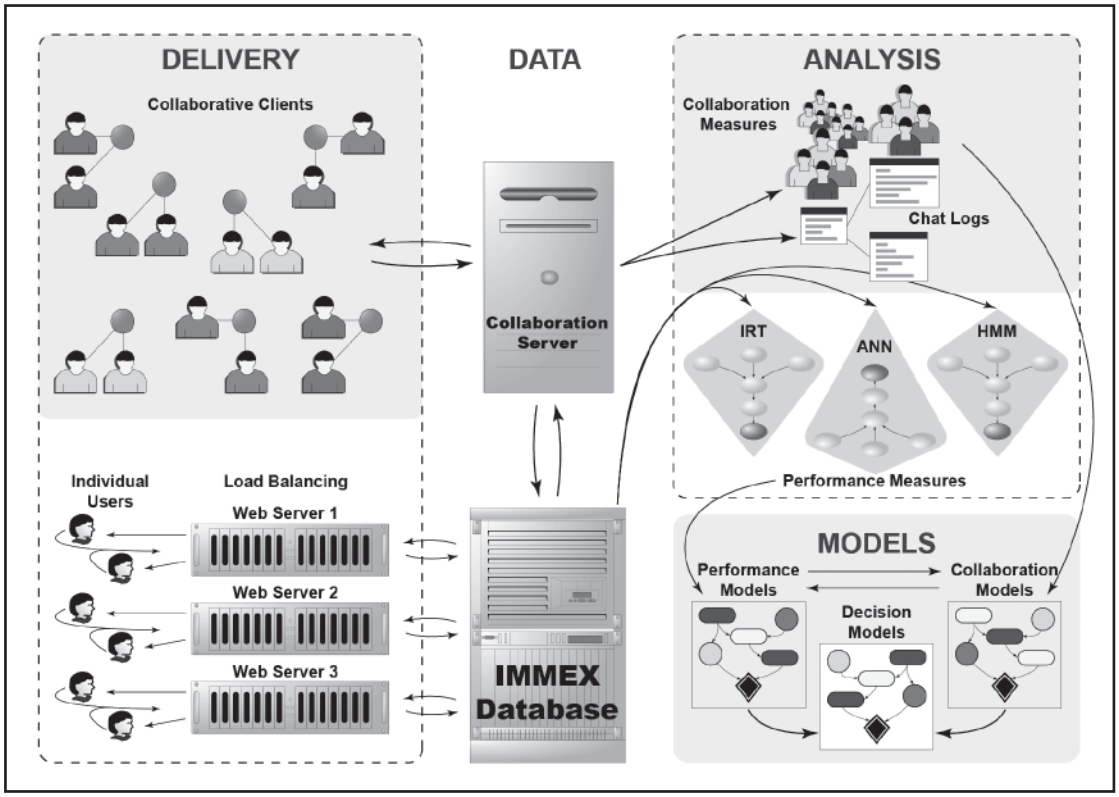
The first strand addresses the expansion of the problem-solving program from a medical school environment to one covering the K-16 educational continuum. Events in this strand include adaptations of the interface and presentation modes to accommodate the needs of different audiences as well as different forms of professional development activities to make the problems and data accessible to a wide range of audiences (students, teachers, parents, administrators). Activities in this strand also relate to the specific types of funding needed to support these events.

From a systems architecture perspective, IMMEX is a data-centric system centered around an SQL database containing problem components and performance data. It consists of Delivery, Data, Analysis, and Modeling components, connected with direct or web services communications. In the delivery component, options are currently available for 1) large scale (~400 concurrent users) individual problem solving and 2) for pilot testing (~20 concurrent student groups) for collaborative problem solving.

The IMMEX database server records the student performance data and the collaboration server records the student chat log. These are subsequently merged during the chat modelling process to associate chat

segments with the test selections. For collaborative studies, the collaboration client runs in a browser and is managed through Java applets that communicate with the IMMEX Collaboration Server. The Collaboration Server is an HTTP server acting as a proxy, that filters, edits, and synchronizes the IMMEX HTML pages through JavaScript, and sends them to the clients.

Figure 1. Overall Architecture for IMMEX delivery and Assessment Modules



The analytic models that provide the engine for suggesting interventions, focus on 1) effectiveness, as measured by Item Response Theory (IRT) analysis, and 2) strategies, as modeled by Artificial Neural Network (ANN) analysis and Hidden Markov Modeling (HMM). We have chosen to model both in real time, but in different software modules, as we think they may be assessing different constructs [11]. The analyzed data can then be propagated and integrated back into the decision models as described below, for providing or triggering interventions as needed.

Strand 2: Development of a Layered, Multidimensional Assessment System

The second strand traces the evolution of a layered system of data analysis and visualization techniques that facilitates the extraction of information from the ever-expanding dataset. We believe that the paths that students employ while navigating an IMMEX task provide evidence of a strategy, which we define as a sequence of steps needed to identify, interpret, and use appropriate and necessary facts to reach a logical conclusion or to eliminate or discount other reasonable conclusions. From these strategies a student demonstrates understanding by consistently, and efficiently deriving logical problem solutions.

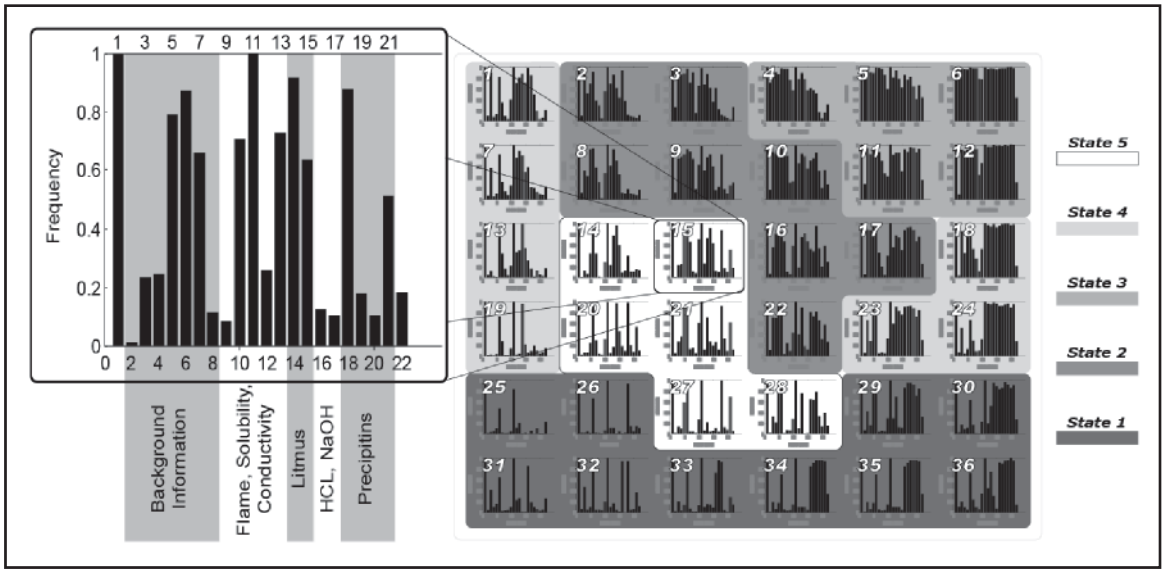
In addition to providing students with opportunities to engage in scientific problem solving, IMMEX supports detailed assessments of students' problem-solving skills and learning trajectories [12,13,14]. Although there are dozens of ways to approach and solve IMMEX problems, the cases are closed ended or well defined in that a problem is either solved or not solved. Each IMMEX "problem set" contains 6 to 50 "cases" of the problem with

the same interface and resources, but with different unknowns and supporting data [15]. IMMEX includes a layered system of analytic tools to dynamically model various aspects of students' problem-solving performance, including: 1) The strategic sophistication of a student at a particular point in time (a performance measure), 2) How the student arrived at this level (a progress measure), 3) How s/he will likely progress with more practice/experience (a predictive measure), 4) Whether this strategic level will be maintained (a stabilization measure), and 5) What instructional interventions could most effectively accelerate each student's learning. And then, of course, comes the next challenge of how to generalize such assessments across domains and educational systems [14].

IMMEX modeling utilizes serialized timestamps of how students use the resources that are available in an individual problem set, and across the problem's cases. For each problem, models are formed in real time based on 1) estimates of student ability, 2) the strategies used, and 3) estimates of future performance. These metrics are generated through IRT analysis, ANN analysis, and HMM, respectively [12,13,14]. For IRT analysis the problem difficulties are first estimated for all cases using the solve rates from a large number of student performances. Then, using this model, the ability of each student is estimated based not only on whether or not the case was solved, but also the relative difficulty of the case.

As students solve IMMEX cases, the menu items selected are also used to train competitive, self-organizing ANN analyses [16,17]. Self-organizing maps learn to cluster similar performances in such a way that neurons near each other in the neuron layer respond to similar input vectors [18]. The result is a topological arrangement of performance clusters where geometric distance between these clusters becomes a metaphor for strategic similarity. We frequently use a 36-node neural network and train with between 5,000 and 10,000 performances derived from students with different ability levels and where each student performed at least 3-4 cases of the problem set [19]. The components of each strategy in this classification can then be visualized for each of the 36 nodes by histograms showing the frequency of items selected (Figure 2).

Figure 2. Sample Neural Network Nodal Analysis for Identifying Strategies. The selection frequency of each action (identified by the labels) is plotted for the performances at node 15, thus characterizing the performances for this node and relating them to performances at neighboring nodes. The nodes are numbered in rows, 1-6, 7-12, etc. This figure also shows the item selection frequencies for all 36 nodes [12,13,14].



Most strategies defined in this way consist of items that are always selected for performances at that node (i.e. those with a frequency of 1) as well as items that are ordered more variably. For instance, many Node 15 performances shown in Figure 2 (left side) contain the items13-15, whereas few contain items 16 and 17 or 19 and 20. Also shown is a composite ANN nodal map, which illustrates the topology generated during the self-organizing training process. Each of the 36 graphs in the matrix represents one node in the ANN, where each individual node summarizes groups of similar students' problem-solving performances automatically clustered together by the ANN procedure.

As IMMEX problem sets contain many parallel cases, learning trajectories can be developed through HMM that not only reflect and model students' strategy shifts as they attempt series of cases, but also predict future problem-solving performance [12].

Prior analyses indicate that many students begin by selecting many test items as they attempt to solve IMMEX problems. Consistent with models of skill acquisition [20], over time most students refine their strategies and select fewer tests, and eventually stabilize with a preferred approach often used on subsequent problems. As expected, with practice, students' solve rates increased from 35 percent to 63 percent ($\chi 2 = 121.8$, df=10, $p<0.000$). The rate of stabilization, and the strategies stabilized with are influenced by gender [21], experience [12], and individual or group collaboration [22], etc. Students often continue to use these stabilized strategies for prolonged periods of time (3-4 months) when serially re-tested [14]. Significant variability of strategy usage is also seen across teachers and classrooms, and detailed videotape analysis of instruction suggests that the ways teachers represent the task to the students has a major effect on the problem-solving outcomes [23].

Strand 3: Scaling Up and Scaling Out

The third strand involves scale and intersects in different ways with the other two strands. Part of the current ASA effort relates to this Strand.

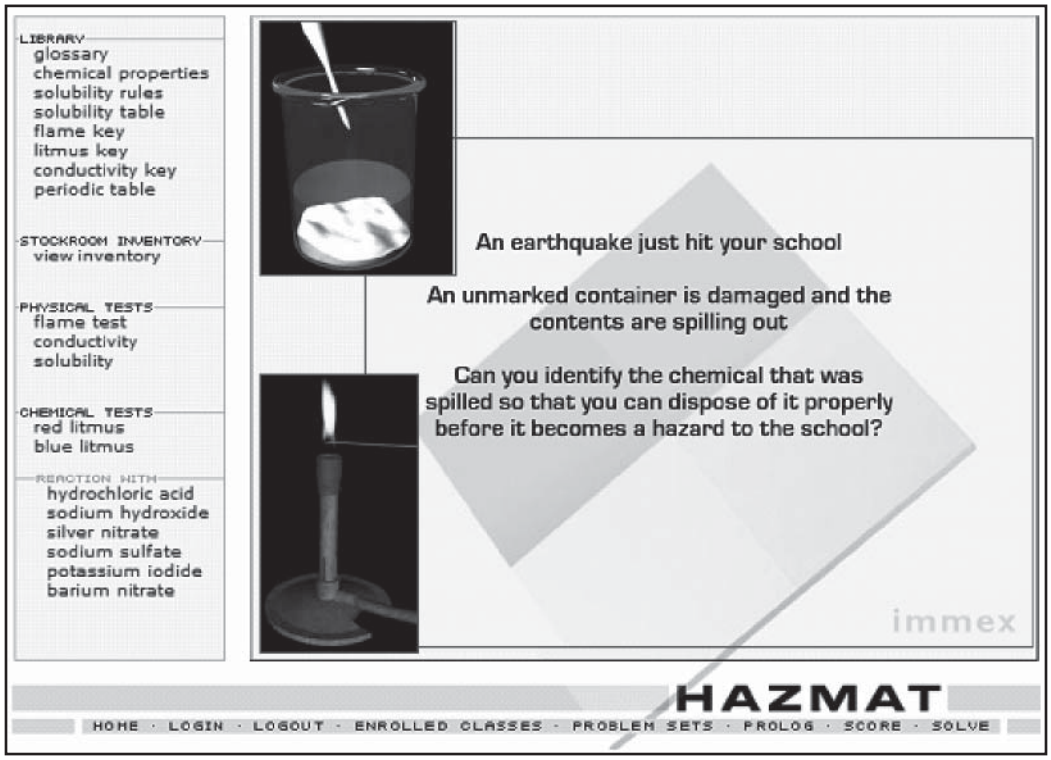
Strand 4: Implications for Teaching and Learning

The development of IMMEX problems and the suite of assessment tools that accompany them has profound implications for both the summative and formative assessment of student problem solving.

SUMMATIVE ASSESSMENT

We now have over thirty IMMEX problem sets covering topics in general and organic chemistry (see http://chemed.ces.clemson.edu/chemed/IMMEXWorkbook.pdf for a workbook detailing the problems in these subjects. There are also many other problems in other fields.) and it is now possible to rapidly assess student performance on complex open-ended problems. An example of a typical IMMEX problem is given in Figure 3. This problem, named Hazmat, is a qualitative analysis problem, in which students must choose appropriate tests to identify an unknown. The results of the tests are presented as animations, so that the student has to understand what the results mean, and act accordingly.

Figure 3. The prolog for Hazmat, a qualitative analysis problem.



The ability to assess complex problem-solving abilities should have a major impact on teaching and learning, because it is a truism that students will learn in response to how we assess them, or to put it another way, assessment often drives the curriculum. If we assess students with multiple-choice problems that emphasize recall and simple one-step problem solving, students will not learn the skills which are so valued in scientists; that is synthesis evaluation and critical thinking. Most online assessment systems grade only the answer to the problem, and even though many now allow free-form input of answers, most systems are still very limited in the complexity of the problem that can be posed. In fact, the systems that pose a different problem set to each student actually encourage algorithmic problem-solving behaviors, since all that changes from problem to problem is the numbers. We now have available to us a system that will allow educators to assess students' complex problem-solving abilities and strategies, while monitoring the strategy that the student chooses to navigate through the problem. The output is seamless and can be used to assess the achievement of the student in a particular topic or course.

FORMATIVE ASSESSMENT

The ongoing assessment and feedback capabilities of IMMEX are probably even more important than the end-of-course type of traditional exam or test. IMMEX provides us with tools to probe how a student solves complex problems, whether that student changes strategy over time, and whether the student actually improves in ability over time. It can give us insight into how students solve problems and it can tell us whether an intervention or teaching strategy is effective in producing improvements in ability and strategy choice. We can measure the effects of interventions and have direct evidence that student problem-solving behaviors can be affected by relatively minor interventions. We can use IMMEX to measure improvements in problem solving and provide feedback and support to students who are not progressing well.

RESEARCH RESULTS

- *Students stabilize at a particular ability level and strategy after about five attempts at a problem*

We have found that a student will often stabilize on a problem-solving strategy (as modeled by ANN and HMMs, [12,24,25]) and at a particular ability level (from IRT, [22,26]) by the time she has performed five problems from that particular problem set. That is, even if the student is not improving, and is using an inefficient and/or ineffectual problem-solving strategy, that student will rarely change strategy once they have stabilized. Nor will the student improve in ability. This finding seems to belie the commonly held belief that if only the student would work harder or longer, then they would improve.

- *Student problem-solving ability and strategy choice can be changed permanently by interventions such as collaborative learning*

One area of our research that has profound implications for teaching and learning is the investigation of interventions to determine whether we can perturb the stabilization levels of problem-solving strategy and ability. For example, we investigated the effect of collaborative grouping on students' problem-solving strategies and abilities [22,25,26]. This research was performed with 713 students enrolled in the first semester of a general chemistry course for science and engineering majors.The experimental design was a modified pre-test, intervention, post-test design, in which students were asked to perform five or more IMMEX problems to provide a baseline stabilization of strategy and ability. This was followed by the intervention in which students worked in pairs on five more problems, and then the "post-test" in which students again worked five or more problems individually. In this way we were able to identify any changes that took place as a result of grouping, and whether these changes remained after grouping.

Figure 4. Strategies used by individuals who have stabilized, before and after grouping. Key shows strategies 1-5 resulting from ANN analysis followed by HMM of the output. The strategies are significantly different (Chi square = 227, p<0.001)

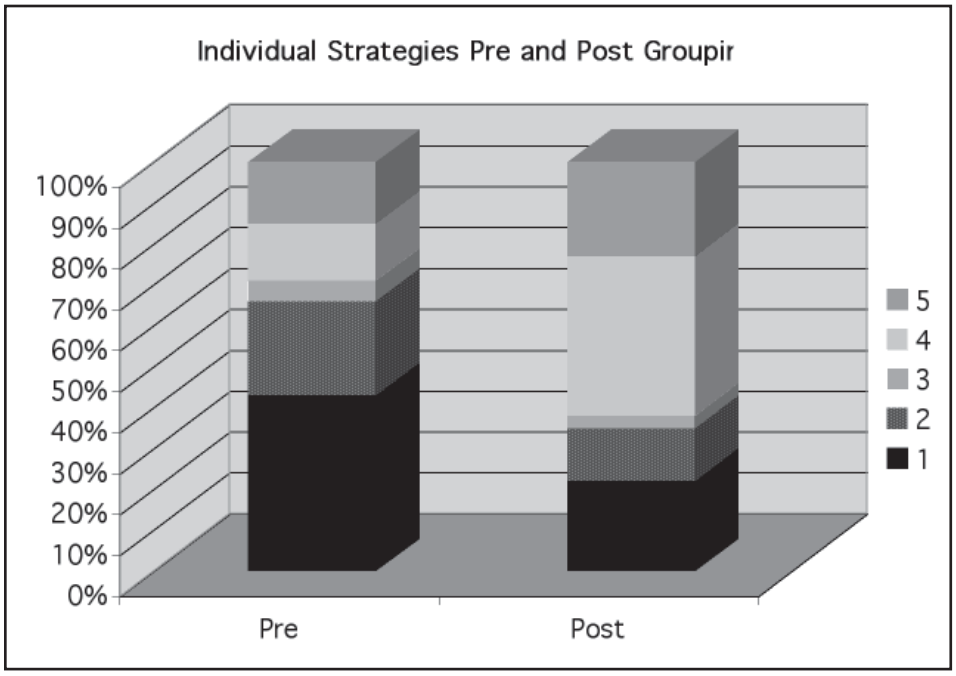
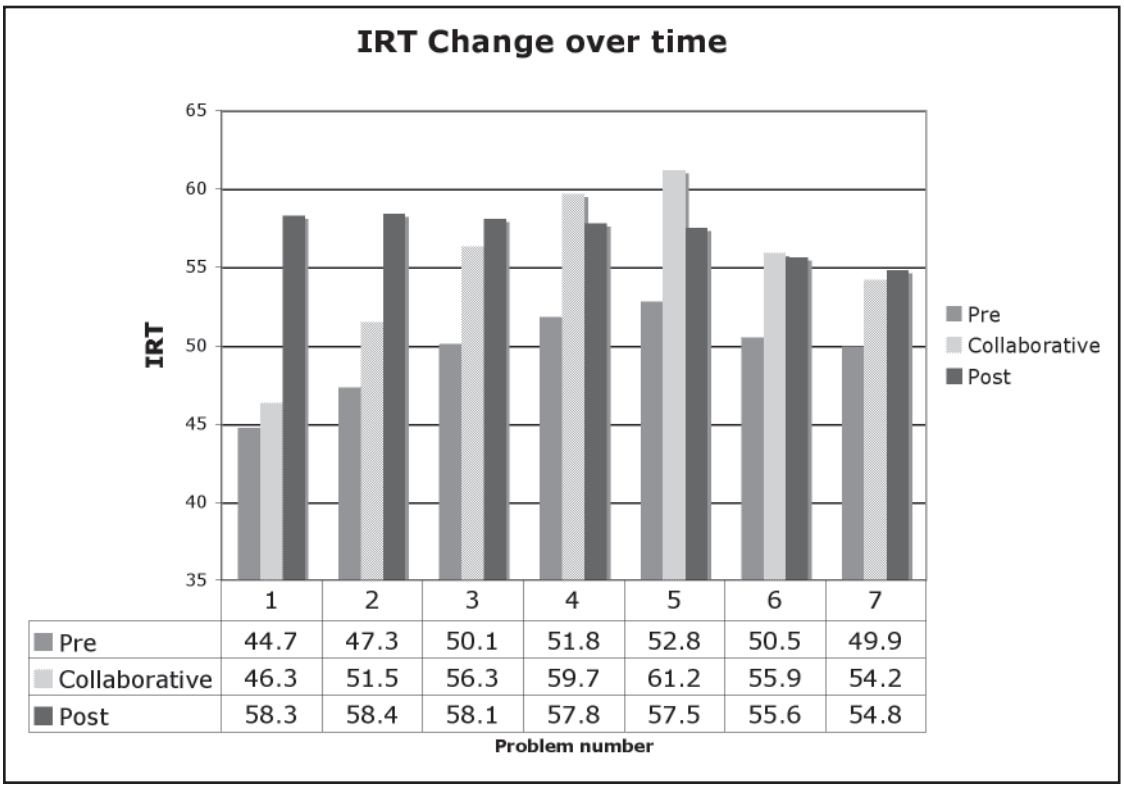


Figure 5. A comparison of student abilities for pre-collaborative, collaborative, and post-collaborative performances. The table shows the actual abilities from IRT analysis, which are reported on a scale ranging from 20-80.



	1	2	3	4	5	6	7
Pre	44.7	47.3	50.1	51.8	52.8	50.5	49.9
Collaborative	46.3	51.5	56.3	59.7	61.2	55.9	54.2
Post	58.3	58.4	58.1	57.8	57.5	55.6	54.8

Although the results reported here are from a study involving 713 students, we have now performed this type of experiment with literally thousands of students, and the results are always that students tend to change strategy and increase in ability during grouping, and that the strategy changes and ability improvement remain after the intervention, as shown in Figures 4 and 5.

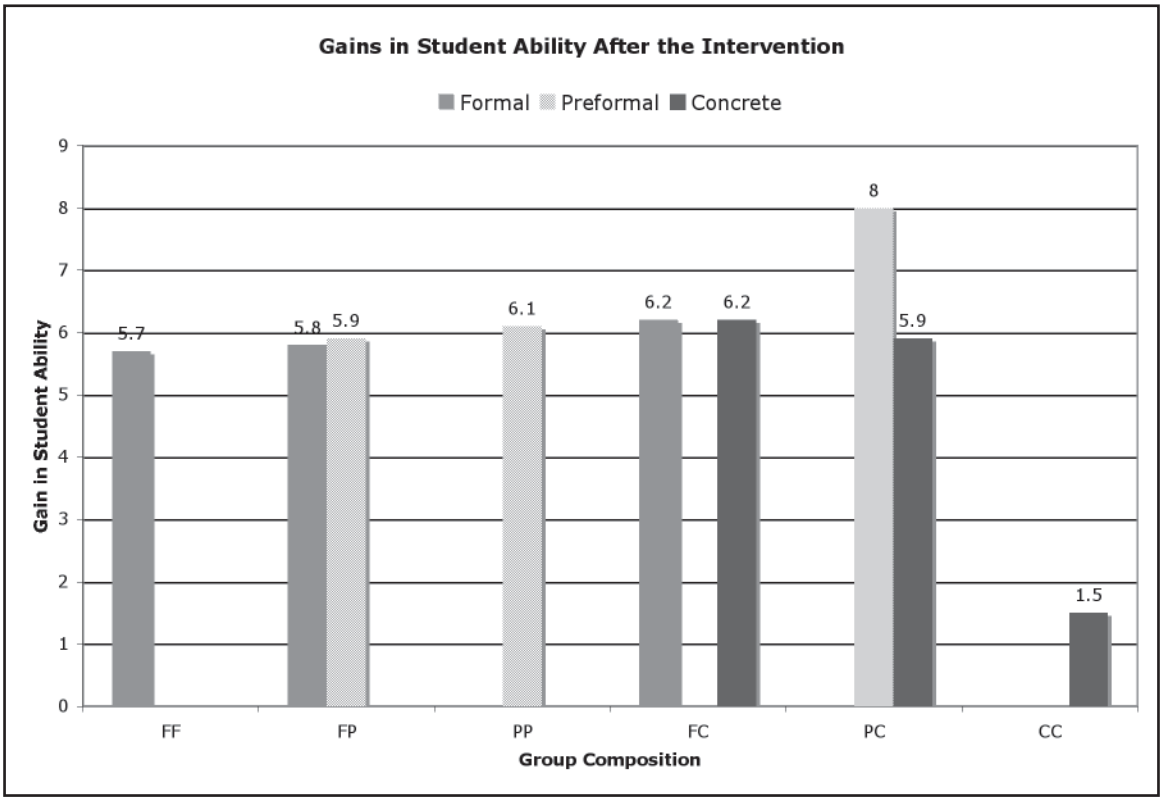
Thus, one intervention consisting of a one- to two-hour problem-solving session with an unstructured group, had the effect of improving the students' ability by about 6 IRT ability units, or about 10 percent. This is direct evidence that an unstructured group problem-solving session can be beneficial. The results indicate that for most students, some part of their time is much better spent solving problems in a group setting rather than plugging away at a problem set alone. This is not to say that group work can substitute for individual effort, but clearly it is beneficial for most students.

- *The student and group makeup can affect improvements*

Another interesting and important finding from this research was that student improvement depends upon the makeup of the group. In these experiments students were paired with another student by their level of logical thinking, as determined by the GALT test. Students were classified as Formal, Transitional, or Concrete, and paired up in all possible combinations. Remarkably, the makeup of the group did not seem to affect the improvement much, except for two types of combination. For most students the improvement is about 10 percent, or 5-6 IRT ability units, as can be seen from Figure 6, which shows the improvement for students in IRT ability units

(which generally range from 20-80 with an average in our studies of 50 units). However, if a transitional student was paired with a concrete student, the improvement was much greater, while if two concrete students were paired there was no significant improvement at all.

Figure 6. Gain in student ability for each type of group and the thinking level of the students in the group. The gain is statistically significant for every group at the p<0.001 level, except for the C-C group.



PROGRAM ASSESSMENT AND THE FUTURE OF IMMEX IN CHEMISTRY

A growing number of educational reports address the need to consider assessment within higher education. The recent report from the Spellings Commission on Higher Education [27], for example, calls for improved accountability within higher education, among several recommendations. As in many disciplines, chemistry departments are searching for ways to implement enhanced assessment prerogatives. In addition to pressure arising from accreditation and national educational policy directives, the process by which the American Chemical Society approves undergraduate chemistry curricula for students is undergoing change as well [28]. This development also places additional impetus on chemistry departments to establish the efficacy of their educational efforts for outcomes-based reporting.

The American Chemical Society, through its Division of Chemical Education, also provides the chemistry education community with assessment materials on a national scale via the Examinations Institute, currently housed at the University of Wisconsin – Milwaukee. The Exams Institute produces exams for most undergraduate chemistry courses through a committee writing process [29]. These exams, however, tend to present snapshots at the end of a course, rather than a measure across sub-disciplinary boundaries of student growth.

Thus, one envisioned pathway for the expansion of IMMEX is to provide a means for chemistry departments to measure student problem-solving strategies in several courses of the undergraduate curriculum. Comparison of how those strategies change with increased exposure to content knowledge and skills development associated with the undergraduate curriculum would provide a measure of growth. Insofar as the strategy measures made possible by IMMEX assess generalized knowledge (not constrained by the sub-discipline of chemistry, in which the problem is set) the use of IMMEX potentially provides a powerful tool for measuring growth of knowledge. Efforts are currently underway to establish the requisite concepts of skill-transfer in a specific category of chemistry problems – those involving the relationship between structure and function of molecular systems. Transfer is a cognitive issue with significant complications [30,31], but the idea of measuring strategies for students as they progress in the chemistry major is straightforward in the context of IMMEX. Thus, as more IMMEX problems are staged for later courses in chemistry, the prospect that meaningful metrics for measuring growth in the sophistication of problem-solving strategies remains.

SUMMARY

This paper has shown that within a complex, online environment where student actions can be followed, it is possible to derive a large database of student performances on a class of chemistry problems and subsequently use data-mining methodologies to establish general strategies that are represented among all the thousands of pathways for solving the problem. The IMMEX system that was devised to carry out this type of measurement has established that valid and reliable measures of problem-solving strategies can be made within this environment. Chemistry education research has shown that the IMMEX system can make measures of student strategies and then be used to determine if teaching interventions are helpful in improving those strategies. This methodology shows promise for future expansion to allow the IMMEX system to be incorporated as part of an overall assessment plan for an academic chemistry department that wishes to measure the change in student problem-solving skills over time.

ACKNOWLEGMENTS

This project has been funded by several grants from the National Science Foundation (DUE- 01236050, REC-02131995, NSF-ROLE 0528840, HRD-0429156, DUE-0512526), whose support is gratefully acknowledged.

REFERENCES

[1] Pelligrino, J.W., Chudowsky, N., and Glaser, R., *Knowing what students know: The science and design of educational assessment*, (National Academy Press, Washington, D.C.), 2001.

[2] Fennema, E., Carpenter, T.P, and Jacobs, V.R. , "A longitudinal study of gender differences in young children's mathematical thinking," *Educational Researcher*, vol. 27, pp. 6-11, 1998.

[3] Mayer, R.E., "Cognitive, metacognitive, and motivational aspects of problem solving," *Instructional Science*, vol. 26, pp. 49-63, 1998.

[4] Lajoie, S.P., "Transitions and trajectories for studies of expertise," *Educational Researcher,* vol. 32, pp. 21-25, 2003.

[5] Bennett, R.E., "Reinventing Assessment. Speculations on the Future of Large-Scale Educational Testing. A Policy Information Perspective," *Educational Testing Services*, 1998, Available on ERIC, ED424254.

[6] Atkin, J.M, Black, P., and Coffey, J., *Classroom Assessment and the National Science Education Standards,* (National Science Foundation, Arlington, Va.), 2001.

[7] Spillane, J.P., Reiser, B.J., and Reimer, T., "Policy implementation and cognition: Reframing and refocusing implementation research, " *Review of Educational Research*, vol. 72, pp. 387-431, 2002.

[8] Lawson, A.E., *Science Teaching and the Development of Thinking,* (Wadsworth Publishing, Belmont, Calif.), 1995.

[9] Stevens, R.H., Kwak, A.R., and McCoy, J.M., "Evaluating preclinical medical students by using computer-based problem-solving examinations,"*Academic Medicine*, vol. 64, pp. 685-687, 1989.

[10] Stevens, R.H., McCoy, J.M., and Kwak, A.R., "Solving the problem of how medical students solve problems," *MD Computing*, vol. 8, pp. 13-20, 1991.

[11] Stevens, R.H. and Thadani, V., "A Bayesian network approach for modeling the influence of contextual variables on scientific problem solving," Lecture Notes in Computer Science, vol. 4053, pp. 72-84, 2006.

[12] Stevens, R., Soller, A., Cooper, M., and Sprang, M., "Modeling the development of problem-solving skills in chemistry with a web-based tutor," *Lecture Notes in Computer Science*, vol. 3220, pp. 580-591, 2004.

[13] Stevens, R.H., Johnson, D.F., Soller, A., "Probabilities and Predictions: Modeling the Development of Scientific Competence," *Cell Biology Educ*., vol. 4, pp. 42 – 57, 2005.

[14] Stevens, R.H., "Constructing Machine Learning Systems for Quantifying Student's Scientific Problem-solving Efficiency and Effectiveness," *Technology in Cognition and Instruction,* in press.

[15] Stevens, R., Palacio-Cayetano, J., "Design and performance frameworks for constructing problem-solving simulations," *Cell Biology Education*, vol. 2, pp.162–179, fall 2003.

[16] Stevens. R., and Najafi, K., "Artificial Neural Networks as Adjuncts for Assessing Medical Students' Problem-solving Performances on Computer-based Simulations," *Computers and Biomedical Research,* vol. 26, no. 2, pp. 172-187, 1993.

[17] Stevens, R., Wang, P., Lopo, A., "Artificial neural networks can distinguish novice and expert strategies during complex problem solving," *JAMIA*, vol. 3, no. 2, pp. 131-138, 1996.

[18] Kohonen, T., *Self Organizing Maps*, 3rd extended ed., (springer, New York), 2001.

[19] Stevens, R., and Casillas, A., "Artificial Neural Networks," in Mislevy, R.E., Williamson, D.M., and Bejar, I. (eds.), *Automated Scoring*, (Lawrence Erlbaum, Mahwah, N.J.), 2006.

[20] Ericsson, K.A., "Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains," *Academic Medicine*, vol. 79, pp. S70-S81, 2004.

[21] Stevens, R.H., and Soller, A., "Machine learning models of problem space navigation: The Influence of Gender," *ComSIS*, vol. 2, no. 2, pp. 83-98, 2005.

[22] Cooper, M.M., Cox, C.T., Nammouz, M., Case, E., and Stevens, R., "An Assessment of the Effect of Collaborative Groups on Students' Problem-solving Strategies and Abilities," *J. Chem Education*, submitted.

[23] Thadani, V., and Stevens, R.H., "Teaching Practices and Their Relationship to Students' Scientific Problem-solving Skills", submitted.

[24] Case, E.L. Cooper, M.M. Stevens, R.H., "Is Collaborative Grouping an Effective Instructional Strategy," *J. Coll. Sci. Teach.*, vol. 36, pp. 42-47, 2007.

[25] Nammouz, M. Ph.D. Dissertation, Clemson University, 2005

[26] Cox, C. T. Ph.D. Dissertation, Clemson University, 2006

[27] U.S. Department of Education, *A test of leadership: Charting the future of US higher education*, 2006. Available at http://www.ed.gov/about/bdscomm/list/hiedfuture/index.html

[28] American Chemical Society, "Proposed Revision to the ACS Guidelines," 2006, available online at http://www.chemistry.org/portal/a/c/s/1/acsdisplay.html?DOC=education\cpt\Proposed_Revision_to_the_ACS_Guidelines.html

[29] Holme, T.A., "Assessment and Quality Control in Chemistry Education," *J. Chemical Education*, vol. 80, pp. 594-598, 2003.

[30] Catrambone, R., "Generalizing Solution Procedures Learned From Examples," *J. Experimental Psychology,* vol. 22, pp. 1020-1031, 1996.

[31] Sohn, Y.W., Doane, S.M., and Garrison, T., "The impact of individual differences and learning contexts on strategic skill acquisition and transfer," *Learning and Individual Differences*, vol. 16, pp. 13-30, 2006.

Ed's Tools: A Web-based Software Toolset for Accelerated Concept Inventory Construction

Kathy Garvin-Doxas, Isidoros Doxas, and Michael W. Klymkowsky*
The University of Colorado, Boulder

ABSTRACT

Ed's Tools is a suite of web-based programs designed to facilitate the research that underlies, and the construction of, concept inventory (CI) questions. Ed's Tools was developed for the Biology Concept Inventory (BCI) project, but has since been used in different fields, from engineering to space physics. It facilitates the collection, coding, and analysis of text-based student data, as well as the synthesis of the results. Multiple, geographically separated coders can analyze and share data, and the results are available in a simple database of coded student responses. Ed's Tools is freely available via the BCI website, bioliteracy.net.

INTRODUCTION

For the last fifteen years or so, a revolutionary transformation has been taking place in the way Physics is taught at the college level. Physics courses at all levels, from the standard introductory course for non-science majors to upper-level courses for physics majors, are being transformed from primarily lecture-based to a more "interactive engagement" style of instruction. At some schools this has led to the complete redesign of the classroom layout, while at others the emphasis has been more on the introduction of student response systems ("clickers") that facilitate real-time interactions between instructors and students. Finally, a number of integrated instructional packages have been published [1,2] that combine meticulously researched materials and tightly scripted instructional methodologies.

The revolutionary transformation of the way physics is being taught can be traced in large measure to a particular catalytic event: the development of the Force Concept Inventory (FCI) [3]. Concept inventories (CIs) are validated and reliable multiple choice instruments that explore students' conceptual understanding in a given subject area; they provide researchers with a common instrument that can be used to measure the improvement in student achievement after a single course. More general tests, like the SAT or the GRE, have long been used as measures of achievement, but they are inappropriate for evaluating the change in student conceptual knowledge after a single course, and in general are not designed to specifically measure student mastery of particular ideas. Armed with pre- and post-instruction FCI data, Hake [4] found that courses that use collaborative learning techniques (referred to generically as "interactive engagement") have learning gains that are often twice those seen in classes that use only traditional lectures. This study is commonly cited as the impetus driving the subsequent transformation of a large, and ever increasing, number of physics courses.

Spurred on by the impact of the FCI, a large number of efforts have begun to create similar instruments for other subject areas. A non-exhaustive list includes the Astronomy Diagnostic Test [5], the Thermodynamics Concept Inventory [6], the Natural Selection Concept Inventory [7], the Fluid Mechanics Concept Inventory [8], the Geoscience Concept Test [9], the Basic Biology Concept Inventory [10, 11], the Materials Concept Inventory [12], and the Chemistry Concept Inventory [13]. At the same time, it is worth noting that the impact of these various concept inventories on their respective target fields has been rather less dramatic than that of the FCI. This raises

a practical question of what factors determine whether a concept inventory has a significant impact on teaching and learning within its target area. Although a thorough discussion of this question is beyond our scope here, one can speculate as to the factors that lead to the widespread effects of the FCI. These include the championing of FCI data by prominent educators such as Eric Mazur and Richard Hake as a valid reflection of learning. In addition, one might argue that physics, as a discipline, is distinct in the near universal acceptance of key ideas (e.g. Newton's laws of motion and Maxwell's equations) as organizational principals in the introductory course sequence. In others areas, such as biology, finding a similar consensus upon which to focus conceptual assessment appears to be much more problematic. In our own experience, much time and effort is required simply to identify the foundational concepts to be evaluated.

That said, even with clear conceptual targets, CI research, construction, validation, and reliability requires considerable time and effort. The creation of the FCI [3] took almost a decade, not withstanding the fact that it addresses an extremely limited and well-accepted cluster of concepts in classical physics, namely Newton's laws of motion. Moreover, no substantial advance has been made in the amount of time, effort, and expense required to develop a validated, reliable, relevant, and widely accepted instrument. Precisely because of the effort required to produce them, existing instruments are jealously guarded to protect their validity. The FCI, for instance, although freely available, makes a point of admonishing users to not include its laboriously researched questions in any for-grade exams lest students' very efficient "answer dissemination" networks make the answers widely available, and so significantly degrade the validity of a very expensive instrument. Even so, as the content of a particular instrument becomes widely known, there is a serious possibility that instructors will, often subconsciously, begin to "teach to the test."

CONCEPT INVENTORIES VERSUS STANDARD TESTS

While an experienced teacher may be able to assess a student's understanding of a particular concept through a relatively brief one-on-one examination, this approach is not feasible for the assessment of a large number of students. In the same way, essay-type exams are equally time consuming to develop and to analyze objectively, and it is often the case that such tests are "inauthentic" in the sense outlined by McClymer and Knoles [14]. The same reasoning that underlies the use of double blind, placebo-controlled trials in the biomedical sciences also applies to educational assessment. Instructors have a vested interest, both personally and professionally, with regards to the evaluation of student learning – they cannot be assumed to be objective, no matter how hard they try.

Developing and validating reliable assessment instruments is an arduous and time-consuming task, qualitatively different from that involved in the generation of a standard exam. In the case of a multiple-choice type instrument, there are two main obstacles involved in designing a valid, reliable instrument: the first is the design of the questions and the second is the choices supplied to students. For example, the question cannot be so detailed and jargon-laden that it artificially limits responses, whereas the correct choice cannot provide irrelevant "clues" that flag it as correct. Similarly, the incorrect choices, known as distractors, should reflect commonly held student misconceptions – such distractors lure students away from the correct answer. In contrast, obviously wrong distractors degrade the instrument's validity, since they improve the odds of guessing the correct response, a phenomenon known as "construct irrelevant easiness".

Once questions and choices have been generated, it is critical that the instrument be validated, to avoid false readings [15]. This involves extensive interviews, to ensure that students who score highly actually understand the concepts embedded in the test, while those who do not (and picked the distractors) actually hold the misconception represented by the incorrect answers. By repeated trial-and-error, during which the exact wording of each question and answer is finely tuned, an experienced team of content and assessment experts can increase the validity of a test score inference. Reliability has to do with the instrument's ability to consistently produce the same result when administered to the same population (independent, internal consistency measures like Cronbach's alpha, are used to measure the reliability of instruments) and can be degraded when the distractors are arbitrary, thus introducing random elements into subjects' responses. As with validity, an experienced team of content and instrument experts can produce a reliable instrument after only a few iterations.

CONCEPT INVENTORY CONSTRUCTION AND ED'S TOOLS

Developing questions, answers and distractors for a concept inventory has to be repeatedly anchored in students' conceptual understanding of the subject (Figure 1). It follows, rather rigidly, a pathway that ensures the adherence of the development process to what the students think, as opposed to what the developers (or content experts) believe they think. These steps consist of:

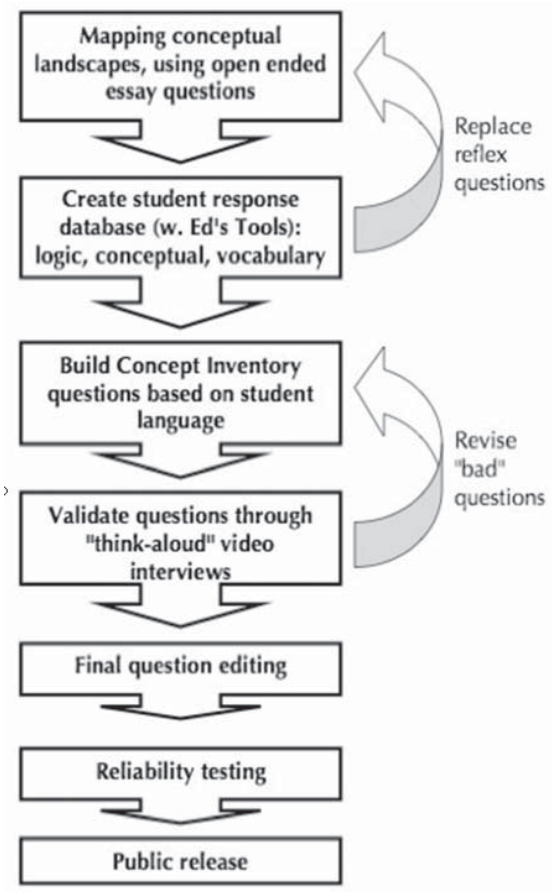
1) Assigning essay questions to students. These are typically board questions designed to elicit open-ended responses that represent the spectrum of conceptual understanding of the responders on the subject matter. These responses are the start of a process of mapping the students' conceptual landscape, which is often poorly appreciated by instructors and experts in the subject matter.

2) Analyzing language of respondents. The essays are coded (not "graded") for all concepts, both correct and incorrect, present in the answers. This "catalog of concepts" provides the raw material for questions, answers, and distractors.

3A) Using student language to develop candidate questions and answer choices. This intermediate step leads to the use of largely verbatim quotes of respondents' language as the possible answers in the development versions of the inventory. It accelerates the production time by allowing the developers to quickly isolate the type of language that elicits valid responses from the respondents.

3B) Producing written explanations for the purpose behind each question, and the concept each answer is meant to probe. This is a critical step, which requires coders to make explicit the connection between the question or answer they propose and the coded concepts.

Figure1. The Development Process



4) Conducting validity tests. Validity testing involves a variety of standard methods, like interviews, both individual and group, think-alouds, etc. This process enables the interviewer to gauge how well the question and answer represent the student's conceptual understanding.

5) Refining questions and answers. The objective of this step is to produce a valid list of prevalent concepts, not to score correct answers.

6) Conducting reliability testing. This step consists of administering the candidate questions to a sample population and using standard statistical techniques to measure reliability.

7) Repeating the process (steps 3-6) until some criterion for reliability convergence is reached.

It was to facilitate steps 1 and 2 of this process that Ed's Tools was developed. In our first pass through a new concept area, we generate what we believe are broad, open-ended essay type questions that we expect will generate responses of between 100 to 200 words. Our goal is to elicit responses that go beyond the usual rhetoric of science employed by our students. Often these questions fail in this regard, and must be reworded (see below).

By its very nature, steps 1 and 2 require that we collect and analyze large numbers (thousands) of essays. Initially, we planned to use the usual means of data collection and analysis: a simple web program to collect data that could then be "dumped" into a computer-assisted text analysis program such as Nudist, and a standard statistical analysis program such as SPSS for tracking pilot test results. While this initially seemed reasonable, in practice it proved awkward to implement.

More than half the functions in the text analysis programs were not useful for concept inventory development, and the functions we did need proved cumbersome, so much so that our initial round of coding was actually conducted using pencil and paper, followed by discussion and comparison of results, deciding on common coding categories, and then following the process again (Figure 1). This was simply too inefficient a process for long-term use, since it required too much face-to-face time.

We quickly realized that we needed tools that better met our needs and increased our efficiency, preferably tools that were platform independent; accessible to all participating coders (faculty volunteers and graduate students working with us); web-based, so that results could be combined and compared easily; and easy to learn and use for non-social scientists and inexperienced graduate students. We therefore developed Ed's Tools (see below), a program that meets all of these needs (and more) since it allows for essay data collection, iterative rounds of coding, actual grading, and tracking student participation by course instructors when they elect to give students credit for responding to essays.

Funding for Ed's Tools, which was named after the undergraduate who coded the program, was provided by the NSF as part of the BCI project. The result is a web-based tool currently being used by STEM educators in the biological sciences, astronomy, and space physics, who have similar needs and are interested in student conceptual understanding. It is hosted by the Center for Integrated Plasma Studies (affiliated with the Department of Physics at the UC Boulder) and available at: https://solarsystem.colorado.edu/conceptInventories/ or through the bioliteracy web site (http://bioliteracy.net).

STUDENT RESPONSE ANALYSIS

We analyzed essay responses using content analysis and beginning without *a priori* assumptions. This was facilitated by the fact that non-instructors did most of the coding. Content analysis is a technique designed to identify recurring patterns in key words, the relationships articulated or implied by the way words and phrases are consistently

used, and oppositions between words and phrases [16,17]. The approach involves the use of multiple coders reviewing essay data and looking for themes and patterns. This gives us a picture of where students are in terms of their thinking and assumptions about specific concepts. Content analysis is reasonably simple to perform using Ed's Tools.

After multiple individual analyses by each coder, patterns and themes were discussed and sorted into clusters or groupings with similar meaning, leading to categories that accurately represent the conceptual landscape as it is experienced and understood by students. Recurring patterns both among individual questions and across different questions made it possible to identify those areas where students experience common difficulties in their understanding. For example, content analysis of essay data indicated that students are able to answer certain types of questions about natural selection correctly (questions asking them to define, label, and categorize), but were unable to demonstrate conceptual-level understanding of natural selection when asked to explain, recognize, or apply their understanding [11].

The research questions addressed during the initial phases of essay analysis are: 1) What recurring themes, patterns, and language do students employ when discussing these broad topics; 2) Are there patterns in their responses that span different questions; 3) What do the recurring themes, patterns, and language tell us about student understanding in these areas; and 4) Do any of these recurring themes, patterns, and/or language indicate common misconceptions or do student responses demonstrate that the majority have solid conceptual understanding in these areas? If results of initial essay analysis are meaningful and complete, we move on to the next step, student interviews, if not, new essay questions need to be developed and asked (step 1). Three types of interviews are involved, again in an iterative process. The initial type of interview is broad and thematic, rather than structured, and focuses on the validation of the interpretation of essay-based content analytic results. These interviews may be held in a small group in order to provide peer discussion from which to draw on, or they may be individual interviews, and deal with a restricted set of themes.

Armed with a reasonable certainty of student conceptual-level understanding and strong indications of the areas where they commonly hold misconceptions as well as how they articulate them, we move on to the second type of interview. These are moderately scheduled think-alouds where students participate in question development and the validation of research results from essay coding and thematic interviews by talking through their thinking processes as they read potential questions and responses. Using as much student language as possible, we create a set of candidate questions along with student expressions of common misconceptions that they then talk through for us. Our research questions at this phase are: 1) Do the students understand the question in the way we intended; 2) When they read each distractor, do they interpret it in the way intended; 3) When they select a response, does it represent their conceptual understanding; and 4) Do we need to modify questions and/or potential responses so that they more accurately reflect student conceptual understanding (particularly their misconceptions)? Again, this is an iterative process, and we may find at this point that we must develop new essay questions or conduct more thematic interviews.

Finally, we refine questions and potential responses based on all interview data, when possible, we compare to essay data (often, there is a large gap between essay responses and the product developed after the second type of interview), and we produce a cluster of questions aimed at a particular conceptual target. This is followed by the final type of student interview, based on scheduled think-alouds. Students first fill out the instrument and then participate in the validation process by talking through what they understood each question asked and how they selected their

responses. Our research questions at this phase are: 1) Do students interpret the questions and potential responses in the ways intended; and 2) Do the students' selected answers accurately represent their conceptual understanding, both at the level of the individual question and at the level of the conceptual cluster?

The final step is the administration of the refined version of the instrument (Figure 1). We look for quantitative results, patterns among answers (e.g., unnecessary questions), test-re-test reliability, etc. and then begin pilot testing. This type of validation process relies heavily on a qualitative component where language and meaning are featured, and is meant to ensure that students believe questions and responses hold the same meaning that developers intended. Reliability is a statistical measure of the repeatability of results in like populations. We find that type 1 interviews typically require between 5 to 10 students; type 2 interviews benefit from between 10 to 15 students; and type 3 interviews require no more than 30 students. This is especially true when the students are offered a reasonable amount of money for their time (about 1 hour), so that a representative sample of students is examined.

Throughout cluster and question development, we ensure that the research justifications for each question and potential responses are clearly articulated and met. Figure 2 shows an example (Question 25 from the BCI) of the resulting articulation of the conceptual understanding that it is intended to explore, and each potential response is directly traced to commonly held misconceptions supported by essay and interview research. The model is also designed so that the purpose of the question and distractors will be clear for all users (who can thus make informed decisions about including a question in their customized instrument).

Figure 2. Question Development Model

Q25. Imagine that you are an ADP molecule inside a bacterial cell. Which best describes how you would manage to "find" an ATP synthase so that you could become an ATP molecule?
This question is designed to test whether students understand that diffusion is caused by random motion of molecules. Our research shows that most students do not understand that molecules are in constant motion. Although they understand the concept of gradients and moving down a concentration gradient, they don't understand what causes this movement.

a. I would follow the hydrogen ion flow.
Students who choose this answer think that ADP somehow can identify where a hydrogen ion gradient is. In our research, students never explain how an ADP would sense an H+ gradient. Many students use language that suggests they think ADP is able to actively seek out an ATP synthase. For example, ADP is described as "looking for" or "noticing." Although the answer is incorrect, students who select may understand correctly that there is a hydrogen ion gradient. They may also correctly understand that the gradient is causing hydrogen ions to flow through the synthase and across the plasma membrane to the area where hydrogen ions are less concentrated.

b. The ATP synthase would grab me.
Students who select this answer think that an ATP synthase senses the presence of ADP and actively grabs it. Excluding some students who said ATP synthase has a receptor that recognizes ADP, most students do not explain how the ATP synthase does this grabbing. Once again, students who select this answer may believe that molecules have the ability to actively seek out or choose other molecules.

c. My electronegativity would attract me to the ATP synthase.
Students who select this answer think that charges cause the ADP and ATP synthase to be attracted to each other. Based on what students know about molecular interactions, it's a good guess. However, it seems to be just that – a guess – because students don't say what charge the ATP synthase is or that the ATP synthase is oppositely charged.

d. I would be actively pumped to the right area.
Students who select this answer think that the ADP is somehow placed in the correct area so that it is close to the ATP synthase. Students may have no explanation for how this occurs.

e. Random movements would bring me to the ATP synthase.
This is the correct answer. In other words, ADP finds ATP synthase by the random motion of ADP molecules.

THE STRUCTURE OF ED'S TOOLS

Ed's Tools provides a facile system for the collection and analysis of **any** kind of student text online (Figure 3), or any data that can be translated into text (i.e., interview transcripts), and the system allows responses to be transferred into SPSS for analysis. In future versions, we plan to include the ability to upload graphics files since these can also reveal important student assumptions.

Ed's Tools was designed and built on a "just-in-time" basis, meeting project needs as they arose, and so there are still areas that would benefit from additional refinement. In its current form, it: 1) contains a posting place for assigned or optional essay questions; 2) contains a coding site that allows for multiple passes; 3) provides a means for comparing coding results from more than one coder; and 4) includes a site where a selected series of responses can be viewed (e.g., all responses for a particular question can be selected or a sample of responses can be seen and printed if needed). We have found that students, instructors, and researchers can easily master Ed's Tools, and Ed's Tools makes it fast and easy to collect student essay responses and facilitates the iterative process of coding those essays. Interview data is handled in the same way when transcribed. Our ability to rapidly identify expanded BCI concept clusters and create CI questions is largely dependent upon Ed's Tools.

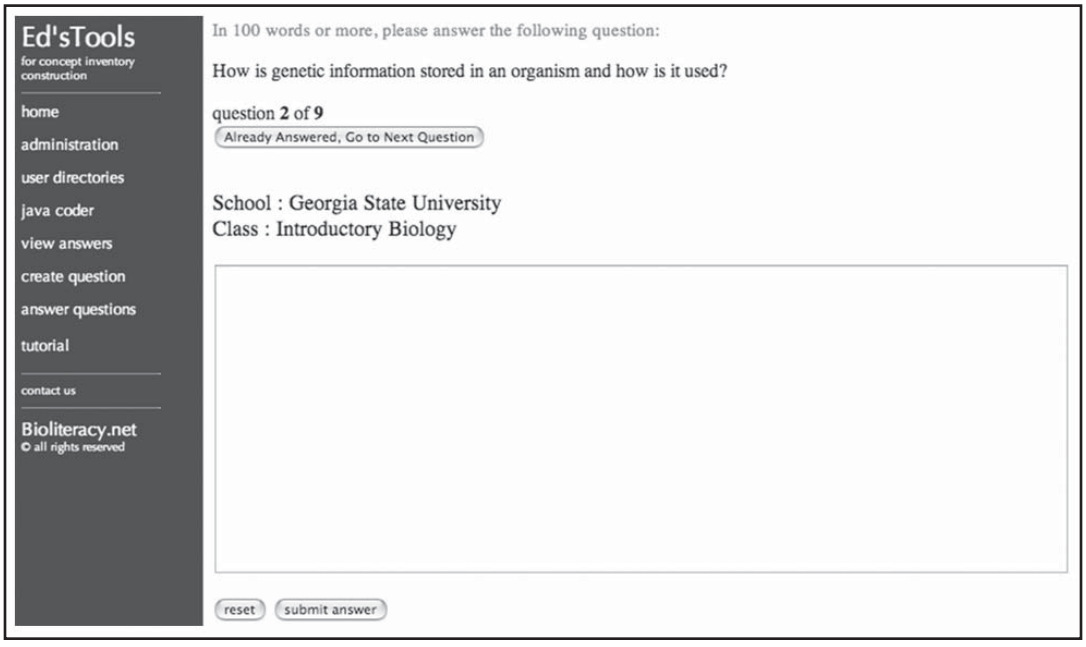


Figure 3. The input field. Instructors can assign any combination of questions to any class. Answers are immediately imported into the answer database.

We have found that Ed's Tools can greatly facilitate collaboration during the iterative process of coding student expressions (verbal or written) of their understanding of the biological sciences. Figure 4 illustrates the Ed's Tools workspace for coding textual data. An individual coder logs on (more than one coder can be logged on at a time without interfering with one another), and then selects a question s/he wishes to work on. The student response then becomes visible within the coding window (font size can be changed for better legibility) and the question is displayed below. The researcher reads through the essay in an effort to identify recurring patterns within the response (usually we need to read through 5 or more essays before we begin to see patterns). While reading through each text, individual coders highlight the portions of text that represent a recurring pattern using the mouse, assign each one a color, and type in a label naming each category of response, theme, or pattern. Each category is identified in the right hand text box, together with its associated color. Multiple categories can be assigned to the same text. During first round coding, these labels tend to be fairly detailed (rather than broad, abstract ideas), so that nothing that may prove to be important is lost.

Once entered, each coder's labels remain with the question set as s/he moves along to new student responses. New labels can be added at any time during the process and text can be labeled as representing multiple categories, themes, and/or patterns. Second round coding generally remains an individual activity in which several of these detailed labels are subsumed under broader, more inclusive (and usually scholarly) names, and labels that do not recur get dropped. None of the first round coding is lost during this process. Third round coding generally involves looking at what the other coders have done with the same data set; considering areas where they overlap and/or disagree; arriving at even more inclusive labels for recurring patterns, categories, and themes; and finally, applying these "third level" codes to an even larger portion (if not all) of the data set. Agreement among coders is much more quickly arrived at using Ed's Tools because each can see what the other coders are doing and have done, consider it asynchronously, and then make group decisions either face-to-face or online. The color-coding scheme, with the ability to call up or print out student language representing each coding category alone, makes the entire process faster.

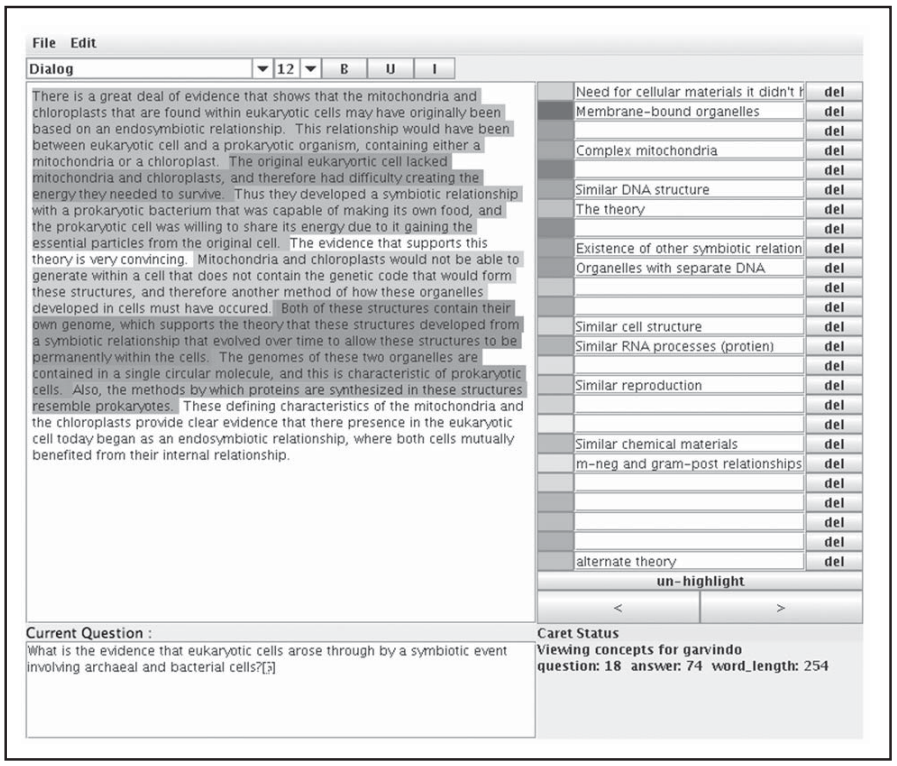


Figure 4. The Concept Coder. The text is imported from a database that associates a set of codes with a coder. Codes can be edited, viewed, or imported. Only the author can change a code, but others can view it and import it (i.e. adopt it for their own).

After the coding of a question is completed, the language (the responder's phrases) that the coders have identified with a particular concept can be aggregated for all (or some) responses using the aggregation tool (Figure 5). The developer specifies the name of the coder, the question, and the concept (top), and the tool aggregates all the text that this particular coder associated with that concept (bottom). A number of permutations of these capabilities are possible, such as showing the text as tagged by all coders, showing all the concepts this coder tagged for this question, etc. Once the language is thus aggregated, it can be used to assign candidate answers (verbatim) to possible Inventory questions. It is important that the answers be used more or less verbatim at this point, as it is the natural language of the students that best represents their misconceptions, and is therefore most likely to resonate with their actual thinking.

Figure 5. The language aggregator. Investigators can select, from the database, the student language that each coder associated with a given concept (top), which is aggregated (bottom), facilitating the construction of the first draft of an Inventory.

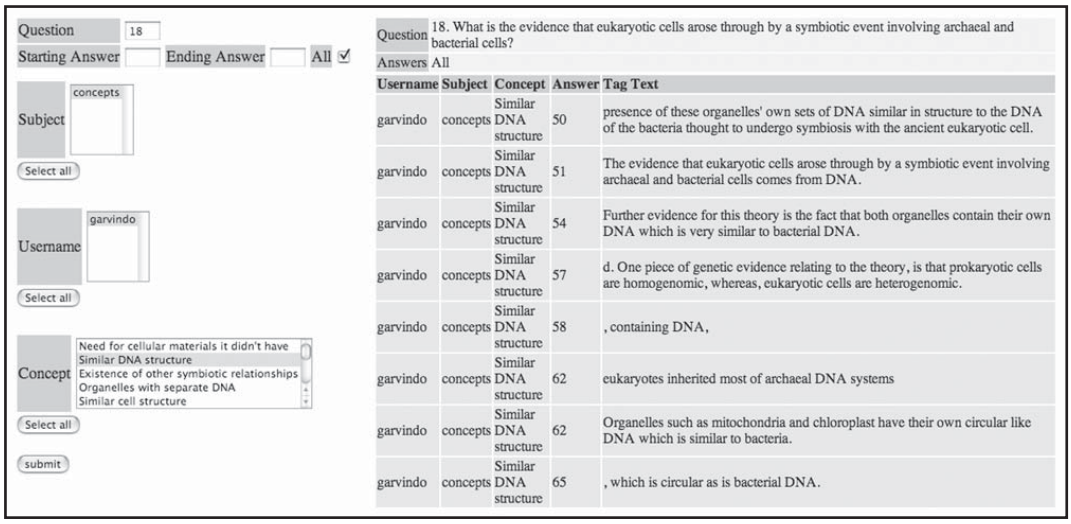
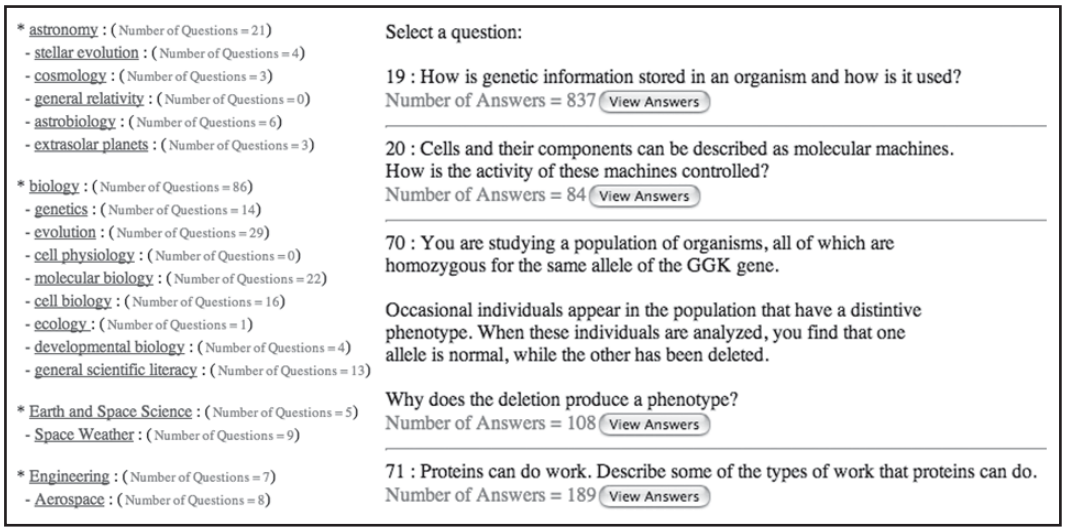


Figure 6. Approximately 10,000 essays, on subjects from astronomy to engineering, have been collected using Ed's Tools for 28 different classes. A sample of 10 answers for each question can be perused at the BCI website.



Although Ed's Tools was developed for the BCI project, it is currently also used for the construction of instruments in astronomy and geoscience (cf. the Space Weather Concept Inventory, SPCI, at solarsystem.colorado.edu/conceptInventories/SPCI), as well as an essay submission and grading tool for courses in engineering. We have collected in all, over 10,000 essays. The topics covered, as well as the number of essays available for each question, can be seen on the public website (Figure 6). A sample of 10 essays for each question can also be perused at the same site.

CONCLUSIONS

Concept Inventories have catalyzed the transformation of teaching of physics, and hold the promise of doing the same for disciplines from biology to engineering, but Concept Inventory construction is a long and laborious process. This is mainly because the validation processes relies heavily on a qualitative component where language and meaning are featured. Ed's Tools greatly facilitates the initial validation stages of the process, and we invite others, both researchers and teachers, to use them.

ACKNOWLEDGEMENTS

This work is part of the NSF-funded Building a Basic Biology Concept Inventory project. While we have had a number of helpful participants, we are particularly grateful to Ed Svirsky for programming, Sara Pallas and Richard Cyr for their support in the early phase of the project, and the cooperation and support of faculty in Molecular, Cellular and Developmental Biology, and the Discipline-based education research group at the University of Colorado, Boulder.

REFERENCES

[1] McDermott, L.C., Shaffer, P.S., and the Physics Education Group, *Tutorials in Introductory Physics*, (Prentice Hall, Upper Saddle River, N.J.), 2002.

[2] Wolfgang, C., and Belloni, M., Physlets: Teaching Physics with Interactive Curricular Material, (Benjamin Cummings), 277 pp., 2000.

[3] Hestenes, D., Wells, M., Swackhammer, G., "Force Concept Inventory," *The Physics Teacher*, vol. 30, pp. 141-158, 1992.

[4] Hake, R., "Interactive engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *American Journal of Physics*, vol. 66, pp. 64-74, 1998.

[5] Hufnagel, B., Slater, T., Deming, G., Adams, J., Adrian, R.L., Brick, C., and Zeilik, M., "Pre-Course Results from the Astronomy Diagnostic Test," *Publications of the Astronomical Society of Australia*, vol. 17, no. 2, p. 152, 2000.

[6] Midkiff, K.C., Litzinger, T.A., and Evans, D.L., "Development of Engineering Thermodynamics Concept Inventory Instruments," *Proceedings of the Frontiers in Education Conference,* Reno, Nev., October 10-13, 2001.

[7] Anderson, D.L., Fisher, K. M., and Norman, G. J., "Development and Evaluation of the Conceptual Inventory of Natural Selection, " *J. Res. Sci. Teaching*, vol. 30, pp. 952-978, 2002.

[8] Martin, J.K., Mitchell, J., and Newell, T., "Development of a Concept Inventory for Fluid Mechanics," *Proceedings of the Frontiers in Education Conference*, Boulder, Colo., 2003.

[9] Libarkin, J., and Anderson, S.W., "The Geoscience Concept Test: Linking grounded theory, scale development, and item response theory," EOS, *Transactions of the American Geophysical Union*, vol. 84, Abstract ED22E-06, 2003.

[10] Klymkowsky, M., Garvin-Doxas, K., and Zeilik., M., "Bioliteracy and Teaching Efficacy: What Biologists can Learn from Physicists," *Cell Biology Education,* vol. 2, pp. 155-161, 2003.

[11] Garvin-Doxas & Klymkowsky, " Lessons from the Biology Concept Inventory: I. Understanding of randomness as a common barrier to biological literacy," in preparation.

[12] Krause, S., Tasooji, A., and Griffin, R., "Origins of Misconceptions in a Materials Concept Inventory From Student Focus Groups," *Proceedings of the ASEE Annual Conference*, 2004.

[13] Jenkins, B.E., Birk, J.P., Bauer, R.C., Krause, S., and Pavelich, M.J., "Development and Application of a Chemistry Concept Inventory," Symposium on Research in Chemistry Education, 227th National Meeting of the American Chemical Society, Anaheim, Calif., Mar. 28-Apr. 1, 2004.

[14] McClymer, J.F., and Knoles, L.Z., "Ersatz Learning, Inauthentic Testing," *Journal of Excellence in College Teaching*, vol. 3, pp. 33-50, 1992.

[15] Messick, S., "Validity," in Linn, R.L. (ed.), *Educational measurement*, 3rd ed., (Macmillan, New York), pp. 13-103, 1989.

[16] De Sola Pool, I., *Trends in Content Analysis*, (University of Illinois Press, Urbana), 1959.

[17] Holsti, O. R., *Content Analysis for the Social Sciences and Humanities*, (Addison-Wesley Publishing Company, Reading, Mass.), 1969.

ClassAction: A Model Rapid-feedback and Dynamic Formative Assessment System

Kevin M. Lee* - *University of Nebraska*
Michael Guidry - *University of Tennessee*
Edward G. Schmidt - *University of Nebraska*
Timothy F. Slater - *University of Arizona*
Todd S. Young - *Wayne State College*

INTRODUCTION

Over the past few years a substantial research base concerning student learning has been developed in university classes of introductory physics, and to a lesser extent, introductory astronomy. This research suggests that science instruction predominantly characterized by professor-centered lectures results in small increases in meaningful understanding of fundamental concepts for most students [1, 2, 3, 4]. Students are not able to sufficiently absorb concepts simply by being informed by authority that the concepts are valid; students must be guided and encouraged to constantly challenge their present state of understanding. In the end, confusion over when and where specific scientific concepts apply can only be resolved through some process of "active engagement" by the learner.

Pedagogical techniques that require students to become active participants in their own learning are collectively known as "active engagement". This modern view of an appropriate pedagogy for introductory science courses is presented in the National Research Council's (NRC) How People Learn [5]. Students' prior knowledge must be challenged, students need repeated exposure to the phenomena in a multitude of contexts, and students' metacognitive skills must be enhanced through critical and supportive feedback. Moreover, an important component of this process is involving students in collaborative group work because most students learn best through social interactions.

Hake [3] completed the most comprehensive study comparing the effectiveness of active engagement techniques and traditional instructional methods. He compared pretest/posttest gains on a physics diagnostic test known as the Force Concept Inventory (FCI) from over 6,000 students in over 60 physics courses. Students in courses making use of active engagement had much larger FCI gains than students in courses taught with more traditional methods.

This paper is concerned most directly with the nearly 250,000 undergraduates who take an introductory astronomy survey course each year in the United States. Fraknoi [6] notes that for many of these students, this course is their only one in astronomy and often marks the end of their formal education in science. Thus, introductory astronomy serves as a unique forum to highlight the intimate relationships between science, technology, and society, while also modeling effective instructional strategies for the numerous preservice teachers who enroll. Lawrenz [7] and colleagues report that as many as 40 percent of students in introductory science survey courses eventually become certified teachers who will serve a critically important role for systemic change in science education in the near future. In combination, these issues elevate introductory astronomy to a critical component in the STEM education system that helps to improve the public's understanding of science, enhance the STEM career pipeline for future scientists and engineers, and support preparation of future teachers.

Nearly half of these students who take astronomy do so in one of more than 500 community and tribal colleges [8]. Fewer than half of introductory astronomy instructors have a formal degree in astronomy and even less have formal training in pedagogy [6]. Contributing to these challenges, the typical university faculty member is necessarily focused on building an active and externally-funded research program and has little time for curriculum development. Both groups need assistance in implementing active engagement in the classroom.

ACTIVE ENGAGEMENT TECHNIQUES

Perhaps the most well-known of the multitude of active engagement techniques is think-pair-share, also known as peer instruction in the vocabulary of Mazur [9]. This is a carefully structured questioning process in which all students in the class participate. In its simplest form this strategy involves halting a lecture to pose a conceptual question (conceptest in the vocabulary of Mazur) to the class that is designed to probe known difficulties with the material. Students then spend several minutes thinking about the question and then respond by voting – either with their fingers, simple flashcards, or some type of electronic personal response system. Based upon the results of this vote, the instructor may choose to immediately move on to new material if the results are very positive, provide additional instruction if the results are poor, or have students discuss the question with their peers and try to reach consensus regarding the correct answer before voting again.

There are numerous advantages to adding think-pair-share to a lecture course. It encourages students to think through the arguments being developed and put them into their own words. It also breaks the monotony of passive lecturing and introduces a social component that most students find enjoyable. Most importantly, it provides both student and instructor with feedback regarding student understanding of the concept. Crouch [11] recently collected data on 30 introductory physics courses using peer instruction techniques at a broad spectrum of colleges and universities. These courses administered the FCI in a pretest/posttest format and showed a normalized gain of 0.39 ± 0.09. A gain over 0.3 is considered quite good and these statistics illustrate the effectiveness of the technique.

The success of peer instruction depends strongly on the quality and relevance of the conceptual questions that are asked. Mazur [9] suggests that the conceptual questions should focus on a single concept, instead of relying on equations, have adequate multiple-choice answers, be unambiguously worded, and have a medium level of difficulty. His book provides a library of conceptual questions encompassing most of the topics of introductory physics. Similar resources are now available for introductory astronomy [12, 13].

Another important technique has been developed by Novak and colleagues and is known as "Just-In-Time Teaching" [13]. This approach blends active learning with web-based technology by having students take pre-class web assignments and using the results to adjust classroom activities. Students receive instruction on the specific questions and problems with which they are having difficulty rather than a generic presentation that would be less likely to meet their needs. This targeted feedback encourages students to be more participatory in the learning process and become interested active learners.

OVERVIEW OF CLASSACTION

ClassAction is a model rapid-feedback and dynamic formative assessment system for promoting interactive engagement in the classroom that builds upon the successes of peer instruction and just-in-time

teaching. It consists of computer databases of conceptual questions for use in collaborative student discussion and interactive voting in introductory astronomy. The databases are programmed in Adobe Flash and are designed to be projected to a class using a computer video projection system. A screenshot of the coordinates and motions module is shown in Figure 1. Separate Flash modules are being created for each major topic area in astronomy, and instructors may conveniently select from a variety of questions in each module. Most question prompts include animations, diagrams, and images that students must interpret when answering. Many questions are carefully designed to focus on tenacious student misconceptions.

Figure 1. The Opening Screen of a ClassAction Module

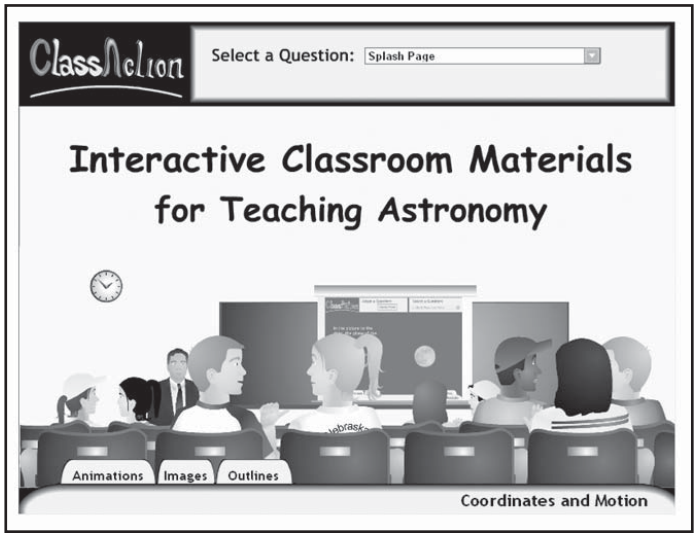
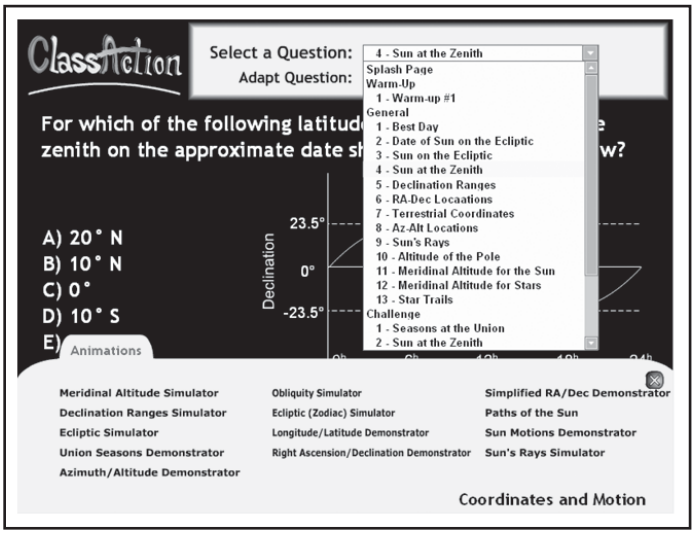


Figure 2. The Question Selection Pull-Down Menu



There are four different types of questions included in ClassAction modules:

1) **Warm-up Questions** do not have voting options and typically take the form of "what's wrong with this picture?" We envision instructors having one of these projected during the few minutes before class starts to serve as an "ice breaker".

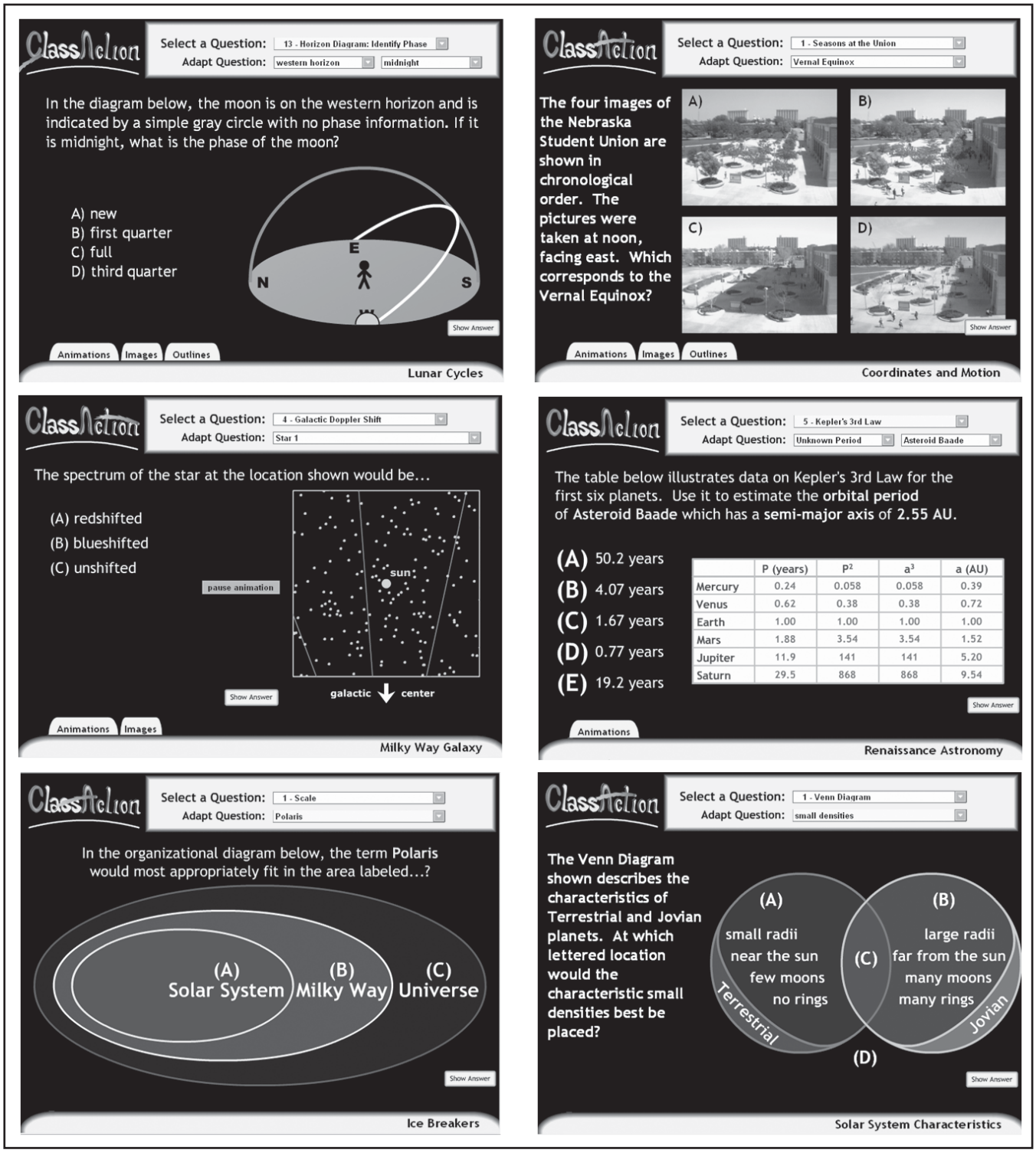
2) **General Questions** are straightforward applications of course principles with voting options and are the most numerous type of ClassAction question. These questions are focused on one (or at most two) astronomical concepts, and while not necessarily easy, the concepts are being tested in the same context in which they are generally covered in a lecture or textbook. We envision instructors using the think-share-pair technique with these questions.

3) **Challenge Questions** require students to transfer knowledge to a new situation. These questions often test several astronomical concepts and students must apply these concepts in a context that would not have been generally covered in a lecture or textbook. We also envision instructors using the think-pair-share method with these questions as well, but budgeting more time for these questions. Hints are often available to allow instructors some control over the difficulty level.

4) **Discussion Questions** do not have simple answers conducive to the think-pair-method. They are open-ended questions that are too difficult for students to work on individually. We envision instructors

encouraging group discussion and then having group representatives report back to the class in a culminating assemblage of wisdom. All four types of questions are available in a pull-down menu shown in the upper right of Figure 2.

Figure 3. Example ClassAction Questions A) Lunar Phases Diagram, B) Images of the Noon Shadow, C) Animation of the Revolution of Stars in the Solar Neighborhood, D) Table of Kepler's Third Law Data, E) Schematic Diagram, F) Venn Diagram of Planet Characteristics.



Many of these questions (and in particular, the general questions) are dynamic in that instructors have the capability to easily recast them into alternate forms based on their own preferences and feedback from the class. For example, one question from the lunar cycles module (shown in Figure 3A) shows the location of a certain phase of the moon in a horizon diagram either on the eastern horizon, meridian, or western horizon. Thus, the instructor can choose from among twenty-four versions of this question. For many questions, the number of permutations is much smaller (i.e. choosing either the greatest or smallest value from a list), but question permutations are made available whenever possible.

Many resources are provided within ClassAction enabling the instructor to provide feedback. *Animations* are programmed in Flash and include everything from simple user controlled diagrams to full interactive simulations. Many of the powerful simulations of the Nebraska Astronomy Applet Project have been incorporated into ClassAction. Instructors will often have the capability to display with a simulation the actual phenomenon being asked about in a question. *Outlines* are one-page summaries of information about a specific astronomical topic most resembling a PowerPoint slide. These will be useful when voting results indicate to an instructor that a short lecture over a topic is necessary. Each different type of resource is available as an expanding panel at the bottom of a ClassAction module with links to the individual resources. The *Animations* panel is shown in expanded view at the bottom of Figure 2.

ClassAction allows instructors to teach interactively in an extremely flexible manner by offering great flexibility in the selection and ordering of questions. They can choose questions based upon student understanding of previous questions. The extensive resources allow instructors to immediately address detected deficiencies in understanding. Instructors can then dynamically create a new permutation of a question and see if their instruction took hold – ClassAction allows instructors to create complete learning cycles. It is anticipated that instructors would plan out a path (with branching possibilities based on student understanding) through the questions of a ClassAction module and identify appropriate resources before class.

REINVENTING CONCEPTUAL QUESTIONS

The majority of published peer instruction questions available for introductory astronomy consist solely of text [12, 13]. The sophistication of questions can be greatly increased by using images, diagrams, tables, and animations to probe student understanding. Another goal of the ClassAction project is to "push the limits" of what constitutes a conceptual question to further the development of student critical-thinking skills.

Although many conceptual questions have their own idiosyncrasies and don't fit nicely into a category, we will discuss some general categories of question prompts and examples from each category.

- *Diagrams of many different types are widely used in ClassAction. An example of lunar phases in a horizon diagram is shown in Figure 3A. Another common technique is placing labeled points on a graph such as the Hertzsprung-Russell Diagram and asking students to identify which point meets a certain criterion. It is very easy to create permutations of questions based on diagrams.*

- *Images are used as the centerpiece for many ClassAction questions. The question in Figure 3B shows a sequence of images of the University of Nebraska Student Union from a western perspective. Students must identify the time of year from the shadow cast by the noontime sun.*

• *Animations are used frequently as feedback tools and occasionally in questions. Figure 3C shows an animation used as a hint to help students determine the Doppler shifts of nearby stars in various directions.*

• *Tables of data are used in many questions such as the example on Kepler's third law shown in Figure 3D. Students are asked to make extrapolations and interpolations, and to identify incorrect entries.*

• *Schematic Diagrams are good at helping students organize and internalize their developing understanding of astronomy. The organizational diagram in Figure 3E is very useful early in an astronomy course to help students classify objects as being in the solar system, Milky Way Galaxy, or the universe. The Venn diagram shown in Figure 3F asks students to identify characteristics associated with terrestrial and jovian planets. Flowcharts and Concept Maps are also used in questions.*

CUSTOMIZATION

Despite the considerable flexibility that ClassAction modules provide, we believe that it is desirable to provide instructors with additional capabilities to further individualize their experience. Recently each ClassAction question and resource has been placed in its own file. Modules are now constructed through extensible markup language .xml files that identify the specific questions and resources to be included. A master module reads the .xml file and assembles the appropriate resources. These .xml files can be edited to delete or add additional content to a module since they are simple text files. Thus, instructors will be able to easily assemble modules containing any ClassAction resources they wish to include. This organization allows materials for concepts like the Doppler Shift that are applicable in many areas of astronomy to be incorporated in multiple modules and yet exist in only one form within the ClassAction project. This is highly desirable from a software engineering perspective.

This organization will allow user-created questions and resources to be included in a ClassAction module. Instructors, with a rudimentary knowledge of Flash, could create questions from Flash template files we would provide or as graphics exported from PowerPoint. Any questions, animations, images, or outlines specified by users in the .xml files would be included in a module. We ultimately plan to create software that will allow instructors to search through all ClassAction materials based on keywords and preview any material before adding it to a module through the drag-and-drop of icons. This software would then save this user-specified module by writing the .xml file.

CLOSING COMMENTS

ClassAction endeavors to extend peer instruction in many meaningful ways through technology. Its organization embodies a new architecture for education: dynamic formative assessment materials surrounded by teaching resources in an extremely easy to use intuitive interface. ClassAction can be effortlessly adopted by astronomy faculty and it is sufficiently flexible to compliment a wide variety of teaching styles. We believe the underlying design will be easily transferable to other disciplines.

ClassAction materials are publicly available at http://astro.unl.edu. This work was funded by NSF grant #0404988.

REFERENCES

[1] Hestenes, D., Wells, M., and Swackhamer, G., "Force Concept Inventory," *The Physics Teacher*, vol. 30, p. 141, 1992.

[2] Redish, E.F., Saul, J.M., and Steinberg, R.N., "On the Effectiveness of Active Engagement Microcomputer-based Laboratories," *American Journal of Physics*, vol. 65, pp. 45-54, 1997.

[3] Hake, R.R., "Interactive-Engagement Versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses," *American Journal of Physics,* vol. 66, no. 1, pp. 64-74, 1998.

[4] Prather, E.E., Slater, T.F., Adams, J.P., Bailey, J.M., Jones, L.V., and Dostal, J.A., "Research on a Lecture-tutorial Approach to Teaching Introductory Astronomy for Non-science Majors," *Astronomy Education Review*, vol. 3, no. 2, 2004.

[5] Bransford, J., (ed.), *How People Learn: Brain, Mind, Experience, and School,* National Research Council Committee on Development in the Sciences, (National Academy Press, Washington, D.C.), 2000.

[6] Fraknoi, A., "Enrollments in Astronomy 101 Courses: An Update," *Astronomy Education Review*, vol. 1, no. 1, pp. 121-123, 2001.

[7] Lawrenz, F., Huffman, D., and Appeldoorn, K., "Enhancing the Instructional Environment: Optimal Learning in Introductory Science," *Journal of College Science Teaching*, vol. 34, no. 7, pp. 40-44, 2005.

[8] Brissenden, G., Prather, E. E., Slater, T. F., Greene, M. W., and Thaller, M., "The NASA Center for Astronomy Education (CAE): Where We Have Been & Where We Are Going," poster abstract, *Bulletin of the American Astronomical Society 208*, no. 17.5, 2006.

[9] Mazur, E., *Peer Instruction: A User's Manual,* (Prentice-Hall, Upper Saddle River, N.J.), 1997.

[10] Crouch, C. H., Watkins, J., Fagen, A. P., and Mazur, E., "Peer Instruction: Engaging Students One-on-One, All At Once," preprint.

[11] Green, P. J., *Peer Instruction for Astronomy,* (Prentice-Hall, Upper Saddle River, N.J.), 2003.

[12] Slater, T. F., and Adams, J. P., 2003, *Learner-Centered Astronomy Teaching: Strategies for Astro 101*, (Prentice-Hall), 2003.

[13] Novak, G. M, Patterson, E. T., Gavrin, A. D., and Christian, W., *Just-In-Time Teaching: Blending Active Learning with Web Technology,* (Prentice-Hall), 1999

Development of the Geoscience Concept Inventory

Julie C. Libarkin* - *Michigan State University*
Steven W. Anderson - *University of Northern Colorado*

OVERVIEW

This article describes the method that was used to develop, pilot, analyze, revise, and validate Geoscience Concept Inventory (GCI) items. GCI research was unique in its attention to grounded theory, scale development theory, and item response theory during item generation and validation. The GCI was evaluated by both scientists and science educators, and was piloted with over 3,500 students enrolled in 60 college geoscience courses nationwide. These data provided invaluable information regarding GCI validity and reliability, and served as the basis for generation of statistically similar GCI sub-tests. The GCI is currently being used for course assessment and research by over 100 faculty and researchers nationwide.

THE IMPETUS FOR CREATION OF THE GCI

The development of the Force Concept Inventory [1] in the early 1990's dramatically changed the way scientists viewed teaching and learning in college-level physics courses, and marked a new beginning for concept inventory development and use in higher education. Research involving the FCI and related instruments has led to significant changes in physics instruction, as well as a new perspective of the importance of physics education research in academic physics [2]. Subsequent development of quantitative instruments in other disciplines, including biology [3], physics [4], astronomy [5,6], and the geosciences (Table 1) [7], has resulted in burgeoning research in conceptual change across science disciplines.

In geoscience, we set out to design an assessment instrument that would be a valid tool for use with all entry-level college students nationwide, and which could be applied to a wide range of courses covering a variety of topics relevant to the Earth sciences. The resulting instrument, the Geoscience Concept Inventory, is truly unique in higher education because of its use of Rasch approaches during development, validation, and use [8]. ***The ability to statistically place all GCI items on a single Rasch scale means that all GCI sub-tests can be tied to a single scale, allowing faculty nationwide to have the freedom of designing tests relative to their course content, without sacrificing the ability to compare student learning and instructional approaches used in different courses!*** Development of the GCI built upon existing studies and incorporated additional methodologies for development and validation, blended three theoretical bases (scale development theory, grounded theory, and item response theory), and utilized a diverse population of students and institutions during piloting (Table 1).

A number of conceptual assessment instruments in other science, mathematics, engineering, and technology (STEM) disciplines have been developed in the past two decades. Faculty and researchers alike have used results from these conceptual inventories to ascertain the effectiveness of new instructional interventions, to evaluate student learning, to compare innovative and traditional pedagogies, and to test the transferability of curricula from one discipline to another. "Learning" has been typically evaluated through comparison of pre vs. post-test averages, use of raw or normalized gain scores, comparison of change in individual scores, implementation of model analysis theory [9], or comparison of scores scaled via Rasch analysis [7,8]. Many concept inventories have been placed online through the Field-tested Learning Assessment Guide (FLAG; www.flaguide.org) or through independent websites (e.g. Geoscience Concept Inventory or Lunar Phases Concept Inventory).

Table 1. Comparison of existing concept inventories and the Geoscience Concept Inventory (GCI)

CONCEPT INVENTORY DEVELOPMENT*	DEVELOPMENT OF THE GCI	COMMENTS ON GCI APPROACH
Predetermined content	Test content is based upon ideas presented by students	Questions are grounded in data gathered from college students
Narrow content focus (e.g., "force" or "lunar phases")	Content covers multiple topics	Use of subtests allows comparison of outcomes across content areas
Alternative choices based on existing studies, questionnaires, and/or interviews (N<30) from limited population	Data from ~20 existing studies 1000+ questionnaires 75+ interviews 10 institutions	Analysis (coding) of qualitative data allows development of authentic "incorrect" choices
50-750 college students tested during piloting	fall 2002: N = 2,219 pre-tests F2003: N = 1,376 pre-tests	For N >~300, statistical sampling of sub-populations is usually possible
Institutions of similar type or locality (N = 1-8)	Colleges: 7 community or tribal, 44 public or private, 60 courses, 8-250 students per course.	The GCI should be generalizeable to all populations of students.
Statistical analyses either not performed, or reliability scores only ‡	Item Response Theory (Rasch) and differential item functioning (DIF) analyses performed.	Raw scores can be re-scaled relative to test difficulty, providing a more accurate measure of changes. DIF suggests bias in items.

*Notes: * Blend of development strategies utilized by Hestenes et al. [1], Zeilik et al. [5], Yeo and Zadnick [4], Anderson et al. [3], and Lindell and Olsen [6]. ‡Anderson et al. [3] perform a factor analysis to ascertain internal validity. Modified from Libarkin and Anderson [8].*

As we have documented previously [8], recognition of student alternative conceptions has led to the development of a number of concept inventories in higher education science. Inventory developers utilize several common approaches, including (see Table 1): 1) Using a pre-determined content focus, usually through expert opinion or a review of texts; 2) Designing alternative responses to multiple-choice items based upon developer experiences in the classroom, a review of existing literature, open-ended questionnaires, and/or interviews with students; and 3) Choosing participating institutions according to type (e.g. large state schools) or similar geographical area. For existing studies discussed in Table 1, the number of students tested during initial piloting ranged from 50 to 750, and measurement of reliability, factor analysis, and model analysis were performed in some studies but not others. A wide range of individual item difficulties is reported for each of these instruments.

THEORETICAL FOUNDATIONS

This significant research into concept test development in some STEM areas provides an excellent basis for test development in other disciplines (Table 1). The development of high-quality tools for assessing student learning is clearly needed to accurately reveal links between learning and teaching. Recent advances in the field suggest that the incorporation of cognitive models into assessment instrument development holds the greatest promise for developing tools that will yield useful results. Valid and reliable assessment tools are most applicable to a wide variety of college students when 1) test questions and answers are developed through careful qualitative evaluation of student conceptions, following pioneering work in other fields; 2) careful attention is paid to scale development, particularly issues related to validity and construction of items; and 3) modern statistical techniques, including item response theory (IRT)[10] approaches, are used to ensure a high degree of validity and reliability [8].

The Cognition-Observation-Interpretation assessment triangle is a validated approach to assessment developed by the National Research Council [11] that considers the link between thinking, measurement, and statistical analysis. In this approach, *Cognition* is a model for student knowledge or skill development, *Observation* is a situation through which student performance can be observed, and *Interpretation* is a method for scoring observations and drawing conclusions. In reference to the GCI one could consider 1) conceptual change in the geosciences as the model being evaluated; 2) the GCI itself as the task that elicits thought; and 3) Rasch and differential item functioning (DIF) analyses as the interpretive agents [8]. A theory of cognition specific to the domain of geosciences is not well established, and should be the basis of future work in geoscience education.

In order to collect the most accurate observations possible using the GCI, we paid careful attention to scale development methodologies and grounded theory. Our use of grounded theory specifically related to the mining of interview and survey data for conceptual questions and responses, rather than driving the research design itself [12,13]. We hoped to embed the development of the GCI in the experiences and perspectives of the population being studied. As a consequence, the GCI test items were only partially based upon predetermined content, and were primarily developed from coding of questionnaire and interview data [14]. Measurement of psychological phenomena is a well-established field of research (psychometrics) and has specific procedures for ensuring validity and reliability. A variety of forms of validity were incorporated into the development of the GCI, including construct validity, content or face validity, criterion validity, external validity, and internal validity (Table 2).

Table 2. Validity and reliability measures used in developing the GCI

VALIDITY/ RELIABILITY*	Exemplar Question	Example of method used for GCI development
Construct Validity	Is there strong support for content of items?	1) Multimethod: GCI stems and items are based upon large interview data set (n=75) and questionnaires (n=1,000); items developed naturally from data (grounded); Think-aloud interviews with students 2) Multitrait: Each concept covered by multiple questions
Content (Face) Validity	Do items actually measure conceptions related to "geoscience"?	1) Review of each question by 3-10 geologists/science educators 2) Review of revised items by 10 to 21 faculty for content and correctness of responses
Criterion Validity	Correlation between GCI and other measures?	1) Trends in quantitative GCI data correlate strongly with conceptions revealed in qualitative data 2) Preliminary GCI 15-item sub-test results show correlation between sub-tests
External Validity	Are results generalizable to other populations?	1) Piloting with wide range of students from 49 institutions 2) Calculation of bias relative to gender and/or ethnicity of subjects via DIF; caution with four items suggested by Mantel-Haenszel DIF approximation.
Internal Validity	Random sample? Do researcher expectations or actions bias results?	1) Items were reviewed by experts in both geology and education 2) GCI administered by participating faculty; no administration bias on part of GCI developers 3) Rasch scales are very similar for pre- and post-tests, suggesting that student attrition and changes made to items during revision do not affect stability of questions on Rasch scale.
Reliability (repeatability)	One example: Are test results repeatable?	1) Administration to multiple populations yielded similar results 2) Classical reliability and Rasch scale stability 3) Internal consistency of items (KR-20) = 0.69 4) The item separation reliability of Rasch scale = 0.99

**Validity/reliability terms differ slightly across disciplines; terms used here reflect general usage. Not all forms of validity and reliability are described here.*

DEVELOPMENT STRATEGY

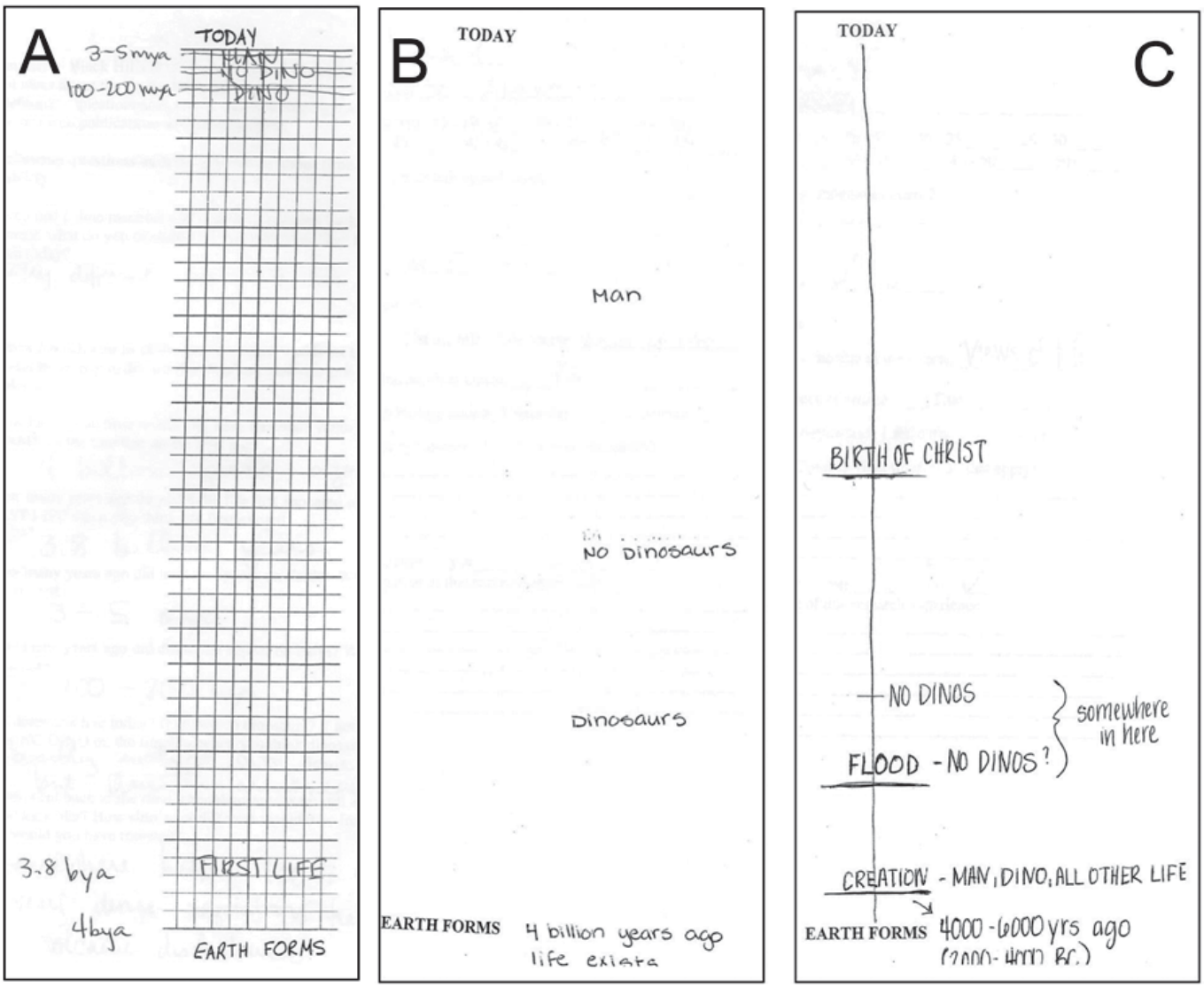
The existing work on concept inventory development provided an excellent foundation from which the GCI could be developed. We were fortunate to receive funding for development of the GCI from the National Science Foundation's Assessment of Student Achievement program. This funding gave us the opportunity to carefully evaluate existing theories and methodologies that might be important to consider in higher education concept inventory development. Seventy-five GCI questions were developed and evaluated over two years and in several iterations. Over the course of this study, we developed a methodology for concept inventory development that acknowledges existing work, incorporates scale development and grounded theory, and utilizes psychometric approaches to modeling behavior. Development of the GCI involved an iterative process of qualitative data collection (both open-ended questionnaires and interviews), question development, review by educators and geoscientists, pilot testing, statistical analysis, revision, testing, and further collection of Think-Aloud Interviews [5] to ascertain the reasons behind student responses. Each step in the process was designed to achieve the highest reliability and validity possible. Test development was an iterative process, such that we returned to a step more than once:

1. **Review of Concept Goals.** A general list of topics that could be covered on a geosciences concept test was generated based on the expertise of the authors, both with Ph.D.s in geology, as well as a review of common entry-level texts. This list was not used to write test items directly, but rather helped frame the development of a semi-structured interview protocol and an open-ended questionnaire.

2. **Qualitative Data Collection.** Open-ended questionnaires were disseminated to students at 17 institutions. Interviews from five sites (including public, private, and community colleges) were used in conjunction with the questionnaire data in developing test items. For example, students were asked to place events on a geologic timeline, yielding a range of responses [14,15].

3. **Generate Test Items.** Test items were generated based on coding of questionnaires and interviews. This direct correlation between qualitative data and item development forced the test to assess fundamental concepts, rather than simply higher-order ideas. Wording of questions was carefully chosen to be unambiguous and non-technical. Finally, question stems were written to be short and clear, although a few readable scenario questions with longer stems were included. All answers and distractors were written in similar language or presented as similar figures. An example of a question that was developed based partially upon data represented in Figure 1 is shown in Figure 2.

Figure 1. Representative student timelines. A) Student timeline where absolute dates and relative positioning are both close to the scientific model. B) Student with mixed scientific and non-scientific ideas. In this case, the student believes that life existed when the Earth first formed. C) Student with strict creationist perspective of Earth's history.



4. **Reduction of Choices to Five or Fewer.** Scale development theory and simple statistical analysis suggests that multiple-choice items should contain at least three answers [16]. Similarly, items containing more than five answers may become artificially difficult as students are faced with too many options. We chose up to five student-derived distractors to reduce the possibility that guessing would influence test results. While some concepts did not lend themselves to more than four conceptual ideas, several, such as age dating, yielded a wide range of alternative ideas.

5. **Pre-testing.** Over 3500 pre-tests were collected from 60 courses at 49 institutions located in 23 states across the U.S. Tested courses were introductory level, and included physical geology, oceanography, environmental science, historical geology, and specialty topic courses. Details about the test population can be found in [7, 8].

6. **Review of Test Items Both Internally and Externally.** Test items were reviewed and revised by the two authors. Test items were then externally reviewed by both education and geoscience faculty, with as many as 20 faculty reviewing a common subset of questions. Each question was reviewed by at least three experts in addition to the authors.

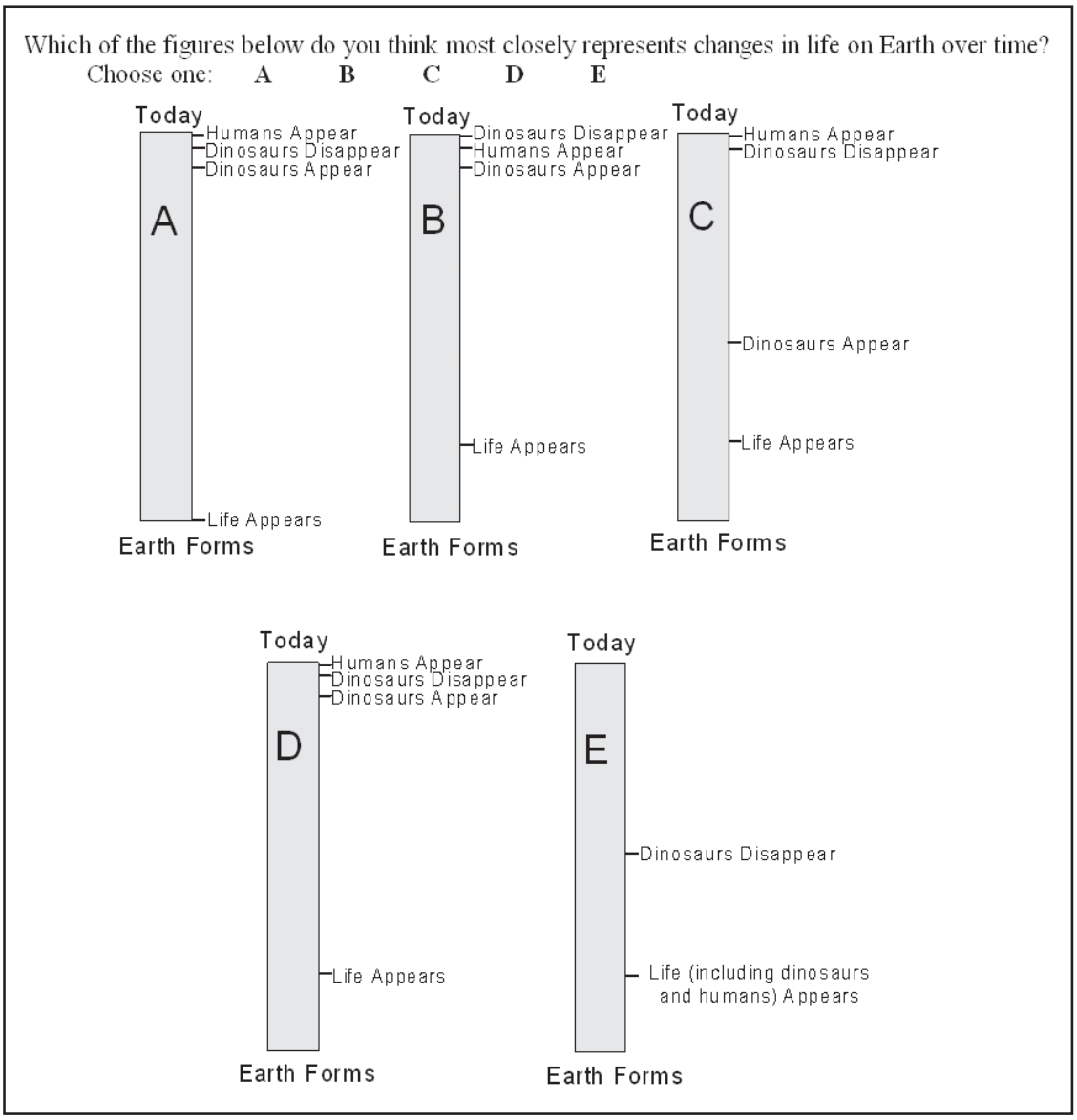
7. **External Review by Participating Faculty.** Geoscience faculty, an expert population, teaching courses participating in the study were invited to provide feedback about the test items and complete the test themselves. Ninety percent of faculty received perfect scores on items they completed, and faculty provided comments about the appropriateness of each item. Items were revised to account for those questions that a few faculty had difficulty with and to accommodate suggested revisions.

8. **Revision of Test Items.** Significant comments from faculty prompted the revision of some test items. These revisions included modification of figures, simple rewording of items, and extensive rewording to accommodate logical inconsistencies or semantic issues.

9. **Post-Testing of Pilot Courses.** About 2,200 students from 35 courses participated in post-semester testing. This second testing was vital to ensure that revision of items did not affect Rasch and DIF relationships (see step 11).

10. **Think-aloud Interviews.** During post-testing, twenty students from two institutions participated in think-aloud interviews. Students were prompted to provide explanations for their responses to test items. This allows researchers to 1) identify any mismatch between student understanding of test items and the intent of those items; 2) identify revisions that need to be made to test items resulting from misunderstandings identified during interviews; and 3) gauge the extent to which students were guessing as opposed to demonstrating alternative conceptions.

Figure 2. Example GCI question. This question was based in part upon student responses as shown in Figure 1.



11. **Item Response Theory (Rasch) Analysis.** Rasch analysis was performed on pre-test and post-test results to ensure the validity of item revisions. Items that maintained their relative positioning on the Rasch scale across testings (pre and post) were considered stable. In addition, we evaluated all items for DIF relative to gender and race. One item has been completely removed from the set of GCI questions based upon gender discrimination, and several more are under further statistical evaluation. Scaling functions for conversion of raw to scaled Rasch scores have also been developed and disseminated to faculty currently using GCI in course assessment.

12. **Creation of GCI sub-tests.** Ongoing activity involves the creation of a methodology whereby faculty and researchers can create 15-item GCI sub-tests that are 1) specific to the content of their courses/research needs; and 2) are statistically similar to each other. All GCI items and the methods for creating sub-tests have been disseminated to over 100 college and high school faculty thus far via the GCI website at http://newton.bhsu.edu/eps/gci.html. Development of an appropriate scaling function and evaluation of the similarity of created sub-tests is ongoing.

Participants

Data were collected from institutions scattered across the U.S. as an approach to ensuring external validity, and the generalizability of the GCI to entry-level college students nationwide. GCI testing was completed at 51 institutions nationwide, with 3,595 students participating in the pre-test (usually on the first day of class and no later than two weeks into the year). We also have post-test data from about 1,750 students, collected during the last week of class or during the final examination. In all, 60 courses in physical geology, environmental science, oceanography, and historical geology, and with class sizes ranging from 8 to 210, were included in the study. These courses stemmed from 51 different institutions in 23 states across the country. Of these institutions, six were community colleges, one was a tribal college, and 44 were public or private four-year institutions. Participants were almost equally split between men and women, and about 20 percent of the students were non-Caucasians. Specific information about courses and students participating in this study are detailed in [7,8].

Rasch Analysis and Differential Item Functioning

We adopted the one-parameter logistic Rasch model for simple item response theory analysis (see [8] for details). The one-parameter logistic model is described by Embertson and Reise [17]:

$$P_i(\theta) = \frac{e^{\left(\theta - b_i\right)}}{1 + e^{\left(\theta - b_i\right)}}, \quad i = 1,2,\ldots \mathrm{n} \qquad (1)$$

where $P_i(\theta)$is the probability that a random examinee with ability, θ, answers item i correctly. $\theta - b_i$ is the difference between ability and item difficulty. Variable b_i is the difficulty parameter, or threshold, for item i. This threshold is the point where the probability of a correct response to the specific item i is 50 percent. Eq. (1) implies an s-shaped curve between 0 and 1 over the ability scale. The one-parameter logistic model assumes unidimensionality in the data, suggesting that items are measuring a single latent trait. This unidimensionality can be evaluated through factor analysis. The presence of a single dominant factor, even given the existence of less important sub-factors, satisfies the condition of unidimensionality [17]. Differential item functioning, or DIF, was used to evaluate differences in item performance relative to the demographic variables of race and gender. In general, individuals at the same level of understanding should perform similarly on unbiased items regardless of other variables (e.g., age, gender, race). DIF compares test-takers with similar abilities and different demographics, looking for dramatic differences in difficulty that are unrelated to the overall test performance of an individual [18,19]. In our case, we specifically used the Mantel-Haenszel statistic to approximate DIF. This approach was chosen because it is a widely-used and common approach employed in the research literature [20], and is also a routine calculation easily performed using QUEST software. For the GCI, we were interested in developing an instrument that would be applicable to the widest range of students possible, and used DIF to identify items that performed differently for different populations of students [8].

FINDINGS

As described in previous publications [7,8], Rasch analysis was used to facilitate the development of statistically similar 15-item GCI sub-tests from the bank of GCI questions. The statistical similarity derives from the calculation of a single scaling function that applies to all sub-tests. Four items were chosen as anchor items for the sub-tests, specifically the two most difficult items, the easiest item, and one intermediate item. The remaining 65 validated items were divided into 11 bins, where each bin is made up of items that are closely grouped on the Rasch scale. The first step in estimating a scaling function for the sub-scales is the averaging of item difficulty

estimates within each bin. Item difficulty estimates within a single bin generally have standard deviations on the order of ±0.15, with highest deviations related to the ends of the Rasch scale. Average item difficulties for each bin were then used to determine a relationship between true score, on a 0-15 scale, and theta. The resulting relationship between raw score and scaled score, as fit by the statistical package JMP, looks like (R^2 =1):

$$S_{GCI} = 16.76 + 4.30R_{GCI} + 0.115(R_{GCI}-7.5)^2 + 0.042(R_{GCI}-7.5)^3 - 0.0017\,(R_{GCI}-7.5)^4 \qquad (2)$$

where S_{GCI} is the scaled score on a 0-100 percent scale and R_{GCI} is the raw score on a 15-item GCI sub-test. For more details on development and use of the GCI, please see http://www.msu.edu/~libarkin.

The GCI is now being used by over 100 faculty and researchers across the U.S., and has been translated into both Spanish and Thai by graduate students (not affiliated with these authors) investigating geoscience conceptions in Puerto Rico and Thailand. The GCI has also been used as a primary assessment tool for a variety of projects [21], has been proposed as a springboard for development of geoscience assessments for elementary students, and has been applied to assessment of high school students. This extensive use of the GCI indicates a major area of need for scientists and educators, and suggests that development of new geoscience instruments would be well received by the community.

CONCLUSIONS

The findings presented above represent a significant leap forward for the field of geoscience education research. The qualitative and quantitative approaches utilized in developing the GCI allow comprehensive analysis of student conceptual understanding of Earth phenomena across a variety of populations. In addition, we developed several unique analytical approaches, including the incorporation of IRT and DIF approaches into concept inventory development in higher education. The methodology outlined above, most importantly the adoption of multiple theoretical foundations and the use of Rasch analysis, allowed for creation of a highly reliable concept inventory. Our experience with the GCI will hopefully aid other researchers in developing similar concept inventories for higher education.

REFERENCES

[1] Hestenes, D., Wells, M., and Swackhamer, G., "Force Concept Inventory," *The Physics Teacher*, vol. 30, pp. 141-158, 1992.

[2] Gonzales-Espada, W.J., "Physics education research in the United States: A summary of its rationale and main finding," *Revista de Educacion en Ciencias,* vol. 4, pp. 5-7, 2003.

[3] Anderson, D.L., Fisher, K.M., and Norman, G.L., "Development and validation of the conceptual inventory of natural selection," *Journal of Research in Science Teaching*, vol. 39, pp. 952-978, 2002.

[4] Yeo, S., and Zadnik, M., "Introductory thermal concept evaluation," *The Physics Teacher*, vol. 39, pp. 496-503, 2001.

[5] Zeilik, M., Schau, C., and Mattern, N., "Conceptual astronomy. II. Replicating conceptual gains, probing attitude changes across three semesters," *American Journal of Physics*, vol. 67, no.10, pp. 923-972, 1999.

[6] Lindell, R.S., and Olsen, J.P., "Developing the Lunar Phases Concept Inventory," *Proceedings of the 2002 Physics Education Research Conference*, 2002.

[7] Libarkin, J.C., and Anderson, S.W., "Assessment of Learning in Entry-level Geoscience Courses: Results from the Geoscience Concept Inventory," *Journal of Geoscience Education*, vol. 53, pp. 394-401, 2005.

[8] Libarkin, J.C., and Anderson, S.W., "The Geoscience Concept Inventory: Application of Rasch Analysis to Concept Inventory Development in Higher Education," in Liu, X., and Boone, W. (eds.), *Applications of Rasch Measurement in Science Education*, (JAM Publishers), pp. 45-73, 2006.

[9] Bao, L., and Redish, E.F., "Concentration analysis: A quantitative assessment of student states," *American Journal of Physics*, vol. 69, no. S1, pp. S45-S53, 2001.

[10] Hambleton, R.K., Swaminathan, H., and Rogers, H.J., *Fundamentals of item response theory*, (Sage, Newbury Park, Calif.) 173pp., 1991.

[11] Pellegrino, J.W., Chudowsky, N., and Glaser, R., (eds.). *Knowing what students know: The science and design of educational assessment*, 2001, (National Academy Press, Washington, D.C.), 366 pp., 2001.

[12] Glaser, B. and Strauss, A., *The discovery of Grounded Theory*, (Aldine Press, Chicago), 1967.

[13] Haig, B.D., "Grounded theory as scientific method," in *Philosophy of Education 1995: Current Issues,* (University of Illinois Press, Urbana, Ill.), pp. 281-290, 1996.

[14] Libarkin, J.C., Anderson, S.W., Dahl, J., Beilfuss, M., Boone, W., and Kurdziel, J.P., "College students' ideas about geologic time, Earth's interior, and Earth's crust", *Journal of Geoscience Education*, vol. 53, no. 1, pp. 17-26, 2005.

[15] Libarkin, J.C., and Kurdziel, J.P., "Ontology and the teaching of earth system science," *Journal of Geoscience Education*, vol. 54, pp. 408-413, 2006.

[16] Haladyna, T.M., Downing, S.M., and Rodriguez, M.C., "A review of multiple-choice item-writing guidelines for classroom assessment," *Applied Measurement in Education,* vol. 15, pp. 309-334, 2002.

[17] Embretson, S.E. and Reise, S.P., *Item Response Theory for Psychologists*, (Lawrence Erlbaum Associates, Mahwah, New Jersey), 2000.

[18] Zumbo, B.D., A Handbook on the Theory and Methods of Differential Item Functioning (DIF): *Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*, (Directorate of Human Resources Research and Evaluation, National Defense Headquarters, Ottawa, Canada), 1999.

[19] Bond, T.G. and Fox, C.M., *Applying the Rasch model: Fundamental measurement in the human sciences,* (Lawrence Erlbaum Associates, Mahwah, N. J.), 2001.

[20] Clauser, B.E. and Mazor, K.M., "Using statistical procedures to identify differentially functioning items," *Educational Measurement: Issues and Practice*, vol. 17, pp. 31-44, 1998.

[21] Lyle-Elkins, N.M., and Elkins, J.T., "Assessment of the effectiveness of interdisciplinary expeditionary field trips on student misconceptions in geosciences," *Geological Society of America Abstracts with Programs*, vol. 36, no. 5, p. 554, 2004.

Measuring What Students Know about How to Learn Chemistry

Nathaniel Grove and Stacey Lowery Bretz* - *Miami University*

INTRODUCTION

Elaine Seymour and Nancy Hewitt's *Talking About Leaving: Why Undergraduates Leave the Sciences* [1] offers a provocative examination of the factors which drive talented, would-be scientists to abandon majors in chemistry, physics, biology, and mathematics for other disciplines outside STEM. *Talking* gives voice to the perspectives of bright young undergraduates who switch their majors from science:

> *"Another common assumption encountered among faculty and challenged by our data is that switching involves the discovery of errors in student choices, judgments or self-perceptions, and represents logical action to correct these … [the] view of students as either 'scientists' or 'not scientists' obscures the loss of many students with the ability to do science - including some who could have done it very well. The nature of the switching process revealed in students' accounts is very different from that imagined by [STEM] faculty. We found the decision to leave [a STEM] major was always the culmination … the process began with poor experiences in [STEM] classes in their first year … Students began to experience self-doubt and lowered confidence in their ability to do science. They became disillusioned with science and the science-based careers to which they had aspired, and questioned whether getting the degree would be worth the effort and distress involved."*

Seymour and Hewitt found that faculty attribution of the switching or loss of science majors to under-preparation, poor judgment, and/or misinformation about what to expect as a STEM major in college did not coincide with the voices of the students themselves as they explained their choice to switch away from science. Rather, students spoke eloquently about the significant effort and distress involved in sacrificing oneself to become a scientist.

In *They're Not Dumb, They're Different*, Tobias [2] also examined the factors that drove bright, inquisitive students from pursuing careers in science and in doing so, coined the construct of the "second tier":

> *The second tier … includes a variety of types of students not pursuing science in college for a variety of reasons. They may have different learning styles, **different expectations**, different degrees of discipline, different 'kinds of minds' from students who traditionally like and do well at science. ([2], p. 14, emphasis added)*

The role and significance of expectations in learning chemistry have not been previously examined. While research abounds on what students already know about chemistry when they come to college, there exist no companion studies to reveal what students know about *how to learn chemistry*. Both Seymour/Hewitt's and Tobias's works stand as important cautions to faculty about the tangible consequences when students whose expectations for learning chemistry are at odds with the reality of the course as structured by the faculty member. Consider, for example, one of Tobias's research subjects' damning expectations about learning chemistry:

Chemistry is a very hard and fast science. Facts are facts and at the introductory level there is little debate about what is presented to students. Basically, there is a body of information that has to pass from professor to student and there is no room for interpretation or creative thought. ([2], p. 52)

It is not surprising that such a student, while bright and capable of doing science, chooses to switch out of STEM. Imagine being a student of a subject that leaves no opportunity for creative thought. What would thinking in that subject look like? What would a student expect learning chemistry to involve? Such are the questions that shaped this research – what do students expect to do to learn chemistry?

THEORETICAL FRAMEWORK

Ausubel and Novak's construct of meaningful learning [3,4,5] provided an operational definition of learning for this research. Meaningful learning is the process of making substantive connections between what a learner already knows and what the learner needs to know. In contrast, rote memorization stands as the polar opposite of meaningful learning. There are three prerequisite conditions for meaningful learning: relevant prior knowledge, meaningful material, and the meaningful learning set, i.e., the affective commitment on the part of the learner to seek and form connections among concepts.

The role of prior knowledge in chemistry has typically focused upon both desirable (e.g., prerequisite mathematics) [6] and undesirable (e.g., misconceptions) prior knowledge [7,8]. However, prior knowledge certainly includes more than *what chemistry a student knows*, but also what a student knows about *how to learn chemistry*.

Tobias's finding that students bring differing expectations about how to learn to their study of chemistry highlights the importance to chemists of a tool that could assess such expectations. If teachers of chemistry had the tools to measure student expectations about learning chemistry, they could better structure the learning experience to bring student expectations into alignment with the reality of the discipline.

PHYSICS EDUCATION RESEARCH

The physics education research community developed such an assessment. The Maryland Physics Expectations Survey (MPEX), developed by Redish and co-workers [9] measured the construct known as 'cognitive expectations':

It is not only physics concepts that a student brings into the physics classroom. Each student, based on his or her own experiences, brings to the physics class a set of attitudes, beliefs, and assumptions about what sorts of things they will learn, what skills will be required, and what they will be expected to do … We will use the phrase ***expectations*** *to cover this rich set of understandings. We focus on what we might call students'* ***cognitive expectations*** *– expectations about their understanding of the process of learning physics and the structure of physics knowledge rather than about the content of physics itself. ([9], p. 213, emphasis, original)*

MPEX contained 34 items across 6 dimensions of learning (e.g., the independence of the learner relative to the teacher) to which students responded on a five-point Likert scale of agree/disagree. The expectations of students (N=1500) enrolled in first-semester calculus-based physics at six institutions were compared to those of physics faculty, showing that student expectations were significantly lower than those of faculty at the start of the

semester. While this was not surprising in and of itself, what was unexpected was that the data at the end of the semester did not support Redish's hypothesis that students' cognitive expectations would improve after 15 weeks spent learning physics. Rather, the students' cognitive expectations declined. Did the students actually know less about learning physics having spent a semester doing so? Or, as inexperienced physics students, had they been naive at the start of the semester and responded with overly optimistic expectations? Perhaps they had given responses more grounded in the reality of their experiences 15 weeks later?

Having learned of Redish's findings in physics, we wondered whether the same phenomenon would exist for students of chemistry. Would there be a gap between faculty and students? Would it grow over time? What are the cognitive expectations for learning chemistry? While physics and chemistry share common methodologies and views of the universe, they are not isomorphic disciplines. We could not simply substitute the word 'chemistry' for each instance of 'physics' in the MPEX instrument. Rather, we needed a framework which conveys the complexities of chemistry and the structure of this discipline and in which to ground the development of our own instrument.

COGNITIVE EXPECTATIONS IN CHEMISTRY

To develop the items in our measure of cognitive expectations for learning chemistry, we turned to Johnstone's domains [10] for inspiration. To learn chemistry, one must not only master information in the three domains of chemistry (macroscopic, particulate, and symbolic), but must also form connections amongst the concepts and ideas in each of these domains [11]. Certainly the most critical venue for making these connections is in the chemistry laboratory where sensory information is used to infer the behavior of microscopic atoms and molecules that is then communicated in an intricate language of symbols and numbers.

Accordingly, we developed an instrument called CHEMX to measure cognitive expectations for learning chemistry [12]. CHEMX consists of 47 items across 7 clusters, or dimensions, of learning chemistry: concepts, mathematics, visualization, reality, effort, outcomes, and laboratory. Responses on a five-point Likert scale range from strongly agree to strongly disagree, and responses are collapsed into favorable or unfavorable positions based on the desired response found in the chemistry education research literature. Table 1 summarizes these clusters and dimensions and provides a sample item from each cluster. The instrument was validated through individual interviews with both students and faculty. The calculated reliability (Cronbach α) was 0.97 overall, with the reliability of the individual clusters ranging from 0.73 to 0.89.

Table 1. CHEMX Clusters and Dimensions.

Cluster	Favorable Dimension	Unfavorable Dimension	Representative Statement
Concepts	Stresses understanding of the underlying ideas and concepts	Focuses on memorizing and using formulas	"When I solve most exam or homework problems, I explicitly think about the concepts that underlie the problems."
Math Link	Considers mathematics as a convenient way of representing physical phenomena	Views chemistry and math as independent with little relationship between them	"In this course, I do not expect to understand equations in an intuitive sense, they just have to be taken as givens."
Visualization	Considers visualization of atoms and molecules in three dimensions essential to learning chemistry	Views visualization as unnecessary for learning chemistry	"Solving a chemistry problem may require me to be able to draw molecules in more than one way."

Reality Link	Believes ideas learned in chemistry are relevant and useful in a wide variety of real contexts	Believes ideas learned in chemistry have little relation to experiences outside the classroom	"It is unnecessary for me to have to relate chemistry to the real world."
Effort	Makes the effort to use information available and tries to make sense of it	Does not attempt to use available information effectively	"I read the text in detail and work through many of the examples given there."
Outcome	Believes learning chemistry is essential to ultimate career goals	Believes chemistry is simply another obstacle to endure before getting to the "important" material	"Only a very few specially qualified people are capable of really understanding chemistry."
Laboratory	Stresses importance of understanding chemical concepts behind experiments	Views laboratory experiments as steps to follow and data to collect with little connection to lecture	"I really don't expect to understand how laboratory instruments work – they are just tools that help me complete the lab."

CHEMX RESULTS

We administered an online version [13] of CHEMX to two sample populations: 1) a stratified random sample of faculty from 50 chemistry departments across the country (N=157) and 2) a purposefully sampled selection of undergraduate students from four diverse institutions (N=597). The four diverse institutions represented a small, selective public university, an open-admission, medium-sized public university, a highly selective, liberal arts college, and a public community college. Keeping in mind the results of the original MPEX study, we hypothesized that a statistically significant gap would initially exist between teacher and student – a gap that would widen as the first semester of general chemistry progressed.

Although our data did show a statistically significant gap between teacher and student at the beginning of general chemistry, no significant increases or decreases in cognitive expectations developed as the semester progressed. Indeed, the hypothesized widening of the gap between faculty and student expectations was not observed until the end of General Chemistry II, when significant declines were revealed across most of the individual CHEMX clusters. After General Chemistry, students' expectations began to improve as they continued into the sophomore and junior years. By the time they finished their junior years, student expectations reached a level comparable to the chemistry faculty sampled.

Gender

In order to gain better insight into these trends, we conducted additional analyses using demographic information such as major, gender, and ethnicity as covariants. The results of these analyses were quite informative and showed not only significant differences between males and females, but also striking differences in how the cognitive expectations of chemistry (including biochemistry) majors changed in relation to non-chemistry majors. At the beginning of General Chemistry I, our CHEMX data indicated that females scored significantly higher on questions contained within the laboratory, effort, and outcome clusters. Although the females in our study consistently maintained their high expectations for these three clusters, it was not until the end of the junior year that males had sufficiently improved so that the significant differences were no longer detectable.

Majors

Further analyses showed that the cognitive expectations of students majoring in chemistry (including biochemistry) steadily improved over the first two years of college chemistry. Simultaneously, non-chemistry majors experienced sharp, prolonged declines. Interestingly enough, our analyses revealed that non-majors actually started their study of chemistry with expectations that more closely agreed with those of chemistry faculty than did the chemistry/biochemistry majors. In addition, their views were significantly higher than those held by chemistry or biochemistry majors. Yet, as the non-majors continued their study of chemistry during the next two years, their views became less and less favorable, digressing further from faculty, to a point that in the case of some clusters, they disagreed with faculty 100 percent of the time. On the other hand, students who intended to major in chemistry or biochemistry began with cognitive expectations far removed from faculty. Despite this sizable gap, many continued to major in chemistry/biochemistry and moved closer to the views held by faculty in the subsequent four semesters. These findings offer quantitative insight into the departure of both Tobias's second tier and Seymour/Hewitt's 'switchers'. CHEMX is a powerful tool to monitor the changing expectations of learners in our chemistry courses.

DISSEMINATION

CHEMX is unique within the chemistry education research community in that it provides chemists with an easy-to-use instrument to "take the cognitive temperature" of their students, and as such, there has been great interest in its use since its introduction. To date, nearly three dozen instructors from 15 institutions have registered to use the online version of CHEMX in their classrooms. Many of these individuals use the results they gather from CHEMX to quantify changes in what students know about how they learn. CHEMX provides a quantitative measure for faculty to gauge how new curricula and pedagogical innovations are affecting how students perceive their role in learning chemistry. Two such examples are highlighted below.

DEVELOPMENT OF SPATIAL ABILITY AND ATTITUDES IN CHEMISTRY COURSES

Principal Investigator - Dr. Renée Cole (Central Missouri State University)

All students enrolled in any chemistry course are administered the CHEMX survey both pre- and post-instruction in the course. During the semester, activities pertaining to computation and visualization are documented for each course. Current seniors will have had very little exposure to these activities, while each subsequent class will have experienced increasing exposure to activities intended to increase visualization skills. CHEMX will allow us to document changes in student thinking about the learning of chemistry, including specifically the mathematics and visualization clusters. In addition, CHEMX will enable us to conduct a longitudinal study to measure the impact of visualization activities across the curriculum [14].

IMPLEMENTING STUDENT-CENTERED GUIDED INQUIRY IN THE CHEMISTRY LABORATORY

Principal Investigator - Dr. Jamie Schneider, (University of Wisconsin, La Crosse)

The introductory chemistry laboratories will be transformed from a teacher-centered, 'cookbook' approach to a more student-centered guided inquiry approach. POGIL [15] will be used to guide this transformation. CHEMX will allow us to assess students' cognitive abilities before, during, and after changes are implemented.

DISCUSSION

CHEMX drew its inspiration from work in the physics education research community. Therefore, it seems appropriate to note that CHEMX has itself crossed disciplinary boundaries to inspire the development of assessment tools in another discipline, namely economics education. [16] Indeed, all disciplines should see the explication of what understandings within the field are paramount to the process of learning as a valuable step forward in promoting learning for their students.

Vast resources in both time and money are spent each year by faculty committed to improving the learning of their students in science, and particularly in chemistry. Creative, novel pedagogies and curricula are field-tested with each new semester. However, the development of valid and reliable measures that characterize multiple dimensions of student knowledge and learning have not kept pace with that of innovations in the classroom. CHEMX is one such tool that can be used in any chemistry classroom or laboratory setting to more closely examine the relationship between the quality of student learning and their thinking about how their learning occurs.

REFERENCES

[1] Seymour, E., and Hewitt, N.M., *Talking about Leaving: Why Undergraduates Leave the Sciences*, (Westview Press, Boulder, Colo.), 1997.

[2] Tobias, S., *They're Not Dumb, They're Different: Stalking the Second Tier,* (Research Corporation, Tucson, Ariz.), 1990.

[3] Ausubel, D., Novak, J., and Hanesian, H., *Educational Psychology*, (Werbel & Peck, New York), 1978.

[4] Novak, J. D., *Learning, Creating, and Using Knowledge*, (Lawrence Erlbaum, Mahwah, N.J.), 1998.

[5] Bretz, S.L., "Human Constructivism and Meaningful Learning," *J. Chem. Educ*., vol. 78, no. 8, p. 1107, 2001.

[6] Ozsogomonoyan, A., and Loftus, D., "Predictors of General Chemistry Grades," *J. Chem. Educ*., vol. 56, pp. 173-175, 1979.

[7] Taber, K., *Chemical Misconceptions - Prevention, Diagnosis and Cure,* Vol. 1, (Royal Society of Chemistry, London), 2002.

[8] Wandersee, J.H., Mintzes, J.J., and Novak, J.D., "Research on Alternative Conceptions in Science," in Gabel, D. (ed.), *Handbook of Research on Science Teaching and Learning,* (Macmillan, New York), pp. 177-210, 1994.

[9] Redish, E.F., Saul, J.M., and Steinberg, R.N., "Student Expectations in Introductory Physics," *Am. J. Phys*., vol. 66, pp. 212-224, 1998.

[10] Johnstone, A.H., "Why is Science Difficult to Learn? Things are Seldom What they Seem," *J. Computer Assisted Learning*, vol. 7, pp. 75-83, 1991.

[11] Gabel, D., "Improving Teaching and Learning through Chemistry Education Research: A Look to the Future," *J. Chem. Educ.*, vol. 76, pp. 548-554, 1999.

[12] Grove, N., and Bretz, S.L., "CHEMX: An Instrument to Assess Students' Cognitive Expectations for Learning Chemistry," *J. Chem. Educ.*, in press.

[13] CHEMX Online Database, http://www.chemx.org

[14] Linenberger, K., and Cole, R.S., "Development of Spatial Ability and Attitudes in Chemistry Courses," Proceedings of the 231st American Chemical Society National Meeting, Atlanta, Ga., March 26-31, 2006.

[15] Farrell, J.J., Moog, R.S., and Spencer, J.N., "A Guided Inquiry Chemistry Course," *J. Chem. Educ.*, vol. 76, pp. 570-574, 1999.

[16] Simkins, S., and Maier, M., "Implications of STEM Discipline Research for Instructional Innovation and Assessment in Economic Education," NSF-CCLI-DUE#0411037.

National Dissemination and Research on Professional Development: Supporting Assessment in Undergraduate Mathematics

Bernard L. Madison* - *University of Arkansas*

OVERVIEW

This paper describes how a dozen mathematics faculty investigators supported efforts of mathematics faculty across the United States in planning and implementing programs to assess student learning in various coherent blocks of undergraduate mathematics courses. This national dissemination project doubled as a research project on professional development programs for college faculty. Because of a decade of work in supporting assessment for program improvements, the grantee institution, the Mathematical Association of America (MAA), was well positioned to make a proposal when the NSF/DUE Assessment of Student Achievement (ASA) program first solicited proposals in 2001. The resources that MAA had developed, including guidelines for assessment and a volume of case studies, constituted the basis for initial national dissemination, and the growing program of college faculty professional development at MAA provided the impetus for research on what works. Additional resources, including another volume of case studies and an information-rich website, were developed and disseminated during the five-year project.

BACKGROUND

In 2001, after a decade of encouraging and supporting comprehensive assessment of learning in undergraduate mathematics, the MAA was well positioned to seize an opportunity for funding from NSF to intensify and extend this work. As a result, NSF awarded MAA $499,928 for Supporting Assessment in Undergraduate Mathematics (SAUM) that provided a much-needed stimulus for assessment at the departmental level. The need for such a program is rooted in the various and often conflicting views of assessment stemming from worry about uses of the results, difficulties and complexities of the work, and possible conflicts with traditional practices. Faculty navigating through these views to develop effective assessment programs encounter numerous tensions between alternative routes and traditional practices that restrict options. Against this background the MAA launched SAUM in January 2002.

In the late1980s, as a result of discussions about the growing assessment movement in higher education, MAA appointed a 12-member Subcommittee on Assessment of MAA's Committee on the Undergraduate Program in Mathematics (CUPM). The subcommittee, which I chaired, was charged with advising MAA members on policies and procedures for assessment of learning in the undergraduate major for the purpose of program improvement. Few of the subcommittee members had experience in or knowledge of assessment, and we struggled with the multiple meanings and connotations of the assessment vocabulary.

In retrospect, the subcommittee's work developed in three phases: 1) understanding the assessment landscape including opposition to assessment; 2) developing guidelines for assessment; and 3) compiling case studies of assessment programs in mathematics departments. A fourth phase was the extensive faculty professional development made possible by SAUM.

Two vehicles proved very helpful in Phase 1. First, in 1991, I moderated an e-mail discussion among fourteen academics (twelve mathematicians and two non-mathematicians) on assessment that included four members of the Assessment Subcommittee. Some of the discussants were opposed to assessment as it was then evolving, with worries ranging from operational issues like extra work to fundamental issues like academic freedom. A report of this discussion is in the MAA volume *Heeding the Call for Change* [1] along with the text of Grant Wiggins' seminal keynote address to the 1990 assessment conference of the American Association for Higher Education (AAHE) [2].

Phase 2 of the Assessment Subcommittee's work consisted of producing a document on assessment that would both encourage assessment and guide department faculties in their efforts to design and implement assessment programs. By 1993, the Subcommittee had a draft ready to circulate for comment. Aside from being viewed as simplistic by some because of inattention to research on learning, the guidelines were well received and CUPM approved them in January 1995 [3].

Further plans of the Subcommittee included gathering case studies as examples to guide others in developing assessment programs. One of the Subcommittee members, William Marion, teamed up with Bonnie Gold and Sandra Keith to gather and edit case studies on more general assessment of learning in undergraduate mathematics. In 1999, *Assessment Practices in Undergraduate Mathematics*, containing seventy-two case studies, was published [4].

Two years later, in 2001, NSF announced the first solicitation of proposals in the new ASA program. The MAA proposal, sponsored within MAA by CUPM, was built around individuals who had been involved in the MAA assessment work or curricular reform through MAA committees such as CUPM and the CUPM Subcommittee on Curriculum Renewal Across the First Two Years (CRAFTY): Thomas Rishel, William Haver, Bonnie Gold, Sandra Keith, William Marion, Lynn Steen, and me. Peter Ewell agreed to serve as SAUM's evaluator, and Michael Pearson came on board a year later when he replaced Rishel as MAA Director of Programs and Services.

SAUM'S GOALS AND STRATEGIES

The goal of SAUM was to encourage and support faculty in designing and implementing effective programs of assessment of student learning in some curricular block of undergraduate mathematics. Consistent with the CUPM guidelines, the purpose of assessment of student learning was program improvement. Types of assessment were not restricted and would vary considerably. SAUM leaders were reasonably sure that many faculty would welcome help with assessment because they were under mandates to develop and implement programs to assess student learning—mandates originating in most cases from external entities such as regional accrediting bodies. Our expectations were accurate: many faculty were willing to tackle assessment, but unenthusiastic and even skeptical about the work.

College mathematics includes several curricular blocks that are potential targets for assessment. These include courses for the mathematics major, courses for future teachers, developmental courses, general education courses, pre-calculus courses, courses for mathematics-intensive majors, reform or innovative courses, and courses for graduate degrees. The last of these, graduate courses, was not a focus of SAUM, but each of the other areas was the target of multiple teams in SAUM workshops, forums, and mini-courses. By far, the most common target area was the major.

During the five years (including two one-year, no-cost extensions) of SAUM, we promoted assessment to hundreds of faculty in professional forums, and worked directly with 68 teams of one to five faculty from 66 colleges or universities in SAUM workshops. Most of the 68 teams had two or three members, with two usually attending the workshop sessions. As these teams worked in the face-to-face workshop sessions, as they continued their work back home, and as we promoted assessment to the larger audiences in professional forums, skepticism about assessment was evident. Nevertheless, we encouraged teams to implement meaningful and effective assessments rather than opting for easier modes. Often this entailed challenging and changing existing practices such as reliance on traditional in-course testing.

FROM AWARENESS TO OWNERSHIP

The SAUM strategy was based on an unarticulated progression of steps to get faculty fully committed to meaningful and effective assessment of student learning. The first step is *awareness*, the second, *acceptance*; next comes *engagement,* and finally ownership.

First, we aimed to make faculty aware of the nature and value of assessment by stimulating thought and discussion using three principal vehicles: 1) panels at national and regional professional meetings; 2) ninety-minute forums at seventeen of the twenty-nine MAA Sections, and 3) distributing the case studies volume, *Assessment Practices in Undergraduate Mathematics*, to the chair of each of the 3,000 plus departments of mathematics in two-year and four-year colleges or universities in the United States.

Second, we encouraged *acceptance* through knowledgeable and respected plenary speakers at workshops, and collegial interaction with others interested in and sometimes experienced in assessment. Examples of the plenary presentations, documented on the SAUM website (www.maa.org/SAUM), are presentations and writings by Lynn Steen (SAUM senior personnel) and Peter Ewell (SAUM evaluator). Their combined overview of how assessment is positioned in the larger arena of federal, state, and university policies and practices can be surmised from their article, *The Four A's: Accountability, Accreditation, Assessment, and Articulation* [5]. This article is based on a presentation by Peter Ewell at a combined face-to-face session of two SAUM workshops at Towson University in January 2003.

Peter Ewell was an unexpected and valuable resource at workshops, giving plenary presentations and generously agreeing to consult with individual teams. His broad historical perspective, vast experience in consulting with and advising colleges and universities, and intimate knowledge of policies of accrediting bodies gave teams both encouragement and helpful advice. Further, Peter's view as a non-mathematician was helpful both for his questioning and his knowledge of other disciplines. Peter's expertise was nicely complemented by Lynn Steen's wide experience with mathematics, mathematics education, and mathematics and science policy issues.

Third, we urged workshop participants to *engage* in designing and implementing an assessment program at their home institutions. Face-to-face workshop sessions required exit tickets that were plans for actions until the next face-to-face session. Teams presented these plans to their workshop colleagues and then reported at the next session on what had been done. As noted by Peter Ewell in his evaluator's report, this strategy provided strong incentive for participants to make progress at their own institution so that they would have something to report at the next session of the workshop.

Finally, we promoted *ownership* by requiring that each team describe their assessment program by writing a case study or presenting a paper or poster at a professional meeting. To facilitate this requirement, SAUM sponsored national meeting paper sessions at MathFest 2003 in Boulder, the 2004 Joint Mathematics Meetings in Phoenix, the 2006 Joint Mathematics Meetings in San Antonio, and the 2007 Joint Mathematics Meetings in New Orleans. SAUM also sponsored a poster session at Phoenix. These sessions comprised 43 paper presentations and 18 posters. This activity started by SAUM will continue with a paper session scheduled for the 2008 Joint Mathematics Meetings in San Diego.

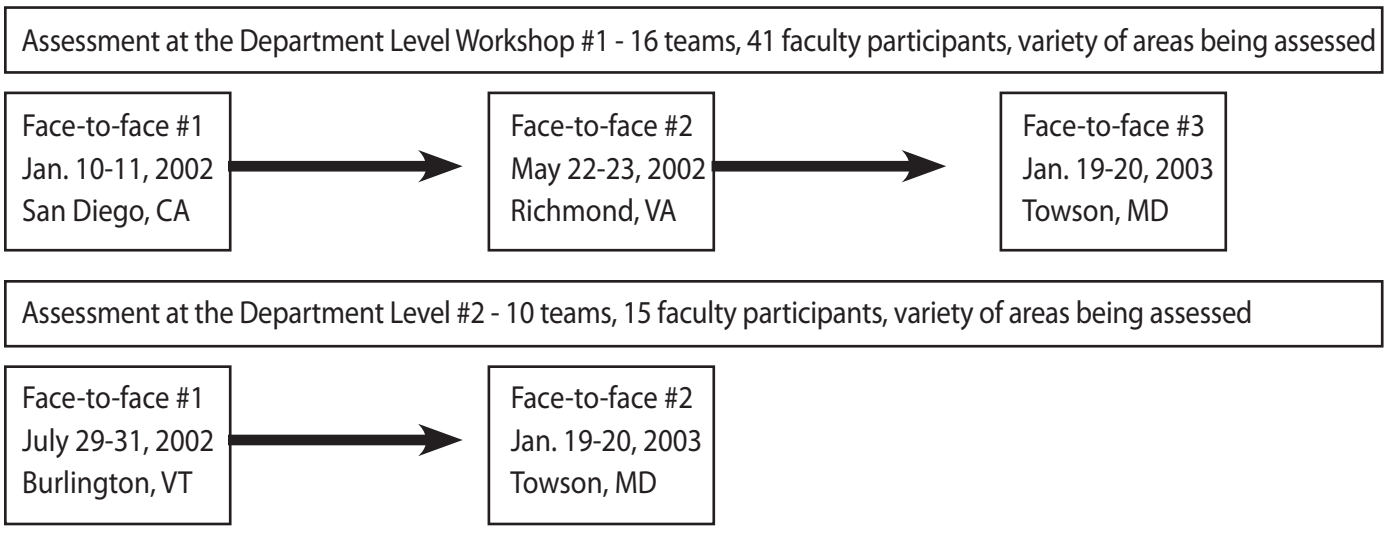
RESOURCE DEVELOPMENT AND NATIONAL DISSEMINATION

Twenty-six of the case studies workshop teams were encouraged to write were compiled in a volume produced by SAUM [6], posted on its website, and distributed as a hard copy to any department that requested one. An additional volume [7] of ten more detailed case studies was published in 2006 by the Association for Institutional Research. Although this volume was not a direct product of SAUM, much of it derived from work by SAUM leaders and SAUM assessment teams. All three case study volumes are available from MAA, and two of them are available on the SAUM website, which is a part of MAA Online (http://www.maa.org/saum), and also includes the following:

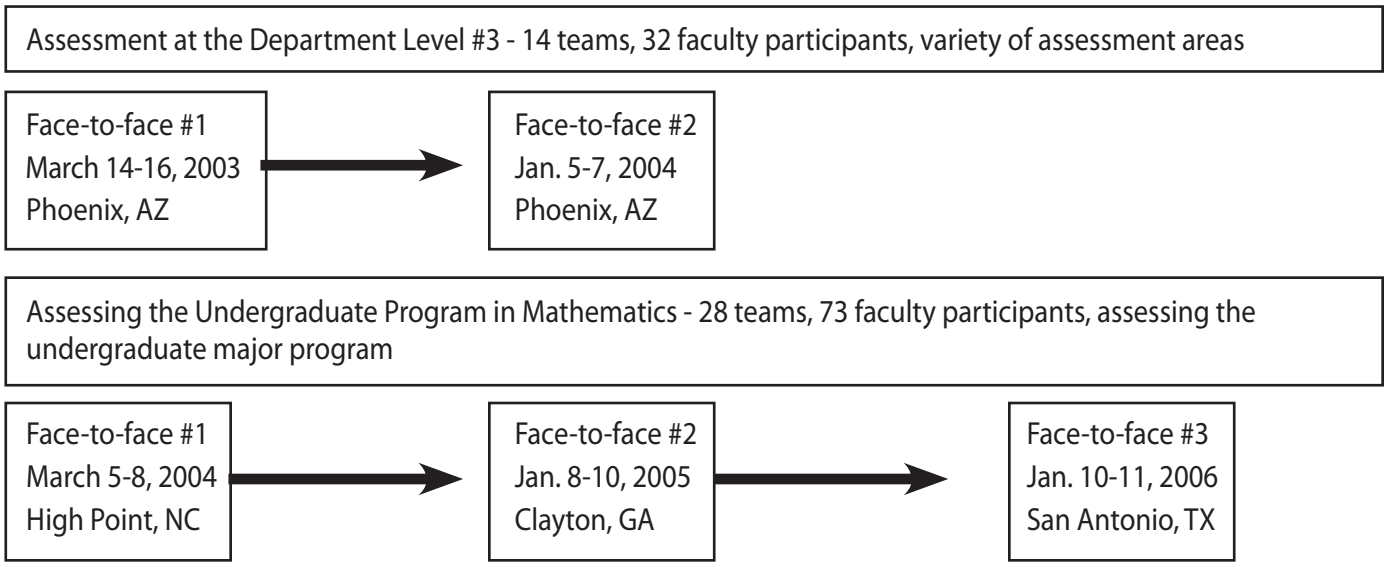
- *An annotated bibliography on assessment with 233 entries*
- *Links to other sites on assessment*
- *A self-paced online workshop on assessment*
- *Some additional case studies*
- *Frequently asked questions on assessment with answers*

PROFESSIONAL DEVELOPMENT WORKSHOP MODEL EVOLUTION AND EFFECTS

The core activity of the movement of faculty from awareness to ownership of assessment programs was multi-year, multi-session workshops. We began SAUM with workshops planned for teams of two or three faculty from institutions, to include two or three face-to-face sessions over one or two years. We scheduled some of the face-to-face sessions just before or after a national meeting of the MAA to minimize extra travel costs and to encourage assessment activities at the national meetings. The models of the four workshops of SAUM are illustrated below.



We combined the two workshops above for the 2003 face-to-face session. This allowed for a richer program and also accentuated the fact that the teams from Workshop #1 had participated in two previous face-to-face sessions while those from Workshop #2 had participated in only one. The participants strongly supported at least three spaced face-to-face sessions, and, as a consequence, we added an Assessment Evening at the Joint Mathematics Meetings in January 2004 for both these workshops and for workshop #3. This was minimally successful, and we had considerable evidence that three spaced face-to-face sessions of at least two days were needed to effectively support assessment activities. This was supported by the findings of the project evaluator, Peter Ewell, as workshop participants repeatedly reported that progress in their assessment activities back in their departments was strongly influenced by their having to report at the next workshop session. Further, the multiple face-to-face sessions were instrumental in establishing networks on assessment and communities of practice and learning about assessment. As we had anticipated, workshop participants learned from each other, especially after we had established common language and principles in the initial face-to-face session. We had already planned Workshop #3 with two face-to-face sessions, but we moved to three for Workshop #4 as indicated below.



The team concept, as opposed to a single faculty participant from an institution, proved to be critical for several reasons. Participants reported that the team basis meant automatic colleagueship and mutual support, which kept the department program moving. Some of the sessions at the workshops were consulting sessions where a team would work with a SAUM leader. Faculty participants who were alone at the workshops (and we did allow that) faced additional challenges in such sessions, and they were sometimes isolated back home in their departments when assessment activities were planned and executed. Too often, a lone faculty worker on assessment becomes the assessment person and is expected to satisfy mandates with little involvement of other faculty.

In addition to the effects of the SAUM workshops on faculty participants cited above, follow-up with workshop participants by Peter Ewell revealed other responses in four basic areas as follows:

NETWORK DEVELOPMENT

- *A network of mathematicians, not just "people doing assessment," was important in grounding departmental efforts.*

- *SAUM workshops reassured participants that assessment remains important at the national level and that other mathematicians are already involved in assessment.*
- *SAUM departments borrowed liberally from one another.*
- *Peer pressure from fellow workshop participants provided impetus for "getting it right."*

PRACTICAL EXPERIENCE

- *Time invested in a messy first effort is better than spending too much time "perfecting" design.*
- *Assessment can be illustrated in manageable ways to show colleagues tangible results. As one participant noted, "the most important lesson I learned was to just get started doing something." Another said, "begin with a manageable project, but begin." A third put it, "start small and stay small."*
- *Assessment was a good deal more time consuming than many first imagined, even for relatively simple things.*
- *Participants learned that local variations in approach are both legitimate and effective.*
- *Participants learned that local rules or customs were not necessarily based on literature about assessment, and, on the other hand, that there were aspects of local culture that could not be addressed by formulaic methods.*

GROWING MATURITY

- *Because of their SAUM experience, participants reported assisting faculty in other disciplines with assessment.*
- *Participants reported success in assessment because it should be done not because it has been mandated, generating a doing it for ourselves attitude.*
- *SAUM experiences provided departments with assessment advocates who could set timelines, enforce deadlines, and provide visibility.*

CHANGING DEPARTMENTAL CULTURE

- *Participation in an NSF-funded project (such as SAUM) helped legitimize efforts back in departments and helped weaken resistance to assessment.*
- *Commitment to assessment by SAUM participants increased participation by other departmental faculty.*
- *SAUM participants became resources for improving departmental courses but also became the assessment workhorses, which often increased cooperation by other faculty.*

CULTURE OF THE DISCIPLINE

The SAUM evaluator, Peter Ewell, administered two surveys to samples of departments, one in the first year of SAUM and a second in 2006, near the end of SAUM. The first survey, to get baseline data, was emailed to a sample of 200 departments stratified by size, institutional type, and location, via MAA departmental liaisons. The end-of-project survey was administered in fall 2006 to a sample of 500 departments similarly stratified. There were 112 responses to the baseline survey and 168 to the end-of-project survey. More details of the survey results are in the final report to NSF on SAUM [8], but highlights of the conclusions include the following:

- *Many departments continue to be in the initial stages of developing a systematic assessment approach, and little overall movement appears to have taken place.*
- *The mathematics major is the most popular target for assessment, with approximately 75 percent of the responding departments indicating activity here.*
- *Ph.D.-granting departments were somewhat less likely than others to be undertaking assessment in any area.*

• *Departments are using a wide variety of assessment methods, with the most popular being faculty-designed examinations.*

• *Approximately one-third of the responders reported awareness of or use of SAUM resources, including the volume of case studies* ***Assessment Practices in Undergraduate Mathematics****, with about half reporting no awareness.*

In Ewell's words, "The results are both positive and sobering. Clearly the SAUM project enjoyed success in the most immediate of its objectives. Important assessment resources were developed for the mathematics community and growing proportions of its member are aware of them and are using them. On the other hand, there does not appear to have been a sea change in the state of assessment in undergraduate mathematics more generally."

REFERENCES

[1] Madison, B. L., "Assessment of undergraduate mathematics," In *Heeding the Call for Change: Suggestions for Curricular Action*, Steen, L.A. (ed.), (Mathematical Association of America, Washington, D.C.), pp. 137-149, 1992.

[2] Wiggins, G., "Toward assessment worthy of the liberal arts: The truth may set you free, but the test may keep you imprisoned," In *Heeding the Call for Change: Suggestions for Curricular Action*. Steen, L.A., (ed.), (Mathematical Association of America, Washington, D.C.) pp. 150-162, 1992.

[3] Committee on the Undergraduate Program in Mathematics, "Assessment of student learning for improving the undergraduate major in mathematics," *Focus: The Newsletter of the Mathematical Association of America*, vol. 15, no. 3, pp. 24-28, 1995. Reprinted in *Assessment Practices in Undergraduate Mathematics,* Gold, B., Keith, S.Z., and Marion, W.A., (eds.), (Mathematical Association of America, Washington, D.C.), pp. 279-284, 1999. Revised version published in *Supporting Assessment in Undergraduate Mathematics*, Steen, L.A., (ed.), (Mathematical Association of America, Washington, D.C.), pp 229-239, 2006.

[4] Gold, B., Keith, S.Z., and Marion, W.A. (eds.), *Assessment Practices in Undergraduate Mathematics*, MAA Notes No. 49, (Mathematical Association of America, Washington, D.C.), 1999.

[5] Ewell, P. T., and Steen, L.A., "The four A's: Accountability, accreditation, assessment, and articulation," *Focus: The Newsletter of the Mathematical Association of America*, vol. 23, no. 5, pp. 6-8, 2003.

[6] Steen, L. A., (ed.), *Supporting Assessment in Undergraduate Mathematics,* (Mathematical Association of America, Washington, D.C.), 2006.

[7] Madison, B. L., (ed.), *Assessment of Student Learning in College Mathematics: Towards Improved Programs and Courses*, (Association for Institutional Research, Tallahassee, Fla.), 2006.

[8] Final report of Supporting Assessment in Undergraduate Mathematics, DUE-0127694, January 2002 to December 2006, Fastlane, (National Science Foundation, Arlington, Va.).

Formative Assessment Materials for Large-Enrollment Physics Lecture Classes

David E. Meltzer* - *University of Washington*

SYNOPSIS: REAL-TIME IN-CLASS FORMATIVE ASSESSMENT

How can an instructor assess students' thinking during class and modify in-class learning activities accordingly? Finding ways to address this question is the objective of our project. Our goal is to develop and test materials that provide a basis for in-class instructional activities, and that also assist the instructor in monitoring student thinking on a moment-to-moment basis.

OVERVIEW

The materials we have developed consist of carefully sequenced sets of multiple-choice items that emphasize qualitative, conceptual questions (See Figure 1 for a sample). They are designed to maximize student-instructor interaction and allow rapid assessment of student learning in a large-class environment. This assessment then aids instructors in structuring and guiding their presentations and class activities

The design of the materials is based on the assumption that the solution of even very simple physics problems invariably hinges on a lengthy chain of concepts and reasoning. Our question sequences guide the student to lay bare these chains of reasoning, and to construct in-depth understanding of physical concepts by step-by-step engagement with conceptual sticking points. Carefully linked sequences of activities first lead the student to confront the conceptual difficulties, and then to resolve them. This strategy is based on one developed at the University of Washington over the past 30 years [1,2,3,4]. Complex physical problems are broken down into conceptual elements, allowing students to grapple with each one in turn and then return to synthesize a unifying perspective [5].

Over several years the materials have undergone a continuous process of testing and revision in actual classroom situations. Constant in-class use reveals ambiguous and confusing wording which is then rapidly corrected in new versions of the materials. Analysis of assessment data provides additional guidance for revision.

MOTIVATION

(NB: Here and below, only selected, representative references to the physics education research literature are given. Relevant references to other and earlier work are provided in the Appendix.)

Research in physics education suggests that instructional methods that incorporate in-class problem-solving activities with rapid feedback can yield improved learning gains, in comparison to traditional lecture methods [5,6]. A key to the success of these methods is that instructional activities should elicit and address common conceptual difficulties, difficulties that are often uncovered or probed through in-depth research on student understanding [1,2,3,4]. When students grapple with conceptual issues by thinking about and solving qualitative problems—problems in which straightforward algebraic procedures may be insufficient (or inefficient) solution methods—learning and retention has often been observed to improve. Instructional methods that engage students in problem-solving activities are sometimes called "active-learning" methods. A particular genre of active-learning methods used in physics has often been referred to by the term "interactive engagement" [6].

INTERACTIVE ENGAGEMENT

Traditionally, instructors (and textbooks) have tended to focus on presenting clearly, precisely, and in painstaking detail the concepts and techniques they wish their students to learn. The emphasis is on the thoroughness and clarity of the presentation [4]. However, in recent decades, research into student learning of physics and other technical subjects has demonstrated that for each new concept or technique to be learned, there will often be a number of conceptual "sticking points" for the student [4,7]. Moreover, there has been increasing recognition of the important role of students' prior (i.e., pre-instruction) knowledge in generating these sticking points and in providing a basis for their eventual resolution [1,2,3,4,8]. In addition, more attention has been paid both to the ways in which students' ideas are linked and organized, and to the nature of students' approaches to applying their knowledge and to solving problems [9]. These realizations have led to a revised view of the instructor's role.

In this revised view, the central function of the instructor is to direct the focus of class activities and discussion toward the key sticking points in the students' thought process, and toward specific weaknesses in the organization of students' knowledge. One has to illuminate in a stark and glaring light, so to speak, the phases in the student's thought process where a key concept or organizational link may be lacking, so that in the student's own mind the gap to be filled is clearly sensed, and the eventual synthesis of the concept or link becomes *dramatically* apparent.

Since ideally one must determine where a student stands conceptually—in the process of understanding a particular idea—in order to guide them to the next phase, some form of back-and-forth interchange with them is essential, even in very large classes. The main focus of instruction is first, to identify the ways in which students are putting the idea together in their own minds, so as to pinpoint any errors or gaps that may exist; second, to identify elements of students' thinking that can potentially form useful and productive components of an improved understanding; and third, to allow the students to grapple with a question, example, or problem that requires them to fill out and perfect their understanding. This could be a problem on which they may all work for several minutes, or instead something as simple as the question: "What is the next step here?" The essential point is to ensure their *active* mental participation as thoroughly as is feasible.

The crux of the instructional problem is that students' minds are not blank slates, and they do not absorb concepts simply by being told (or shown) that they are true. They must be guided continually to challenge their present state of understanding, and to resolve conceptual confusion through a process of active engagement [1]. This may occur either by predicting and then personally investigating the outcome of real physical processes in the instructional laboratory, or by a step-by-step confrontation with conceptual sticking points in the context of a theoretical example [3]. Promoting student interaction through the use of cooperative groups can aid this process by having students challenge each others' understanding, and by encouraging them to help each other deepen their comprehension of the subject matter. As any teacher knows, articulating one's thoughts helps improve one's own learning.

These considerations regarding student learning have led to the development and implementation of a variety of instructional methods which, in the context of physics instruction, have often come to be called by the general term "interactive engagement" [6]. It is particularly challenging to specify what is meant by this term, in part because it generally refers not simply to specific behaviors by the instructors and the students, but also

to specific aspects of the *content* of the instructional materials and activities. These aspects of content refer to features that are explicitly based on consideration of students' pre-instruction knowledge and of their typical learning *behaviors*. Research has suggested that instruction which incorporates certain useful behaviors without also utilizing appropriate *content* may fall far short of the outcomes that result from an appropriate combination of these two key elements [9,10,11].

In view of these considerations, I will outline some of the prominent features of interactive-engagement instruction in physics. Interactive-engagement instruction generally refers to:

1) Instruction that is informed and guided by knowledge of students' pre-instruction knowledge state [1,2,3,4,12,13,14], as well as of their learning trajectory [15,16]. This refers to both their pre-existing ideas and to their learning tendencies. These tendencies constitute the ways in which students typically attempt to apply their pre-existing understanding and reasoning processes to issues that emerge during the course of instruction. These include in particular:

- *a) Specific student learning difficulties related to particular physics concepts [1,2,3,4,6, 8,12,14,17]*
- *b) Specific student ideas and knowledge elements that are productive and useful in helping them grapple with new physics concepts [18]*
- *c) Students' **beliefs** about what they need to do in order to learn [14, 19]*
- *d) Students' actual **behaviors** in the context of the learning process [20]*

2) Instruction that guides students to elicit [14] and address specific difficulties typically encountered when studying new concepts, whether by relatively direct methods (in which students are guided to "confront" these difficulties [1-4]) or less direct methods (in which students are guided to "refine" their ideas to "reconcile" them to physics concepts [18]). Other terms that have been applied to this process include "bridging" [21] (i.e., between more familiar and less familiar concepts) and "weaving" (i.e., of loosely connected initial ideas into more complete understanding) [22].

3) Instruction that emphasizes having students "figure things out for themselves" [13] to the extent that is practical and appropriate. This implies that students are guided to reason out concepts and key ideas through a questioning and discussion process ("guided inquiry"), in contrast to receiving these ideas fully and clearly developed in advance of their problem-solving activity [1,2,3,4,13,23]. In the initial stages, instructors tend to ask students many questions rather than provide either direct answers or detailed formulations of generalized principles. Carefully structured question sequences are often used in this process (Detailed formulations of general principles may however be appropriate at a later stage of the process) [3].

4) Instruction that emphasizes having students engage in a wide variety of problem-solving activities during class time, in contrast to spending most of the time listening to an instructor speak [6, 8].

5) Instruction that leads students to express their reasoning explicitly both in verbal form by interacting with instructors and other students, and in written form through explanations written as part of responses to quiz, homework, and exam problems [1,2,3,4,13,14,22,23,24,25,26]. This helps students more clearly expose—and therefore modify—their own thought processes.

6) Instruction that incorporates students working together in small groups in which they are led both to express their own thinking, and to comment on and critique each others' thinking regarding problems and questions posed for their consideration [3,4,14,17,26].

7) Instruction that ensures that students receive rapid feedback in the course of their problem-solving activity [5,6] (rapid in the sense of a minute-to-minute time scale). This includes feedback from instructors through frequent questions and answers, and feedback from fellow students through small-group interaction.

8) Instruction that emphasizes qualitative reasoning and conceptual thinking [1,2,3,4,5,13,14,23,24,25]. Non-quantitative means of problem solving are used to strengthen students' understanding of fundamental ideas, and to avoid having students focus on mastery of mathematical algorithms as a substitute for that understanding.

9) Instruction that seeks to deepen conceptual understanding by posing problems and eliciting solutions in a wide variety of contexts and representations, incorporating diagrammatic, graphical, pictorial, verbal, and other means of representing ideas and resolving questions [2,4,5,14,17,22,23,24,25,26,27,28,29,30,31].

Note that this list emphasizes the *content* of instructional materials and activities (particularly in items 1, 2, 8, and 9) as much as it does the specific instructional *behaviors* (such as those in items 3 through 7). It has become clear that in order to fulfill the objectives of this form of instruction, substantial prior investigation of students' thinking and learning behaviors is required. This type of research lays the basis for, in particular, the first two items in the process outlined above. Instruction that is based on physics education research of this type is often called "research-based" instruction. Instruction that, by contrast, employs some of the same learning behaviors but in which the content does *not* focus on areas identified with specific learning difficulties is not, apparently, as successful.

Several investigations have addressed the issue of ostensibly "interactive," yet not-very-effective learning environments within the context of physics education. A common theme is that such less-effective environments are missing a key element by not addressing students' *actual* learning difficulties. (Such difficulties may be uncovered through research.) In a study by Redish, Saul, and Steinberg [9], even lectures "with much student interaction and discussion" had little impact on student learning. Hake discusses and analyzes courses supposedly based on interactive engagement that produced subpar learning results [6]. In her Ph.D. research, Pam Kraus looked at this issue more systematically [10]. After a lengthy investigation, she arrived at the following conclusion:

In many of our efforts to improve student understanding of important concepts, we have been able to create an environment in which students are mentally engaged during the lecture. While we have found this to be a necessary condition for an instructional intervention to be successful, it has not proved sufficient. Of equal importance is the nature of the specific questions and situations that students are asked to think about and discuss. ([10], p. 286]

Kraus specifies the key criteria she found effective in improving instruction: *eliciting* students' preconceptions with carefully designed questions, guiding them to *confront* these ideas through appropriate discussion and debate involving all the students, and leading students to *resolve* their difficulties with well-chosen alternative models. A somewhat different alternative approach that has been reported as successful is to guide students to

generate and then test their own explanations for patterns observed in simple experiments [28].
In a careful study reported by Cummings et al. [11], "studio" instruction that involved students working together in small groups using computers was compared with research-based instruction in a similar environment. They found that although the studio-physics classrooms appeared to be interactive and students seemed to be engaged in their own learning, learning outcomes were the same as with traditional instruction. By contrast, introduction of research-based techniques and activities generated significant gains in conceptual understanding, although the studio-classroom environment was otherwise the same as before.

INTERACTIVE ENGAGEMENT IN THE CONTEXT OF LARGE CLASSES

A number of workers in recent years have explicitly addressed the challenge of the large-class learning environment in the context of physics. Van Heuvelen [29,32], developed free-response worksheets for use by students during class meetings in the lecture hall. Eric Mazur [33,34] has achieved great success in popularizing "Peer Instruction," the method he developed for suspending a lecture at regular intervals with challenging conceptual questions posed to the whole class. Students discuss the questions with each other and offer responses using a classroom communication system. Sokoloff and Thornton [35] have adapted microcomputer-based laboratory materials for use in large lecture classes, in the form of *Interactive Lecture Demonstrations*. Novak and collaborators [36] have developed *Just-In-Time Teaching*, which makes use of pre-class web-based computer warm-up exercises, and in-class group work by students using whiteboards. To some extent these incorporate similar methods used and promoted by Hake [13,17], and also by Hestenes and his collaborators [30]. The Physics Education Group at the University of Washington has experimented with modifications of its *Tutorials in Introductory Physics* [37], adapted for use in large lecture classes [10]. Other implementations of active learning in large physics classes using classroom communication systems have been described by the group at the University of Massachusetts, Amherst [38], Poulis et al. [39], Shapiro [40], Burnstein and Lederman [41], Lenaerts et al. [42], and Reay et al. [43], as well as others. The "Scale-Up" project at North Carolina State University [44] also makes use of technology-based systems with similar goals in mind.

It is worth emphasizing that extensive empirical evidence of the instructional effectiveness of these various techniques has been published both in the references cited, and in many other sources cited in turn by those references. To choose just one illustrative example, the effectiveness of the *elicit-confront-resolve* method, as implemented in the *Tutorials* developed at the University of Washington [3,37], has been demonstrated repeatedly by multiple investigators at a variety of institutions, including the use of longitudinal studies, with very consistent results [45]. Learning gains generated through use of these materials were clearly superior to those achieved with more traditional instruction. In view of this vast array of direct empirical evidence, the recent finding of only a "weakly positive" relationship between science achievement and loosely defined "reformed-oriented practices" [46] must be taken to reflect limitations either of that particular study, or of the specific instructional practices probed by that investigation.

CURRENT PROJECT: FORMATIVE ASSESSMENT MATERIALS FOR FULLY INTERACTIVE LECTURES

The specific methods we employ and the materials we have developed are, in effect, a variant of Peer Instruction as developed by Mazur. The basic strategy is to drastically increase the quantity and quality of *interaction* that occurs in class between the instructor and the students, and among the students themselves. To this end, the instructor poses many questions. All of the students must decide on an answer to the question,

discuss their ideas with each other, and provide their responses to the instructor using a classroom communication system. The instructor makes immediate use of these responses by tailoring the succeeding questions and discussion to most effectively match the students' pace of understanding.

In an office or small-group environment, the instructor is relatively easily able to get an ongoing sense of where the students are "conceptually," and how well they are following the ideas that are being presented. By getting continual feedback from them, the instructor is able to tailor his or her presentation to the students' actual pace of understanding. The methods we use allow one, to a large extent, to transform the environment of the lecture hall into that of a small seminar room in which all the students are actively engaged in the discussion.

Our methods begin with a de-emphasis of lecturing. Instead, students are asked to respond to questions targeted at known learning difficulties. We use a classroom communication system to obtain instantaneous feedback from the entire class, and we incorporate group work using both multiple-choice and free-response items. We refer to this method as the "fully interactive lecture" and have described it in detail elsewhere [47]. In the remainder of this section I give a brief synopsis of this method. (Note: Since this particular project was restricted to creation of the multiple-choice items, I will not further discuss the free-response items in this paper.)

We ask questions during class and solicit student responses using printed flashcards (containing letters A, B, C, D, E, and F) or with an electronic "clicker" system. The questions stress qualitative concepts involving comparison of magnitudes (e.g., "Which is larger: A, B, or C?"), directions ("Which way will it move?"), and trends ("Will it decrease, remain the same, or increase?"). These kinds of questions are hard to answer by plugging numbers into an equation.

We give the students some time to consider their response, 15 seconds to several minutes depending on the difficulty. Then we ask them to signal their response by holding up one of the cards, everybody at once. Immediately, we can tell whether most of the students have the answer we were seeking—or if, instead, there is a "split vote," half with one answer, half with another. If there is a split vote, we ask them to talk to each other. Eventually, if necessary, we will step in to—we hope—alleviate the confusion. If they haven't already figured things out by themselves, they will now at least be in an excellent position to make sense out of any argument we offer to them.

The time allotted per question varies, leading to a rhythm similar to that of one-on-one tutoring. The questions emphasize qualitative reasoning, to reduce "equation-matching" behavior and to promote deeper thinking. Questions in a sequence progress from relatively simple to more challenging. They are closely linked to each other to explore just one or two concepts from a multitude of perspectives, using a variety of representations such as diagrams, graphs, pictures, words, and equations. We maintain a small conceptual "step size" between questions for high-precision feedback on student understanding, which allows more precise fine-tuning of the class discussion. In line with this objective, we employ a large proportion of "easy" questions, that is, questions to which more than 80 percent of students respond correctly.

We find that easy questions build confidence, encourage student participation, and are important signals to the instructor of students' current knowledge baseline. Often enough, questions thought by the instructor to be simple turn out not to be, requiring some backtracking. Because of that inherent degree of unpredictability, some

proportion of the questions asked will turn out to be quite easy for the students. If the discussion bogs down due to confusion, it can be jump-started with easier questions. The goal is to maintain a continuous and productive discussion with and among the students.

Many question variants are possible. Almost any physics problem may be turned into an appropriate conceptual question. By using the basic question paradigms "increase, decrease, remain the same," "greater than, less than, equal to," and "left, right, up, down, in, out," along with obvious variations, it is possible to rapidly create many questions that probe students' qualitative thinking about a system. By introducing minor alterations in a physical system (adding a force, increasing a resistance, etc.), students can be guided to apply their conceptual understanding in a variety of contexts. In this way, the instructor is able to provide a vivid model of the flexible and adaptive mental approach needed for active learning.

The development and validation of the question sequences is the central task of this project. Many question sequences are needed to cover the full range of topics in the physics course curriculum. (Other materials needed for interactive lecture instruction include free-response worksheets and text reference materials, but these are under development as part of separate projects.)

RESULTS OF ASSESSMENT

In earlier projects related to this one, we have carried out extensive assessment of student learning. We found that learning gains on qualitative problems were well above national norms for students in traditional courses, at the same time that performance on quantitative problems was comparable to (or slightly better than) that of students in traditional courses [47]. These findings are typical of other research-based instructional methods [4,8].

VALIDATION CHECK AND COLLECTION OF BASELINE DATA

Until recently, the process of drafting our assessment items had been based on extensive instructional experience, knowledge of the results of physics education research [48], and experience in the use of previous, related assessment items. As part of the present project, we have begun to include a more systematic validation process to help confirm that the items both test the knowledge they are intended to test, and catalyze students' reasoning process in the manner intended. This validation process employs patient and time-consuming "think-aloud" interviews with individual students, recorded digitally or on audiotape [49]. This particular project has focused on using these interviews for development of the multiple-choice question sequences, although in other work we have used such interview techniques very extensively as part of the development of free-response materials [50].

In this type of process, students are asked to work through the sequence of questions, explaining their reasoning as they go, while the interviewer examines the details of the student's thinking with gently probing questions. This process can be very effective in 1) uncovering confusing or ambiguous language and word usage; 2) confirming that the students interpret the meaning of the question in the manner intended; and 3) determining whether the students make any tacit assumptions intended by the question (e.g., no external electric field), and do not impose any unintended assumptions (e.g., a need to consider very weak forces). The outcome of this process is to substantially strengthen the quality and utility of the collection of assessment items as a whole. Our data from this phase of the project are as yet only preliminary, but we hope to significantly expand this aspect of the work in the future.

One of the goals of this project was to record student responses to each of the assessment items, including those items already developed and class tested, as well as the items that were developed as a result of the present project. These response data will provide a baseline benchmark for comparison when other instructors make use of the assessment materials, and will assist other instructors in planning and interpreting the use of the materials. Samples of these data (obtained at Iowa State University) are shown in Figure 1. They illustrate that correct response rates on the first few questions in a given sequence are relatively high (80 percent or greater); as the sequence progresses to more challenging items, response rates can drop to the 50 to 70 percent level or less. It is these more challenging questions that usually generate the most productive discussions.

As part of previous projects, initial versions of question sequences for topics in electricity and magnetism, optics, and modern physics had been created. During the present project, we have worked on additional materials for magnetism and modern physics, as well as materials for selected topics in mechanics. Ultimately, we intend to complete question sequences for the full two-semester introductory physics course.

CONCLUSION

Although the methods described here have focused on physics instruction, it is clear that they have broad potential applicability to a wide variety of technical fields. As may be verified in part by consulting the rapidly expanding list of citations [51] to Crouch and Mazur's paper on peer instruction [34], similar methods have been embraced and found useful by, among others, astronomers, geoscientists, physiologists, chemists, engineers, and computer scientists.

ACKNOWLEDGMENTS

Thomas J. Greenbowe assumed responsibilities as Principal Investigator for this project upon the author's move to the University of Washington.

Much of the preliminary work on this project was carried out by Ngoc Loan P. Nguyen, a former graduate student at Iowa State University. Mr. Nguyen died unexpectedly in November 2005 as a result of a sudden illness. This was a devastating loss both personally for this author and for the ongoing work of this project, the completion of which is now significantly delayed.

This work has been supported in part by NSF DUE-0243258, DUE-0311450, PHY-0406724, and PHY-0604703.

APPENDIX: HISTORICAL PERSPECTIVE

Although it is not the purpose of this paper to provide comprehensive references regarding the origins and development of the learning methods cited above, it is useful and interesting to offer some historical perspective. The interactive-engagement teaching methods embraced by researchers in physics education are the products of a long chain of developments. These developments are traceable most directly to educational innovations that followed World War II, although they are partially inspired by still earlier work.

The Physical Science Study Committee project initiated in 1956 by MIT physicists Jerrold Zacharias and Francis Friedman was one of the first steps in this process [52]. Eventually involving a broad array of world-famous

physicists, this project resulted in a dramatic rethinking of the high-school physics curriculum and generated a new textbook [53], along with ancillary curricular materials. The new curriculum was distinguished by a greatly increased emphasis—in contrast to traditional curricula—on communicating a deep conceptual understanding of the broad themes of physical principles. It represented a rejection of traditional efforts that had relied heavily on memorization of terse formulations and "cookbook"-style instructional laboratories.

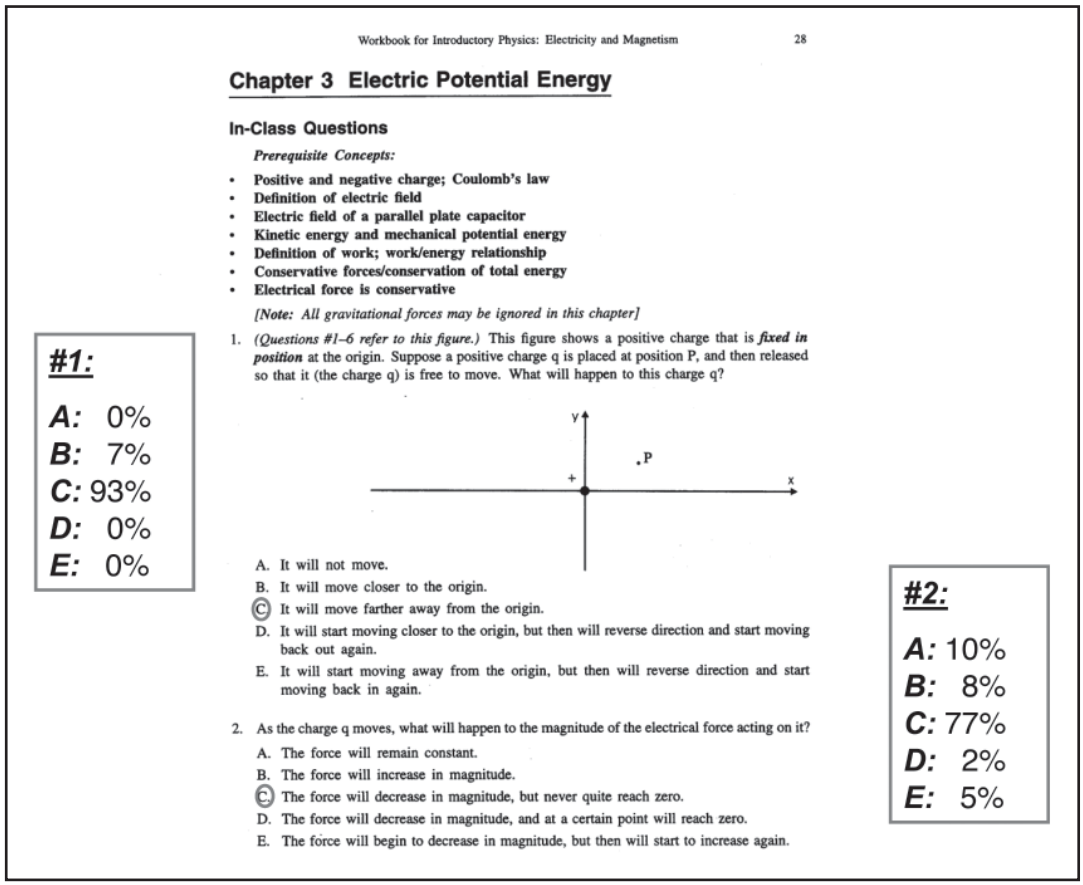
Further catalyzed by the launch of Sputnik in 1957, and with strong funding support by the National Science Foundation (NSF), similar curriculum development efforts were initiated by chemists (in 1957), biologists (in 1959), mathematicians (also in 1959, although preliminary efforts had started in 1952), and earth scientists (in 1962) [54]. A joint conference in 1959 sponsored by the National Academy of Sciences brought the scientists together with prominent psychologists and educators such as Harvard's Jerome Bruner and Piaget collaborator Bärbel Inhelder [55]. General pedagogical principles that emerged from these discussions were enunciated by Bruner [56], Joseph Schwab [57], and others. Soon, the reform effort expanded to include the elementary schools and, backed by the NSF, an explosion of more than a dozen new science curricula aimed at younger students was generated [58]. Prominent physicists again played a central role in several of these curriculum reform projects, notably including Cornell's Philip Morrison (in the "Elementary Science Study" project [59]) and Berkeley's Robert Karplus (a key leader in the "Science Curriculum Improvement Study" [60]). Beginning in the late 1960s and early 1970s, these instructional methods were put into action at the university level by the Washington group led by Arnold Arons [61] and Lillian McDermott [62,63]. In these early efforts, Arons and McDermott put great emphasis on the need for students to formulate and express reasoned responses in written or verbal form to questions that they themselves raised during instruction. Initially, these efforts focused on improving the preparation of prospective K-12 science teachers.

Prominent in all of these efforts was a strong emphasis on learning through guided inquiry (sometimes called "discovery"), utilizing the investigational process of science as a means of teaching scientific concepts themselves [57]. In this process, students would be expected to engage in "discovery of regularities of previously unrecognized relations" [64]. The notion that instructors could guide students through a process of discovery was expressed in the three-phase "learning cycle" propounded by Robert Karplus [65]. In this cycle, students' initial exploration activities led them (with instructor guidance) to grasp generalized principles (concepts) and then to apply these concepts in varied contexts. These ideas of inquiry-based "active" learning could themselves be traced back to workers who came much earlier, including Piaget [66] and his followers, and to proponents of the ancient notions of Socratic dialogue. Piaget's emphasis on the importance of explicitly cultivating reasoning processes that employed hypothesis formation, proportional reasoning, and control of variables, later had an enormous influence on both physics and chemistry educators [67].

Inspired in part by Piaget's earlier groundbreaking investigations, science educators began to perceive the pedagogical importance of the ideas that students brought with them to class. Piaget had emphasized that new ideas being learned had to be "accommodated," in a sense, by a student's already-existing ideas [66]. As Bruner put it, the learning process at first involves "*acquisition* of new information—often information that runs counter to or is a replacement for what the person has previously known implicitly or explicitly. At the very least it is a refinement of previous knowledge" [68]. Later, researchers began systematic efforts to probe students' thinking on a variety of science topics, initially at the elementary and secondary levels [69]. In the late 1970s, Viennot [70] in France, and McDermott and her students in the United States [71], were among the very first to systematically investigate

understanding of science concepts by students enrolled in university-level courses. These investigations led immediately to the development and implementation of research-based instructional methods and curricula. McDermott's research formed the basis for development of curricular materials that explicitly addressed students' pre-instruction ideas. The research-based materials guided students both to elicit their pre-instruction ideas, and then to carry out the thinking processes needed to resolve conceptual and reasoning difficulties that emerged during the instructional process. By doing this research and then bringing to bear on university-level science instruction the pedagogical perspectives and methods employed earlier for younger students, McDermott and other physicist-educators "closed the circle." They had laid the foundation for an ongoing process of research and reform in science education that could engage all participants in the process from the elementary grades on through graduate school. It is on this foundation that the present project is built.

Figure 1. Excerpts from a sequence of "flash-card" questions for interactive lecture, showing student response rates obtained at Iowa State University. The excerpts consist of three (non-consecutive) pages from Chapter 3 of the *Workbook for Introductory Physics* by D. E. Meltzer and K. Manivannan.



Workbook for Introductory Physics: Electricity and Magnetism 28

Chapter 3 Electric Potential Energy

In-Class Questions

Prerequisite Concepts:

- **Positive and negative charge; Coulomb's law**
- **Definition of electric field**
- **Electric field of a parallel plate capacitor**
- **Kinetic energy and mechanical potential energy**
- **Definition of work; work/energy relationship**
- **Conservative forces/conservation of total energy**
- **Electrical force is conservative**

*[**Note:** All gravitational forces may be ignored in this chapter]*

1. *(Questions #1–6 refer to this figure.)* This figure shows a positive charge that is ***fixed in position*** at the origin. Suppose a positive charge q is placed at position P, and then released so that it (the charge q) is free to move. What will happen to this charge q?

A. It will not move.
B. It will move closer to the origin.
C. It will move farther away from the origin.
D. It will start moving closer to the origin, but then will reverse direction and start moving back out again.
E. It will start moving away from the origin, but then will reverse direction and start moving back in again.

2. As the charge q moves, what will happen to the magnitude of the electrical force acting on it?
A. The force will remain constant.
B. The force will increase in magnitude.
C. The force will decrease in magnitude, but never quite reach zero.
D. The force will decrease in magnitude, and at a certain point will reach zero.
E. The force will begin to decrease in magnitude, but then will start to increase again.

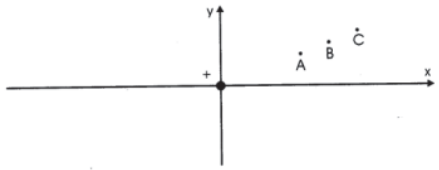
#1:
A: 0%
B: 7%
C: 93%
D: 0%
E: 0%

#2:
A: 10%
B: 8%
C: 77%
D: 2%
E: 5%

#7:

A: 2%
B: 3%
C: 3%
D: 83%
E: 9%

7. *(Questions #7–10 refer to this figure.)* In this figure (as in the previous one) a positive charge is fixed in position at the origin. Suppose a positive charge q is held at rest at position A, and then released and allowed to move freely. It passes through position B, and then moves on toward position C. Which of the following statements about charge q is true?



A. Its kinetic energy is the same at B and A, and its electric potential energy is the same at B and A.
B. Its kinetic energy is larger at B than at A, and its electric potential energy is larger at B than at A.
C. Its kinetic energy is smaller at B than at A, and its electric potential energy is smaller at B than at A.
(D) Its kinetic energy is larger at B than at A, but its electric potential energy is smaller at B than at A.
E. Its kinetic energy is smaller at B than at A, but its electric potential energy is higher at B than at A.

#8:

A: 0%
B: 2%
C: 8%
D: 87%
E: 3%

8. This question again refers to the situation in Question #7. In comparing the energy of the charge q at positions C and B, which of the following statements is true?

A. Its kinetic energy is the same at C and B, and its electric potential energy is the same at C and B.
B. Its kinetic energy is larger at C than at B, and its electric potential energy is larger at C than at B.
C. Its kinetic energy is smaller at C than at B, and its electric potential energy is smaller at C than at B.
(D) Its kinetic energy is larger at C than at B, but its electric potential energy is smaller at C than at B.
E. Its kinetic energy is smaller at C than at B, but its electric potential energy is higher at C than at B.

#9:

A: 0%
B: 13%
C: 7%
D: 53%
E: 22%

9. Again consider the setup shown in Question #7. Suppose now that a positively charged particle is shot from a gun that is located far away from the positive charge at the origin, but which is aimed directly at it. After leaving the gun the particle heads toward the origin, passing first through position C, then position B, and then position A. In comparing its energy at positions C and B, which of the following statements is true?

A. Its kinetic energy and electric potential energy are both the same at C and B.
B. Its kinetic energy and electric potential energy are both larger at C than at B.
C. Its kinetic energy and electric potential energy are both smaller at C than at B.
(D) Its kinetic energy is larger at C than at B, but its electric potential energy is smaller at C than at B.
E. Its kinetic energy is smaller at C than at B, but its electric potential energy is higher at C than at B.

#10:

A: 67%
B: 20%
C: 9%
D: 2%
E: 0%

10. Consider the situation described in #9. Let us call the ***magnitude*** of the change in kinetic energy $|\Delta KE|$, and the ***magnitude*** of the change in electric potential energy $|\Delta PE|$. Which of these is true about the energy of the particle shot from the gun, as it travels from position C to position B?

(A) $|\Delta KE| = |\Delta PE|$
B. $|\Delta KE| > |\Delta PE|$
C. $|\Delta KE| < |\Delta PE|$
D. Not enough information to answer.

REFERENCES

[1] McDermott, L.C., "Millikan Lecture 1990: What we teach and what is learned – closing the gap," *American Journal of Physics*, vol. 59, pp. 301-315, 1991.

[2] McDermott, L.C., "Guest comment: How we teach and how students learn – a mismatch?" *American Journal of Physics*, vol. 61, pp. 295-298, 1993.

[3] McDermott, L.C., "Bridging the gap between teaching and learning: The role of research," in Redish, E.F., and Rigden, J.S., (eds.), *The Changing Role of Physics Departments in Modern Universities: Proceedings of the International Conference on Undergraduate Physics Education, AIP Conference Proceedings 399*, (AIP, Woodbury, New York), pp.139-165, 1997.

[4] McDermott, L.C., "Oersted Medal Lecture 2001: Physics Education Research—The key to student learning," *American Journal of Physics*, vol. 69, pp. 1127-1137, 2001.

[5] Reif, F., "Milliken Lecture 1994: Understanding and teaching important scientific thought processes," *American Journal of Physics*, vol. 63, pp. 17-32, 1995.

[6] Hake, R. R., "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *American Journal of Physics*, vol. 66, pp. 64-74, 1998.

[7] Heron, P.R.L., "Empirical investigations of learning and teaching, Part I: Examining and interpreting student thinking," in Redish, E.F., and Vicentini, M., (eds.), Research on Physics Education: Proceedings of the International School of Physics "Enrico Fermi" Course CLVI, (IOS, Amsterdam), pp. 341-350, 2004; "Part II: Developing research-based instructional materials, Ibid., pp. 351-365; Meltzer, D.E., "The questions we ask and why: methodological orientation in physics education research, in Marx, J., Franklin, S., and Cummings, K., (eds.), *Physics Education Research Conference, Madison, Wisconsin, 6-7 August 2003, AIP Conference Proceedings 720,* (AIP, Melville, New York), pp.11-14, 2004; Heron, P.R.L., "Empirical investigations of student understanding," Ibid., pp. 15-18.

[8] Redish, E.F., and Steinberg, R.N., "Teaching physics: figuring out what works," *Physics Today*, vol. 52, no. 1, p. 24, 1999.

[9] Redish, E F., Saul, J.M, and Steinberg, R.N., "On the effectiveness of active-engagement microcomputer-based laboratories," *American Journal of Physics*, vol. 65, pp. 45-54, 1997.

[10] Kraus, P.A., *Promoting Active Learning in Lecture-based Courses: Demonstrations, Tutorials, and Interactive Tutorial Lectures,* Ph.D. dissertation, University of Washington, University Microfilms, UMI Number 9736313, 1997.

[11] Cummings, K., Marx, J., Thornton, R., and Kuhl, D., "Evaluating innovations in studio physics," *American Journal of Physics,* vol. 67, pp. S38-S44, 1999.

[12] Halloun, I.A., and Hestenes, D., "The initial knowledge state of college physics students," *American Journal of Physics,* vol. 53, pp. 1043-1055, 1985.

[13] Hake, R.R., "Promoting student crossover to the Newtonian world," *American Journal of Physics*, vol. 55, pp. 878-884, 1987.

[14] Goldberg, F., and Bendall, S., "Making the invisible visible: A teaching/learning environment that builds on a new view of the physics learner," *American Journal of Physics*, vol. 63, pp. 978-991, 1995.

[15] Thornton, R.K., "Conceptual Dynamics: Following changing student views of force and motion," in Redish, E.F., and Rigden, J.S., (eds.), *The Changing Role of Physics Departments in Modern Universities: Proceedings of the International Conference on Undergraduate Physics Education, AIP Conference Proceedings 399*, (AIP, Woodbury, New York), pp. 241-266, 1997.

[16] Meltzer, D.E., "How do you hit a moving target? Addressing the dynamics of students' thinking", in Marx, J., Heron, P.R.L., and Franklin, S., (eds.), Physics Education Research Conference, Sacramento, Calif., 4-5 August 2004, *AIP Conference Proceedings 790*, (AIP, Melville, New York), pp. 7-10, 2005.

[17] Hake, R.R., "Socratic pedagogy in the introductory physics laboratory," *The Physics Teacher,* vol. 30, pp. 546-552, 1992.

[18] Elby, A., "Helping physics students learn how to learn," *American Journal of Physics*, vol. 69, pp. S54-S64, 2001.

[19] Redish, E. F., Saul, J.M., and Steinberg, R.N., "Student expectations in introductory physics," *American Journal of Physics*, vol. 66, pp. 212-224, 1998; Hammer, D., "Student resources for learning introductory physics," *American Journal of Physics,* vol. 68, pp. S52-S59, 2001; May, D. B., and Etkina, E., "College physics students' epistemological self-reflection and its relationship to conceptual learning," *American Journal of Physics*, vol. 70, pp. 1249-1258, 2002; Adams, W.K., Perkins, K.K., Podolefsky, N.S., Dubson, M., Finkelstein, N.D., andWieman, C.E., "New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, *Physical Review Special Topics–Physics Education Research*, vol. 2, pp. 010101-1–010101-14, 2006.

[20] Thornton, R.K., "Uncommon knowledge: Student behavior correlated to conceptual learning," in Redish, M., and Vicentini, M., (eds.), *Research on Physics Education: Proceedings of the International School of Physics "Enrico Fermi" Course CLVI*, (IOS, Amsterdam), pp. 591-601, 2004.

[21] Clement, J.J., "Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics," *Journal of Research in Science Teaching,* vol. 30, pp. 1241-1257, 1993.

[22] Minstrell, J.A., "Teaching science for understanding", in Resnick, L.B., and Klopfer, L.E., (eds.), *Toward the Thinking Curriculum: Current Cognitive Research*, (Association for Supervision and Curriculum Development, Alexandria, Va.), pp. 129-149. 1989.

[23] Arons, A.B., "Cultivating the capacity for formal reasoning: Objectives and procedures in an introductory physical science course," *American Journal of Physics*, vol. 44, pp. 834-838, 1976.

[24] Redish, E.F., "Implications of cognitive studies for teaching physics," *American Journal of Physics,* vol. 62, pp. 796-803,1994.

[25] Leonard, W.J., Dufresne, R.J., and Mestre, J.P., "Using qualitative problem-solving strategies to highlight the role of conceptual knowledge in solving problems," *American Journal of Physics*, vol. 64, pp. 1495-1503, 1996.

[26] Heller, P., Keith, R., and Anderson, S., "Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving," *American Journal of Physics,* vol. 60, pp. 627-636, 1992; Heller, P., and Hollabaugh, M., "Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups," *American Journal of Physics*, vol. 60, pp. 637-644, 1992.

[27] McDermott, L.C., "A view from physics," in Gardner, M., Greeno, J.G., Reif, F., Schoenfeld, A.H., diSessa, A., and Stage, E., (eds.), *Toward a Scientific Practice of Science Education,* (Erlbaum, Hillsdale, N.J), pp. 3-30, 1990.

[28] Etkina, E., and Van Heuvelen, A., "Investigative Science Learning Environment: Using the processes of science and cognitive strategies to learn physics, in Franklin, S., Marx, J., and Cummings, K., (eds.), *Proceedings of the 2001 Physics Education Research Conference*, (PERC Publishing, Rochester, N.Y.), pp. 17-21, 2001; Etkina, E., Murthy, S., and Zou, X., "Using introductory labs to engage students in experimental design," *American Journal of Physics*, vol. 74, pp. 979-986, 2006.

[29] Van Heuvelen, A., "Learning to think like a physicist: A review of research-based instructional strategies," *American Journal of Physics*, vol. 59, pp. 891-897, 1991.

[30] Hestenes, D., "Modeling methodology for physics teachers," in Redish, E.F., and Rigden, J.S., (eds.), *The Changing Role of Physics Departments in Modern Universities: Proceedings of the International Conference on Undergraduate Physics Education, AIP Conference Proceedings 399,* (AIP, Woodbury, New York), pp. 241-266, 1997.

[31] Meltzer, D.E., "Relation between students' problem-solving performance and representational format," *American Journal of Physics,* vol. 73, pp. 463-478, and references therein, 2005.

[32] Van Heuvelen, A., "Overview, case study physics," *American Journal of Physics,* vol. 59, pp. 898-907, 1991.

[33] Mazur, E., *Peer Instruction: A User's Manual,* (Prentice Hall, Upper Saddle River, N.J.), 1997.

[34] Crouch, C.H., and Mazur, E., "Peer instruction: Ten years of experience and results," *American Journal of Physics*, vol. 69, pp. 970-977, 2001.

[35] Sokoloff, D. R., and Thornton, R.K., "Using interactive lecture demonstrations to create an active learning environment, *The Physics Teacher,* vol. 35, no. 10, pp. 340-347, 1997.

[36] Novak, G.M., Patterson, E.T., Gavrin, A.D., and Christian, W., *Just-In-Time Teaching: Blending Active Learning with Web Technology,* (Prentice Hall, Upper Saddle River, N.J.), 1999.

[37] McDermott, L.C., Shaffer, P., and the Physics Education Group, *Tutorials in Introductory Physics,* (Prentice Hall, Upper Saddle River, N.J.), 2002.

[38] Dufresne, R.J., Gerace, W.J., Leonard, W.J., Mestre, J.P., and Wenk, L., "Classtalk: A classroom communication system for active learning," J. Comp. Higher Educ., 7, pp. 3-47, 1996; Mestre, J. P., Gerace, W.J., Dufresne, R.J., and Leonard, W.J., "Promoting active learning in large classes using a classroom communication system," in Redish, E.F., and Rigden, J.S., (eds.), *The Changing Role of Physics Departments in Modern Universities: Proceedings of the International Conference on Undergraduate Physics Education, AIP Conference Proceedings 399,* (AIP, Woodbury, New York), pp. 241-266, 1997; Wenk, L., Dufresne, R., Gerace, W., Leonard, W., and Mestre, J., "Technology-assisted active learning in large lectures", in McNeal, A.P., and D'Avanzo, C., (eds.), *Student-Active Science, Models of Innovation in College Science Teaching,* (Saunders College Publishing, Fort Worth, Tex.), pp. 431-452, 1997; Beatty, I.D., Gerace, W. J., Leonard, W.J., and Dufresne, R.J., "Designing effective questions for classroom response system teaching," *American Journal of Physics,* vol. 74, pp. 31-39, 2006.

[39] Poulis, J., Massen, C., Robens, E., and Gilbert, M., "Physics lecturing with audience paced feedback," *American Journal of Physics,* 66, pp. 439-441, 1998.

[40] Shapiro, J.A., "Electronic student response found feasible in large science lecture hall", *J. Coll. Sci. Teach*., vol. 26, pp. 408-412, 1997.

[41] Burnstein, R. A., and Lederman, L.M., "Using wireless keypads in lecture classes," The *Physics Teacher,* vol. 39, pp. 8-11, 2001.

[42] Lenaerts, J., Wieme, W., and Van Zele, E., "Peer Instruction: A case study for an introductory magnetism course," *European Journal of Physics*, vol. 24, pp. 7-14, 2003.

[43] Reay, N.W., Bao, L., Li, P., Warnakulasooriya, R., and Baugh, G., "Toward the effective use of voting machines in physics lectures," *American Journal of Physics,* vol. 73, pp. 554-558, 2005.

[44] Beichner, R. J., Saul, J.M., Allain, R.J., Deardorff, D.L., and Abbott, D.S., "Introduction to SCALE-UP: Student-centered activities for large enrollment university physics," in *Proceedings of the 2000 Annual meeting of the American Society for Engineering Education,* Session 2380, 2000. *http://www2.ncsu.edu/ncsu/pams/physics/Physics_Ed/index.html*

[45] Numerous papers (more than a dozen) by the Washington group itself have documented learning gains resulting from this method, as cited for instance in references 1 through 4 above. Confirmatory evidence in replication experiments has been published in a longitudinal study by investigators at Montana State University [Francis, G. E., Adams, J.P., and Noonan, E.J., "Do they stay fixed?" *The Physics Teacher*, vol. 36, pp. 488-490, 1998.], as well as in other studies at the University of Colorado [Finkelstein, N.D., and Pollock, S.J., "Replicating and understanding successful innovations: Implementing tutorials in introductory physics," *Physical Review Special Topics–Physics Education Research 1: 010101-1–010101-13*, 2005], and the University of Cincinnati [Koenig, K.M., and Endorf, R.J., "Study of TAs' ability to implement the Tutorials in Introductory Physics and student conceptual understanding," in Marx, J., Franklin, S., and Cummings, K., (eds.), *2003 Physics Education Research Conference , Madison, Wisconsin, 6-7 August 2003, AIP Conference Proceedings 720,* (AIP, Melville, N.Y.), pp.161-164,2004.].

[46] Le, V-N, Stecher, B.M., Lockwood, J.R., Hamilton, L.S., Robyn, A., Williams, V.L., Ryan, G., Kerr, K.A., Martinez, J.F., and Klein, S.P., *Improving Mathematics and Science Education,* (RAND Corporation, Santa Monica, Calif.), 2006.

[47] Meltzer, D.E., and Manivannan, K., "Transforming the lecture-hall environment: The fully interactive physics lecture," *American Journal of Physics*, vol. 70, pp. 639-654, 2002.

[48] McDermott, L.C., and Redish, E.F., "Resource Letter: PER-1: Physics Education Research," *American Journal of Physics*, vol. 67, pp. 755-767, 1999.

[49] The model for the interview techniques we employ lies in the "individual demonstration interviews" pioneered at the University of Washington in the late 1970s. See, e.g., references 1 through 4. This interview format in turn was substantially motivated by the "clinical interview" technique employed by Piaget in the early 1900s; see, e.g., Piaget, J., *The Child's Conception of the World,* (Littlefield, Adams, Patterson, N. J.), 1963.

[50] For instance, see: Meltzer, D.E. "Investigation of students' reasoning regarding heat, work, and the first law of thermodynamics in an introductory calculus-based general physics course," *American Journal of Physics*, vol. 72, pp. 1432-1446, 2004.

[51] For instance, using the Thomson "Web of Science" index.

[52] Finley, G., "The Physical Science Study Committee," *The School Review,* vol. 70, pp. 63-81, 1962.

[53] Physical Science Study Committee, *Physics*, (Heath, Boston), 1960.

[54] Deboer, G.E., *A History of Ideas in Science Education: Implications for Practice,* (Teachers College Press, New York), Chaps. 7 and 8., 1991.

[55] Deboer, *ibid,* Chap. 8.

[56] Bruner. J., *The Process of Education*, (Harvard University Press, Cambridge, Mass.), 1960.

[57] Schwab, J.J., "The teaching of science as enquiry, in Schwab, J.J., and Brandwein, P.F., *The Teaching of Science,* (Harvard University Press, Cambridge, Mass.), pp. 1-103, 1962.

[58] Hurd, P.D., and J.J. Gallagher, *New Directions in Elementary Science Teaching,* (Wadsworth, Belmont, Calif.), 1969.

[59] Morrison, P., and Walcott, C., "Enlightened opportunism: An informal account of the Elementary Science Summer Study of 1962," *Journal of Research in Science Teaching*, vol. 1, pp. 48-53, 1963.

[60] Karplus, R., "The Science Curriculum Improvement Study," *Journal of Research in Science Teaching,* vol. 2, pp. 293-303, 1964.

[61] Arons, A.B., "Cultivating the capacity for formal reasoning: Objectives and procedures in an introductory physical science course," *American Journal of Physics,* vol. 44, pp. 834-838, 1976.

[62] McDermott., L.C., "Combined physics course for future elementary and secondary school teachers," *American Journal of Physics*, vol. 42, pp. 668-676; "Practice-teaching program in physics for future elementary school teachers," *American Journal of Physics,* vol. 42, pp. 737-742, 1974.

[63] McDermott, L.C., "Teacher education and the implementation of elementary science curricula", *American Journal of Physics*, vol. 44, pp. 434-441, 1976.

[64] Reference 56, p. 20.

[65] Atkin, J. M., and Karplus, R., "Discovery or invention," *The Science Teacher*, vol. 29, pp. 45-51, 1962; Karplus, R., "Science teaching and the development of reasoning," *Journal of Research in Science Teaching,* vol. 14, pp. 169-175, 1977.

[66] Piaget, J., "The new methods: Their psychological foundations, (1935)", in *Science of Education and the Psychology of the Child,* (Orion, New York), pp. 139-180), 1970.

[67] McKinnon, J.W., and Renner, J.W., "Are colleges concerned with intellectual development?" *American Journal of Physics*, vol. 39, pp. 1047-1052, 1971; Herron, J.D., "Piaget for chemists," *Journal of Chemical Education,* vol. 52, pp. 146-150, 1975.

[68] Reference 56, p. 48.

[69] Driver, R.P., "The Representations of Conceptual Frameworks in Young Adolescent Science Students," Ph.D. dissertation, University of Illinois at Urbana-Champaign, University Microfilms, UMI Number 7412006, 1973; Nussbaum, J., and Novak, J.D., "An assessment of children's concepts of the Earth utilizing structured interviews," *Science Education*, vol. 60, pp. 535-550, 1976; Novick, S., and Menis, J., "A study of student perceptions of the mole concept," *Journal of Chemical Education*, vol. 53, pp. 720-722, 1976. Following on these early studies, an enormous explosion of work on "alternative conceptions in science" arose and expanded throughout the field of K-12 science education. This is documented in exhaustive detail in: Wandersee, J.H., Mintzes, J.J., and Novak, J.D., "Research on alternative conceptions in science," in Gabel, D.L., (ed.), *Handbook of Research on Science Teaching and Learning*, pp. 177-210, (McMillan, New York), 1994.

[70] Viennot, L., Spontaneous reasoning in elementary dynamics," *European Journal of Science Education*, vol. 1, pp. 205-221, 1979.

[71] Trowbridge, D.E., and McDermott, L.C., "Investigation of student understanding of the concept of velocity in one dimension," *American Journal of Physics,* vol. 48, pp. 1020-1028, 1980.

Calibrated Peer Review for ABET Assessment

Frederick C. Berry and Patricia A. Carlson - *Rose-Hulman Institute of Technology*
Don Lewis Millard* *- Rensselaer Polytechnic Institute*

INTRODUCTION

Calibrated Peer Review (CPR) [1] is an online tool with four structured workspaces, listed below, that together create a series of activities reflecting modern pedagogical strategies for using writing in the learning process

- ***Task:*** *Students are presented with a challenging writing task, with guiding questions to act as scaffolding for the demanding cognitive activities.*

- ***Calibration:*** *Students read through three "benchmark" samples and assign each a score based on a series of evaluative questions (a rubric). Students are then given a "Reviewer Competency Index (RCI)" from 1 to 6, based on their demonstrated competency in these exercises. This segment mitigates the common objection to peer review in the undergraduate classroom: that the experience reduces itself to the blind leading the blind.*

- ***Peer Review:*** *After becoming a "trained reader" – and being assigned an RCI – students read and provide written feedback on three anonymous peer essays using the same rubric as used in the calibrations. Students also assign each essay a holistic score from 1 to 10.*

- ***Self-assessment:*** *As a final activity, students evaluate their own essay. As with calibration and peer review, students use the same rubric (set of performance standards for the task). Having "trained" on benchmark samples, and then applied their expertise in evaluating peer text, students engage in a reflective, final activity by assessing their own submission. Students are encouraged at this time to make comments for themselves, which are also available to the instructor, that capture the evolving insights they have gained in the previous two segments. They are also invited to reflect on whether they have gained a deeper level of understanding for the assignment and its outcomes.*

HOW WE APPLIED CPR

After some experimentation with CPR, we found that it could be used to assess the Accreditation Board for Engineering and Technology's (ABET's) Engineering Criterion 3(g) [EC3(g)] communication. Therefore, the Electrical and Computer Engineering Department at Rose-Hulman Institute of Technology built a course using CPR assignments to help our students develop proposals for their senior design projects. This course, ECE362 Principles of Design, is a junior level required course for all computer and electrical engineering students. The topics addressed in ECE362 include intellectual property, research methods, design specifications, conceptual design, scheduling, project management, business plan, market survey, and budgeting. The course culminates in a written proposal and oral presentation requesting funds for development of a product. The following CPR exercises are used in ECE362:

CPR 1: WHAT IS INTELLECTUAL PROPERTY (IP)

This CPR introduces IP in the form of patents, trademarks, industrial designs (trade secrets), and copyright law to the students. Patent protection is the major focus.

CPR 2: WHAT IS AN ANNOTATED BIBLIOGRAPHY

This CPR introduces students to research using the annotated bibliography. The annotated bibliography is used because it adds descriptive and evaluative comments (i.e., an annotation), assessing the nature and value of the cited works.

CPR 3: MARKET ANALYSIS

The students are introduced to two methods of market analysis coupled with project idea generation:

- ***Augmented Projects*** *are existing products that are* added to or supplemented *to extend their functionality. Such projects are the easiest to do because the base product is already developed. It is also easy to get market information on these types of products.*
- ***Bi-associated Projects*** *are projects that* combine two different *products and create a new product from the combination. These projects are more difficult to do because the combination of technologies or products may not be obvious. However, it is still easy to obtain market information for each product and then estimate a market if the two different products were combined into one product.*

CPR 4: PRODUCT DESIGN SPECIFICATION

A Project Design Specification (PDS) is a document that should reflect the common knowledge of the team about the project. The students make use of their preliminary research to develop environmental, performance, and technology specifications for their projects.

CPR 5: SOCIAL IMPACT STATEMENT

This CPR requires the students to reflect on their proposed project and write a social impact document using the IEEE Code of Ethics [3] as the rubric. For this assignment, the students write one or two pages about the impact of their project on society.

CPR 6: PROJECT TECHNICAL DESCRIPTION

The project technical description should provide a concise explanation, which is not overly technical, while frequently emphasizing the key benefits and incorporating appropriate visual elements. The three essential elements of the project technical description are:

1) ***Description:*** *It is important to start the description with a very concise description in order to put the features and benefits into context.*

2) ***Visual Element:*** *Picture, sketch, screen shot, or diagram that shows either the components of the product or how the product fits in its environment is usually helpful for the reader.*

3) ***Key Benefits:*** *State the key benefits of the product early. The use of bullet points is ideal. Then conclude stating the key benefits again in a paragraph form.*

The students produce their first draft of the project technical description using the information from the previous CPRs.

CPR 7: PROJECT TECHNICAL DESCRIPTION, AGAIN

The students next take the feedback from CPRs 1 to 6 and rewrite their project technical description considering these specific elements:

1. *Does the project technical description tell the reader what the product does in the opening paragraph or sentence?*

2. *Does the project technical description use concise and precise sentences along with concrete words to explain the product?*

3. *Does the project technical description use visual elements to help explain the product?*

4. *Does the project technical description present the key benefits of the product early in the description?*

5. *Does the project technical description present an analysis of any competitors?*

6. *Does the project technical description include an explanation of how the parts fit and function together?*

7. *Does the project technical description conclude with the key benefits of the product in paragraph form near the end of the description?*

8. *Does the project technical description convince you this project can be done?*

The students are also using the National Collegiate Inventors & Innovators Alliance (NCIIA) E-Team request for proposal [4] as a format guide for the project technical description.

CPR 8: PRODUCT DESIGN SPECIFICATION, AGAIN

A PDS is a document that will change substantially over the length of the project. There are many factors that will cause a PDS to change. But the one factor that will have the greatest impact is the development of a deeper understanding of the project. As the student teams move forward developing their project proposal, they will always need to think more intensely about their project. The PDS should reflect the common knowledge of the team about the project. Therefore, the PDS needs to be regularly refined during the proposal phase to reflect a deeper understanding of the team's project. The PDS is reviewed again using the following questions:

1. *Is a function list given with a short description for each project-function?*

2. *Are performance specification given for each function?*

3. *Is the operating environment for the project given?*

4. *Are specifications relating to the operating environment provided?*

5. *Are target technologies identified to meet all of the above?*

At this point, the PDS for each student team is very well structured.

CPR 9: SOCIAL IMPACT STATEMENT, AGAIN

This CPR requires the students to reexamine their proposed project and rewrite their social impact statement using the IEEE Code of Ethics as the rubric, especially focusing on item 1 of the code: "to accept responsibility in making decisions consistent with the safety, health and welfare of the public, and to disclose promptly factors that might endanger the public or the environment."

For this assignment the students write one or two pages about the impact of their project on society.

SATISFYING ABET (G)

Driskill [2], in examining how ABET's EC3(g) is addressed in available ABET accreditation plans, noted little evidence in the literature that assessment plans incorporate modern rhetoric pedagogy, contemporary discourse analysis, or the fundamentals of communication theory in their expectations for writing in an engineering education. Thus, the development of a rich definition of "communication" and measuring "effectiveness" by a set of carefully thought out exercises would be needed to assess EC3(g): "ability to communicate effectively". The following description, performance criterion and analysis are included from the ECE Department's ABET report, at Rose-Hulman Institute of Technology.

ABET (g): an ability to communicate effectively.

- ***Description:*** *graduates will demonstrate an ability to communicate effectively with written reports.*
- ***Performance Criterion:*** *70 percent of student-written reports have a low percentage of mistakes and normally contain an executive summary, social impact statement, project technical description, and project design specification.*
- ***Analysis:*** *This performance criterion is being satisfied.*

ECE362	AY 03-04	ECE	AY 04-05	ECE	AY 05-06	ECE
ABET g	Yes*	N**	Yes	N	Yes	N
Annotation	NA	NA	87%	70	87%	78
Project Design Specification Initial	79%	56	81%	48	92%	78
Project Design Specification Final	73%	56	84%	70	72%	78
Project Technical Description Initial	NA	NA	77%	70	91%	78
Project Technical Description Final	80%	56	74%	70	84%	78

* Yes – Percentages Meeting the Standard ** N – Number of Students Responses

CONCLUSION

From our preliminary work, CPR is proving to be an effective tool for assessing EC3(g). Additional research and data analysis is underway that may better frame the effectiveness of CPR as a tool for ABET.

REFERENCES

[1] Chapman, O.L., and Fiore, M.A., "Calibrated Peer Review™," *Journal of Interactive Instruction Development*, vol. 12, no. 3, pp.1-15, 2000.

[2] Driskill, L., "Linking Industry Best Practices and EC3(g) Assessment in Engineering Communication," *Proceedings, American Society for Engineering Education,* (2000).

[3] IEEE Code of Ethics, http://www.ieee.org/web/membership/ethics/code_ethics.html

[4] NCIIA E-Team RFP, http://www.nciia.org/grants_eteam.html.

Computer Science Assessment Instrument Development: Evaluating Attitudes and Outcomes

Barbara M. Moskal* and Nathan Behrens - *Colorado School of Mines*
Mark Guzdial and Allison Elliott Tew - *Georgia Institute of Technology*
Wanda Dann - *Ithaca College*
Steven Cooper - *St. Josephs University*

INTRODUCTION

A goal for many computer science education (CSE) research projects is to determine the extent to which a given instructional intervention has had an impact on student outcomes. Often, these outcomes include student attitudes and student learning. A challenge in these efforts is that valid and reliable assessment instruments that measure the necessary constructs are not currently available. Instead, each CSE research project is left to develop its own instruments. This approach results in several problems. First, most computer scientists are not trained in measurement and, therefore, are not familiar with the psychometric principles that guide the development of valid and reliable instruments. This can result in questionable interpretations. Second, without a common set of instruments, valid comparisons cannot be made across interventions.

The *Collaborative Research: Assessing Concept Knowledge and Attitudes in Introductory Computer Science Courses* (NSF, DUE-0512064) project is developing a set of assessment instruments designed to measure student learning outcomes and student attitudes in introductory computer science courses. These instruments are being designed to be used jointly in a program targeting both student learning and attitudes, or independently in a program targeting only one of these constructs. The validity and reliability of these instruments will be established through extensive testing on undergraduate student populations at four institutions: Colorado School of Mines, Georgia Institute of Technology, St. Josephs University, and Ithaca College. The need for these instruments has been established through the research efforts of the investigative team on several prior National Science Foundation (NSF) grants. This article reviews the motivation for this research effort and the progress of the research team during the first year of implementation. An overriding goal of this research is to provide the CSE research community with two assessment instruments, with established validity and reliability, that may be used to measure changes in student attitudes and learning outcomes in introductory computer science courses.

HISTORY

This project is a joint effort of investigators who are active in three higher education research teams: the Alice Curricular Materials, the Media Computation, and the Colorado School of Mines (CSM) Assessment Research teams. The first two teams have been independently investigating methodologies and approaches for introductory computer science teaching and learning. The current investigation reflects a combining of talents associated with these independent research efforts.

Alice Curricular Materials Team

The Alice Curricular Materials Team consists of researchers from Saint Joseph's University, Ithaca College, and Carnegie Mellon University. This team is working on a sequence of projects supported through NSF that are developing and testing curricular materials for teaching and learning fundamental programming concepts using

simulation and program visualization. The interactive, object-oriented animation environment software used in these projects is named Alice and is freely available via the internet [1]. The CSM Assessment team acts as external evaluators for the Alice Curriculum.

The Alice Curricular Materials Team and the CSM Assessment Team have collaborated on three NSF-sponsored programs since 2002. The overriding goal has been to develop curricular materials for introductory programming courses that are both appealing to students and that result in measurable gains in student learning with respect to basic computer science concepts. Each of the described projects used treatment and control groups to examine the effectiveness of the curricular materials on different demographic populations.

- *Decreasing Attrition Using Animated Virtual Worlds (NSF, DUE-0126833). This project was a proof of concept, designed to develop curriculum materials for introductory programming courses using simulation and visualization. As part of this project, the Alice Curriculum was developed and pilot tested at the investigators' respective institutions. The results of this investigation support that the use of the Alice materials had a positive impact on both student retention and performance in their first year computer science course at the participating institutions. For further details, see Moskal, Lurie and Cooper [2].*

- *Java-based Animation: Building viRtual Worlds for Object-oriented programming in Community Colleges (JABRWOC) (NSF, DUE-0302542). In this project, the Alice Curriculum Materials were implemented in three community college settings. The results of this investigation support the effectiveness of the Alice Curriculum for the improvement of students' understanding of basic computer science concepts. For more information on this research investigation, see Hutchinson [3].*

- *Program Visualization Using Virtual Worlds (NSF, DUE-0339734). This project is implementing workshops that provide faculty with instruction on the use of the Alice curriculum for teaching introductory computer science courses. Early results suggest that faculty positively evaluate both the workshop content and the Alice curriculum.*

Media Computation Team

The Media Computation team at the Georgia Institute of Technology is working on projects to develop curricular materials for teaching and learning fundamental programming concepts using media computation. Thus far, two projects have been implemented by this team.

- Media Computation as a Motivation and Structure for a Non-majors CS1 Class: "Data-first" Computing (NSF, DUE-0231176). The purpose of this poof-of-concept project was to create and evaluate the first offering of a non-majors course for introductory computer science. The content addressed in this course is based on the recommendations of the ACM/IEEE *Computing Curriculum 2001* standards [4]. The approach is to use a communications-oriented context to teach students how to create and manipulate multimedia, e.g., implementing PhotoShop-like image filters, reversing and splicing sounds, creating programs to mine Web pages, and generating animations. The results suggest a dramatic improvement in student retention when compared to prior semesters. For more information, see Guzdial [5] and Guzdial and Forte [6].

- Introduction to Media Computation: A new CS1 approach aimed at non-majors and under-represented populations (NSF, CCLI-0306050). This project extends the proof of concept previously described in two

directions: 1) further analyzing and tracking retention rates, and 2) developing Java-based materials to make the media computation approach accessible to more traditional introductory computer sciences courses. For more information, see Tew, Fowler and Guzdial [7] and Guzdial and Ericson [8].

Stimulus for Current Investigation

In the development and testing of the Alice and Media Computation Curriculums, the research teams came to similar conclusions: there is a lack of up-to-date, validated assessment instruments in the field of computer science education. Specifically, the teams desired instruments that would measure change in students' attitudes and interests in computer science, and students' knowledge of fundamental computer science concepts. If common instruments had existed, cautious comparisons could have been made across various introductory computer science curriculums, including the Alice and Media Computation approaches. The availability of such instruments would, in general, support the field of computer science education in identifying promising instructional approaches for further investigation. This was the stimulus for the combined research effort described here.

CURRENT RESEARCH INVESTIGATION

The purpose of the *Collaborative Research: Assessing Concept Knowledge and Attitudes in Introductory Computer Science Courses* project is to develop and validate two assessment instruments: 1) a computer science attitude survey, and 2) a student outcomes instrument for measuring fundamental computer science concepts. The second instrument is being designed in a manner that seeks to measure fundamental concepts within computer science that are not language specific. The goals and outcomes of this research effort can be summarized as follows. For clarity, goals are labeled with the prefix "G" and outcomes with the prefix "O".

G1. Create a computer science attitudes survey that measures the undergraduate students':

Oi. *confidence in their own ability to learn computer science skills;*
Oii. *perceptions of computer science as a male field;*
Oiii. *beliefs in the usefulness of learning computer science;*
Oiv. *interest in computer science; and*
Ov. *beliefs about professionals in computer science.*

G2. Create a learning outcomes assessment instrument for introductory programming courses that measures the following:

Oi. *fundamental concepts in programming;*
Oii. *algorithmic thinking in programming; and*
Oiii. *problem solving through programming.*

METHODS AND PRELIMINARY RESULTS

The process that is being used to develop and validate the two assessment instruments is outlined in Table 1. These steps are consistent with the recommendations of the *Standards for Educational and Psychological Testing* [9]. To date, the first three steps are underway for each of the instruments.

Table 1. Instrument Development Process

1.	Review of literature to define content and construct domain
2.	Development of questions to match proposed content and construct domain
3.	Expert confirmation of content and construct domain
4.	Pilot testing
5.	Third party expert review of proposed questions and instruments
6.	Large-scale testing
7.	Internal consistency coefficients use to examine internal reliability
8.	Inter-rater and intra-rater reliability examination based on scoring rubric
9.	Factor analysis
10.	Reliability coefficients for equivalent forms examination
11.	Full pilot across all classes, results examination for consequential validity

Attitudes Survey

The important areas in computer science that were reviewed for the purpose of informing the development of the attitudes survey included research on attraction, retention, and gender issues in computer science [10,11]. Abstracts and research articles were also reviewed on the development and use of interest inventories and attitude surveys in computer science [12, 13] as well as research in measurements concerning survey development and analysis [14]. A summary of this effort will be reported in the master's thesis of Nathan Behrens [15].

Based on this research, the team at the Colorado School of Mines created a preliminary set of 50 attitude survey questions designed to be aligned with the previously described attitude outcomes. A discussion ensued within the larger research team concerning the appropriateness of these questions with respect to the defined content and construct domain. Jointly, this team represents expertise in introductory computer science and assessment. Revisions to this instrument were made based on the larger team's feedback before the instrument was pilot-tested on approximately 300 undergraduate students at the Colorado School of Mines. A sample of the piloted questions for each of the outcomes is displayed in Table 2. An analysis of these results is currently underway.

A major result of the previously described process is the determination that a single attitudes instrument is unlikely to serve the needs of the broader computer science community. The current instrument is designed to measure student attitudes toward the field of computer science. However, a major component of computer science is programming. Combining questions that address the construct of computer science as a field with the construct of programming in a single assessment would result in a long instrument that most students would not complete. Furthermore, the two constructs are typically of interest for two different student populations. Therefore, the investigators are currently developing two attitudes instruments. The first is designed to measure the general student populations' attitudes toward computer science. This instrument will be useful for understanding the appeal of computer science as a field of study. The second instrument will measure students' attitudes toward programming. This instrument will be designed to specifically target attitudes once students have entered the field of computer science and have started to complete their programming courses.

Table 2. Sample Attitude Survey Questions *(These questions are in draft form. Send feedback to: nbehrens@mines.edu)*

Outcome	Sample Question
Oi. confidence in their own ability to learn computer science skills	• I am certain that I can apply computer science concepts. • I doubt that I can solve computer science problems.
Oii. perceptions of computer science as a male field	• Men and women are equally capable of solving computer science problems. • Men produce higher quality work in computer science than women produce.
Oiii. beliefs in the usefulness of learning computer science	• Knowledge of computer science will allow me to secure a good job. • Taking computer science courses is a waste of my time.
Oiv. interest in computer science	• Problems I find interesting have computer science components. • I would not take extra computer science courses if I was given the opportunity.
Ov. beliefs about professionals in computer science	• Being good at computer science is a positive quality. • A student who performs well in computer science will probably have a limited social life.

LEARNING OUTCOMES

Research concerning curriculum recommendations for introductory computer science courses was also thoroughly reviewed. To avoid bias from a particular language or pedagogical approach, the research team at the Georgia Institute of Technology identified the two most commonly used introductory computer science textbooks (CS1) from each of the major computer science publishing companies. An analysis was completed on the table of contents of the twelve selected textbooks to identify common content and constructs. The resultant list of concepts was refined based on the following: 1) the organizing framework of the Computer Science volume of Computing Curricula 2001 [4]; 2) a content analysis of the canonical texts representing each of the common introductory approaches (objects-first [16, 17], functional-first [18], and imperative-first [19]); and 3) recent textbooks designed to appeal to a media-driven student population that have been developed by the current collaborative team [20, 21].

Based on this work, the research team identified ten key concepts: programming basics (variable, assignment, integer, mathematical operators), logical operators, selection (if), definite loop (for), indefinite loop (while), function parameters, function return values, array, recursion, and object. Each concept will be tested at three different levels:

- ***Definition*** *– How well does the student understand the definition and use of the concept?*

- ***Trace*** *– How well does the student understand a problem using the concept, and can the student correctly anticipate the execution of written code using the concept?*

- ***Write*** *– Can the student correctly identify which line(s) of code is required to correctly implement a concept?*

Example questions for each type of question are included in Table 3.

Table 3. Sample Learning Outcomes Questions *(These questions are in draft form. Send feedback to: nbehrens@mines.edu)*

Question Level	Sample Question

Definition	Concept: Indefinite Loop (while)
	
	i = 1
	WHILE i <= 10 DO
	  DISPLAY "Number" i
	ENDWHILE
	
	What does the WHILE statement do?

Trace	Concept: Array
	
	array1 = [4, 5, 3, 6, 2, 7, 1]
	array2 = [7, 4, 2, 1]
	
	array1[3] = array1[5]
	
	array1[2] = array2[2]
	
	array1[4] = array2[3] + 5
	
	array1[6] = array1[3]
	
	IF array1[1] > array2[1] THEN
	  array1[1] = array1[1] + 2
	ENDIF
	
	What is the value of array1 after this code is executed?

Write	Concept: Definite Loop (for)
	
	n = 4
	
	FOR XXX1XXX DO
	  FOR XXX2XXX DO
	  DISPLAY XXX3XXX,
	  ENDFOR
	  DISPLAY "!"
	ENDFOR
	
	Which code fragments complete the method such that the output is:
	
	1!
	22!
	333!
	4444!

The research team at the Georgia Institute of Technology created a preliminary set of 30 questions designed to be aligned with the previously described learning outcomes. This instrument is currently being revised and is

expected to be pilot tested at the Georgia Institute of Technology during the 2007-2008 academic year. Based on the results of the pilot test, the instrument will be further revised and a scoring method that attends to validity and reliability will be developed and tested.

CONCLUDING REMARKS

The primary factor that has motivated this research is the desire to provide the computer science education research community with validated assessment instruments that measure students' attitudes toward computer science and their learning within introductory computer science courses. The existence of such instruments will support valid comparisons among various instructional approaches and thereby contribute to the advancement of instruction within the field. Table 1 reflects the broader research plan for this project. One year into the project, the first three steps are in progress with respect to both assessment instruments. The second year is expected to consist of further pilot testing, statistical analysis, and third party review. The deliverables at the conclusion of the project will be the availability of thoroughly tested assessment instruments.

ACKNOWLEDGMENTS

The authors wish to thank Ashlyn Hutchinson and Peter Dewitt, who are members of the Colorado School of Mines Alice assessment team. Their contributions have been essential to this research effort. The authors would also like to thank the NSF, which has provided funding for all of the research projects described here.

REFERENCES

[1] Carnegie Mellon University, Alice Software, http://www.alice.org, 2007.

[2] Moskal, B., Lurie, D., and Cooper, S., Evaluating the effectiveness of a new instructional approach, In *Proceedings of the Special Interest Group on Computer Science Education,* Cincinnati, Ohio, 2004.

[3] Hutchinson, A., *A Statistical Analysis of Outcomes in an Educational Assessment: Impact of the Alice Curriculum On Male and Female Performances and Attitudes at Community Colleges*, Master's Thesis, Colorado School of Mines, 2005.

[4] Joint Task Force on Computing Curricula, Computing Curricula 2001--Computer Science, *Journal of Educational Resources in Computing*, vol. 1 (3es), pp. 1-240, 2001.

[5] Guzdial, M., "A media computation course for non-majors," *Proceedings of the 8th Annual Conference on Innovation and Technology in Computer Science Education, Thessaloniki, Greece*, pp.104-108, 2003.

[6] Guzdial, M. and Forte, A., "Design process for a non-majors computing course," *Proceedings of the 36th Technical Symposium on Computer Science Education, St. Louis, Mo.,* pp. 361-365, 2005.

[7] Tew, A., Fowler, C., and Guzdial, M., "Tracking an Innovation in Introductory CS Education from a Research University to a Two-year College," Special Interest Group in Computer Science Education, 2005.

[8] Guzdial, M. and Ericson, B., *Introduction to computing & programming in Java: A multimedia approach* (Pearson Prentice Hall, Upper Saddle River, N.J.), 2006.

[9] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, (American Educational Research Association, Washington, D.C.), 1999.

[10] Margolis, J., and Fisher, A., *Unlocking the clubhouse: Women in computing*, (MIT Press, Cambridge Mass.), 2003.

[11] Margolis, J., Fisher, A., and Miller, F., "Caring about connections: Gender and computing," *IEEE technology and society magazine*, Winter, pp.13-20, 2000.

[12] Educational Testing Service, http://www.ets.org/portal/site/ets/menuitem.3a88fea28f42ada7c6ce5a10c3921509/vgnextoid=85b65784623f4010VgnVCM10000022f95190RCRD.

[13] Educational Resources Information Center, http://www.eric.ed.gov/.

[14] Suskie, L.A., *Questionnaire Survey Research*: What Works (2nd ed.), (Association for Institutional Research, Florida State, Tallahassee, Fla.), 1996.

[15] Behrens, N., *Examining Student Attitudes in Computer Science Education* (preliminary title), Master's Thesis, Colorado School of Mines, expected 2007.

[16] Deitel, H., and Deitel, P., C++: *How to Program* (5th ed.), (Prentice Hall, Upper Saddle River, N.J.), 2005

[17] J. Lewis and W. Loftus, *Java Software Solutions (Java 5.0 version): Foundations of Program Design* (4th ed.), (Addison Wesley, Boston, Mass.), 2005.

[18] Felleisen, M., Findler, R. B., Flatt, M., and Krishnamurthi, S., *How to Design Programs: An Introduction to Programming and Computing*, (MIT Press, Cambridge, Mass.), 2001.

[19] Zelle, J. M., *Python Programming: An Introduction to Computer Science,* (Franklin Beedle, Wilsonville, Ore.), 2004.

[20] Dann, W., Cooper, S., and Pausch, R., *Learning to Program with Alice,* (Pearson Prentice Hall, Upper Saddle River, N.J.), 2006.

[21] Guzdial, M., *Introduction to computing and programming in Python: A multimedia approach,* (Pearson Prentice Hall, Upper Saddle River, N.J.), 2005.

Designing a Peer Evaluation Instrument that is Simple, Reliable, and Valid: The Comprehensive Assessment of Team-Member Effectiveness

Matthew W. Ohland* - *Purdue University*
Misty L. Loughry - *Clemson University*

INTRODUCTION

The Comprehensive Assessment of Team-Member Effectiveness (CATME) is used for peer-evaluation and self-evaluation of individual team members' contributions to the team. The instrument's development was grounded in theory from the teamwork literature, incorporating all of the team-member behaviors shown by prior research to be important for effective team functioning [1,2,3,4].

Other peer evaluation instruments have been used in engineering education. In 2003, when this work began, the most widely used were the Professional Developer [5,6,7] (which had been released earlier as the Team Developer) and derivatives of an instrument developed by Robert Brown of the Royal Melbourne Institute of Technology [8] made popular by Kaufman, Felder, and Fuller [9].

The Kaufman, Felder, and Fuller instrument was essentially a single-item survey, in which students assigned one of nine descriptive words to describe each team member's level of contribution to the team (no-show, superficial, … up to… very good, excellent). The results from using this instrument were a challenge to process, especially for large enrollment classes. Further, we felt that the single-item nature (where raters make a single rating capturing a wide variety of behaviors) made it unlikely that the instrument could be administered with much reliability. The instrument was paper-and-pencil only.

The Professional Developer asks 50 questions, can be used for 360-degree reviews, and is designed primarily for use in formative assessment—to improve teaming skills. Although the Professional Developer is offered in an online format, the length of the survey is a burden on the students.

The project team for the ASA-funded project "Designing a Peer Evaluation Instrument that is Simple, Reliable, and Valid," including the authors, Lisa F. Bullard, Richard M. Felder, Cynthia J. Finelli, Richard A. Layton, and Douglas G. Schmucker, sought to develop an instrument that was more reliable and valid than the single-item instrument, but was simpler to use than a Likert-scale instrument with a large number of questions.

At the same time as we began to develop CATME, James Sibley of the University of British Columbia was developing the iPeer system [10]. iPeer is an application to develop and deliver rubric-based peer evaluations, but cannot be validated, because it is only a framework—faculty members choose their own questions.

We believed that the most compelling need among engineering education faculty was the creation of a summative instrument that could be used multiple times to gain some formative data and improve teams, but that was primarily designed to measure the team effectiveness of each team member at a given

point in time. We also wanted to create an instrument that was grounded in theory from the teamwork literature. CATME was developed by reviewing existing research on teams in order to incorporate all of the team-member behaviors that research suggests are important for effective team functioning. Empirical studies were then conducted to identify what specific attributes should be measured and to identify which behaviors were most closely linked to those attributes.

This research resulted in three main instruments. The first two, which are Likert-scale measures (statements with which raters indicate their level of agreement or disagreement, see Figure 1), are described in an article that is currently in press with *Educational and Psychological Measurement* (the research was also presented at conferences of the American Society of Engineering Education and Southern Management Association and was published in both conferences' proceedings). The first instrument contains 87 items that measure 29 specific ways to contribute to the team, each with three items that have high internal consistency ($\alpha > 0.7$), meaning that they all measure the same thing. This instrument is intended for use in instructing team members how to contribute to the team, as well as for self- and peer-evaluation of past team member performance. In addition, the instrument was designed so that researchers or instructors who want to measure any of the 29 specific ways of contributing to the team could select from the 29 sub-scales in order to have a parsimonious measure of a relevant behavior. The 29 sub-categories were found to cluster in five broad categories, which were: 1) contributing to the team's work, 2) interacting with teammates, 3) keeping the team on track, 4) expecting quality, and 5) having relevant knowledge, skills, and abilities. The second Likert-scale instrument is a 33-item short version of these five broad categories.

Figure 1. Sample Likert rating describing a student's contribution to the team's work. Students must rate each teammate on each statement.

1 - Strongly Disagree	2 - Disagree	3 - Neither Agree Nor Disagree	4 - Agree	5 - Strongly Agree
Did a fair share of the team's work.				
Fulfilled responsibilities to the team.				
Completed work in a timely manner.				

The third instrument developed as part of this research uses a behaviorally anchored rating scale (BARS) format to assess the five broad categories identified when creating the Likert-scale instrument. A BARS instrument uses descriptions of specific behaviors that are typical of high, medium, and low performance for each category of performance assessed (Figure 2). The rater selects the description from the instrument that most closely matches that behavior displayed by the person being rated.

Figure 2. Sample BARS rating describing a student's contribution to the team's work. Students select the category of behavior that best describes each teammate.

• Does more or higher-quality work than expected.
• Makes important contributions that improve the team's work.
• Helps to complete the work of teammates who are having difficulty.
Demonstrates behaviors described in the row just above and just below.
• Completes a fair share of the team's work with acceptable quality.
• Keeps commitments and completes assignments on time.
• Fills in for teammates when it is easy or important
Demonstrates behaviors described in the row just above and just below.
• Does not do a fair share of the team's work. Delivers sloppy or incomplete work.
• Misses deadlines. Is late, unprepared, or absent for team meetings.
• Does not assist teammates. Quits if the work becomes difficult.

Because this instrument includes five categories of team-member contributions, raters make only five decisions about each person being rated. Therefore, the BARS measure is much simpler and quicker for raters to use. While the Likert-scale instrument contains more information, and could potentially make it clearer to team members in what areas they need to improve, the BARS instrument requires fewer ratings from each student, encouraging students to be more reflective during the rating process. Research also suggests that there are higher levels of inter-rater agreement when raters use a BARS instrument instead of a Likert-scale instrument.

After the development of the instruments was completed, the research team developed a system for the online administration of the BARS version of the instrument. Online administration has several important advantages. First, it is much easier for raters to complete than paper-based scales because of the elegant design of the online system. Second, online administration gives raters greater privacy when rating teammates, which encourages more accurate ratings than when ratings are performed in proximity to teammates, such as in a typical classroom setting. Third, the online system provides timely feedback to team members about how their self-ratings compare to peers' ratings in each category that is measured. Fourth, the system flags special conditions where the ratings indicate that there may be a problem with the team or a team member, making it easier for faculty to intervene in teams that need help. Finally, the online administration greatly reduces the record-keeping burden for faculty members. This allows faculty to administer peer evaluations multiple times in a semester without placing an undue burden on their time.

THE ONLINE INTERFACE

The online system, located at www.catme.org, has been well received by faculty—125 faculty from 40 different institutions are now registered to use the system. Multi-faculty administration was recently added to the system to facilitate its use in large-enrollment classes. A demonstration mode now enables faculty to see the instrument from the student perspective, and a tutorial has been developed to make it easier for faculty (especially those who are not technologically savvy) to get started using the instrument. In the future, Team-Maker, a program developed by Richard Layton and his colleagues, will be added to the website. This program will assign members to teams based on criteria that instructors specify (such as schedule availability, demographic characteristics, or student majors). The use of the system continues to grow, and this new feature will help draw additional users.

One of the most valuable features of the system is that it alerts faculty regarding exceptional conditions that provide information about teams and team members.

- ***Low****—a student who rates him/herself as ineffective and who also receives "ineffective" ratings by teammates.*
- ***Overconfident****—a student rated as "ineffective" by teammates but rates him/herself as much more effective.*
- ***High****—a student who is rated as highly effective according to both teammate and self-ratings.*
- ***Underconfident****—a student rated as highly effective by teammates but who underrates her/himself.*
- ***Manipulator****—a student who rates him/herself as highly effective and who receives "ineffective" ratings by teammates. Such a student may be trying to unfairly influence the distribution of grades.*
- ***Conflict****—a team in which there is considerable disagreement among the various raters about the effectiveness of an individual student.*
- ***Clique****—a team in which cliques appear to have formed. The ratings show that subsets of the team rate members of their subset high and members of other subsets low.*

Most importantly, some of these conditions have more than one explanation. A student flagged as a "manipulator" might actually have performed a disproportionately large amount of the work on the project even though they worked to engage their teammates in the process. Thus, an instructor's involvement and judgment are critical when exceptional conditions are flagged. Though the formal study of these exceptions has not been completed, faculty using the system have reported that both the clique and conflict conditions have accurately provided early warnings of those conditions. Information on the design of the instrument and research supporting its use (including validity studies) can also be found at *www.catme.org*.

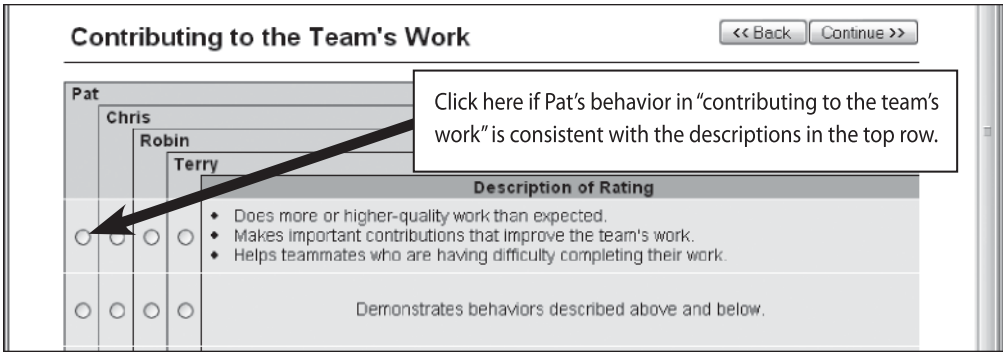
The CATME website is a secure interface for collecting data on team member effectiveness and reporting different views of the data to faculty and students. The CATME system has a number of convenient features—the ability to upload student and team data from files generated by Excel; support for multi-section courses and teaching assistants; and the ability to edit teams, reset surveys, send email reminders, and track survey completion. The system also allows students to make comments for instructors to read and can compute grade adjustments based on how the ratings patterns compare with faculty-specified criteria.

Faculty can request an account at *www.catme.org*. The process of defining a class and setting up teams is wizard-based, but a tutorial is available. Figure 3 shows several typical screen shots from the site.

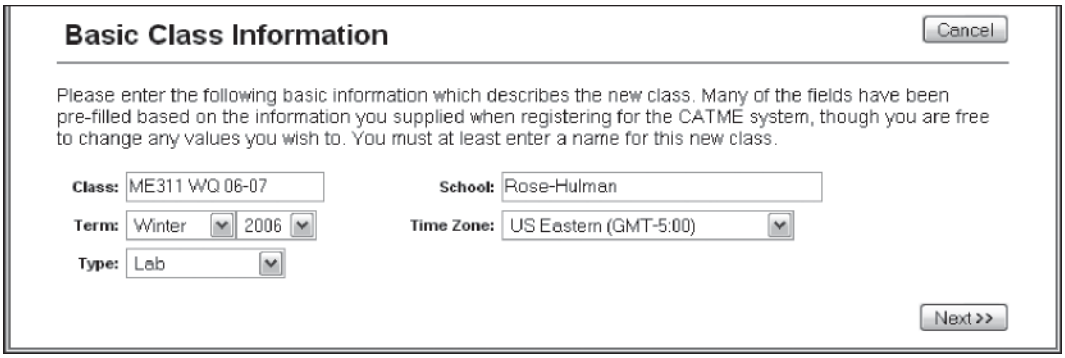
ONGOING WORK

Data analysis efforts from the administration of the BARS instrument at various partner schools are ongoing and are expected to demonstrate the validity of the instrument in different ways. The most interesting finding from a large dataset of Clemson University management classes is that it is vital to consider multi-level issues when analyzing the data. The norms for interpreting and using the rating instrument appear to be established at the team level, but differ among teams. This does not affect the most common use of the instrument, in which student grades are adjusted relative to their teammates. However, it has critical implications for analyzing data for research purposes. Additional data collected from Rose-Hulman, which collects data before and after a training intervention aimed at calibrating student ratings (known as a time-series design), is expected to show an increase in inter-rater agreement after the training.

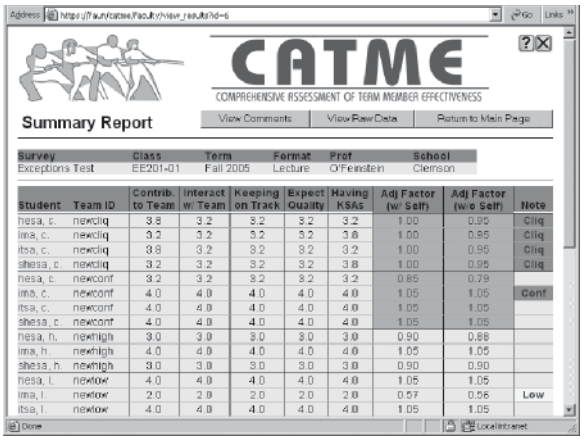
Figur 3. ATME screen shots.



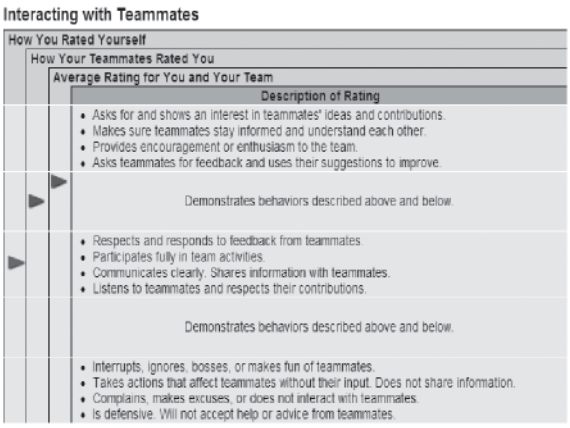
The first of five rating categories: Contributing to the Team's Work.



The wizard-based interface for class creation in CATME is typical of other setup screens.



Faculty summary results (raw data available).



Student results: by self, by team, average

SUMMARY

The research advances the knowledge or understanding across different disciplines by involving the perspectives of the academic communities in engineering and management. An understanding of student teams emerges that transcends either of those disciplines alone. Particularly, the issue of rater accuracy is central in both communities and in a professional context.

The research enhances student learning by developing a peer evaluation instrument that can be used in a formative sense to guide the development of team skills. Further, the incorporation of the Team-Maker team formation system will simplify the process of forming teams through criteria that are known to improve student learning. Training materials are being developed that are particularly targeted at improving teaming skills of individual team members based on feedback from the peer evaluations. Extensions of this project advance discovery and innovation at the frontiers of STEM learning by developing a deeper understanding of team dynamics and generating hypotheses from one discipline (management) that can be tested and refined in another (engineering) in the support of content learning in the STEM disciplines.

The research guides curricular and pedagogical development by supporting a cooperative learning environment through establishing individual accountability [11], providing valuable information about team member performance and behavior [12,13], helping to educate individuals about roles and expectations [14], and facilitating cross-training and the development of shared mental models within the team [15].

The research has a broader impact on issues in higher education by increasing the capacity and breadth of the research and evaluation communities contributing to these frontiers, in that veteran researchers who are well-connected members of other communities are being engaged in research on engineering education. Attracting new faculty to use peer evaluation to facilitate the adoption of cooperative learning has significant benefits in engineering education.

ACKNOWLEDGEMENTS

The other members of the research team and their institutional affiliations are: Lisa G. Bullard and Richard M. Felder of North Carolina State University, Cynthia J. Finelli of the University of Michigan, Richard A. Layton of Rose-Hulman Institute of Technology, and Douglas G. Schmucker of Tri-State University.

REFERENCES

[1] Ohland, M.W., Finelli, C.J., Bullard, L., Loughry, M., and Layton, R.A., "Designing a peer evaluation instrument that is simple, reliable, and valid," NSF Award No. 0243254, Division of Undergraduate Education (DUE), CCLI-ASA Program, Jun 03-May 07.

[2] Loughry, M. L., Ohland, M. W., and Moore, D. D., "Development of a theory-based assessment of team member effectiveness," *Educational and Psychological Measurement*, in press.

[3] Ohland, M.W., Pomeranz, H.R., and Feinstein, H.W., "The Comprehensive Assessment of Team Member Effectiveness: A New Peer Evaluation Instrument," *Proceedings of the American Society of Engineering Education Annual Conference*, Chicago, Ill., June 2006, 5 pp.

[4] Loughry, M. L., Ohland, M. W, and Moore, D. D., "Behaviorally Anchored Peer Evaluation of Team Member Effectiveness," *Proceedings of the Annual Meeting of the Southern Management Association*, 2006.

[5] McGourty, J. and De Meuse, K., *The Team Developer: An assessment and skill building program*, (J. Wiley and Company, New York), 2000.

[6] McGourty, J., Sebastian, C., and Swart, W., "Development of a comprehensive assessment program in engineering education," *J. Engineering Education*, vol. 87, no. 4, pp. 355-361, 1998.

[7] McGourty, J., "Using multisource feedback in the classroom: A computer-based approach," *IEEE Trans. Ed*., vol, 43, no.2, pp.120-124, 2000.

[8] Brown, R.W., "Autorating: Getting individual marks from team marks and enhancing teamwork," *Proc*. *Frontiers in Education Conference*, IEEE/ASEE, Pittsburgh, November, 1995.

[9] Kaufman, D.B., Felder, R.M., and Fuller, H., "Accounting for Individual Effort in Cooperative Learning Teams," *J. Engr. Education*, vol. 89, no. 2, pp. 133-140, 2000.

[10] Sibley, J., iPeer, v. 2.0, http://ipeer.apsc.ubc.ca/ipeer_site/?page=about.

[11] Johnson, D.W., Johnson, R.T., and Smith, K.A., *Active learning: Cooperation in the college classroom*, (Interaction Book Col, Edina, Minn.), 1998.

[12] Viswesvaran, C., Schmidt, F. L., and Ones, D. S., "Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences," *Journal of Applied Psychology,* vol. 90, pp. 108-131, 2005.

[13] Latane, B., Williams, K., and Harkins, S. "Many hands make light the work: The causes and consequences of social loafing," *Journal of Personality and Social Psychology*, vol. 37, pp. 822-832, 1979.

[14] Sheppard, S., Chen, H. L., Schaeffer, E., Steinbeck, R., Neumann, H., and Ko, P., "Peer assessment of student collaborative processes in undergraduate engineering education," Final Report to the National Science Foundation, Award Number 0206820, NSF Program 7431 CCLI-ASA, 2004.

[15] Marks, M. A., Burke, C. S., Sabella, M. J., and Zaccaro, S. J., "The impact of cross-training on team effectiveness," *Journal of Applied Psychology,* vol. 87, pp. 3-13, 2002.

Development of Faculty Collaboratives to Assess Achievement of Student Learning Outcomes in Critical Thinking in Biology Core Courses

Joyce K. Ono* - *California State University, Fullerton*

Project collaborators: Merri Lynn Casem, Bill Hoese, Anne Houtman, Judith Kandel, and Elaine McClanahan - *California State University, Fullerton*

OVERVIEW

From 1997 to 2002, the Department of Biological Science at California State University Fullerton (CSUF) embarked on a revision of the core curriculum for biology majors. The process used involved almost all members of the department and utilized the expertise of a professional facilitator. The end result was not only a revision of the course components of the biology majors' core curriculum, but a transformation in how we teach and develop courses and curricula. We were in the first group of ASA grantees and our project was to continue the use of faculty collaborative groups to develop assessments of critical thinking and problem solving (CT/PS) in biology as part of an ongoing faculty development process initiated with a previous NSF Course, Curriculum, and Laboratory Improvements (CCLI) grant. In addition, we developed a plan to examine the effects of the altered learning environment in our core biology courses on student achievements but were unable to develop a database to maintain assessments of CT/PS that faculty could share and utilize in their classes.

The major achievement of our NSF support has been the development of processes for faculty collaboration in defining CT/PS in biology and a template for creating assessments that incorporate several of the elements of our CT/PS definition. This template, and the resultant assessments, has been utilized in the first two biology core courses and in an introductory non-majors' biology course. We also share some of the approaches and parts of the project that had been proposed but did not work for us in hopes these lessons learned might be helpful to others.

THE IMPETUS TO REQUEST NSF ASA SUPPORT

With the support of an NSF CCLI grant in 1998 [1], the Department of Biological Science at CSUF expanded a major revision of the biology major's core curriculum to transform how we develop new curricula and how we teach. A professional facilitator, Elaine McClanahan, utilizing the "backwards design" approach [2], helped us to reach consensus on how to revise our core curriculum. Faculty with expertise in major areas of biology that corresponded to the four new core biology courses (Evolution and Biodiversity, Cellular Basis of Life, Genetics and Molecular Biology, and Principles of Physiology and Ecology) were grouped into Teaching Collaboratives to identify and develop student learning outcomes, namely, what we wanted students to know and be able to do when they had completed the entire core curriculum and at the end of each core course. We stated these outcomes in the form of "assessable" verbs, as described in a revision of Bloom's Taxonomy [3], to guide instructors about the depth of understanding for each of the major learning goals. This collaborative approach for developing courses and curricula was novel for us and required extensive communication among faculty as well as faculty development, guided by our professional facilitator.

We also realized that many of the learning goals would not be achievable through the typical lecture format. Constrained by the need for large lecture classes to serve the hundreds of biology majors, we began to learn to utilize active learning strategies in these core classes. In the smaller laboratory classes major changes were made to create

more inquiry-based laboratory and field experiences to replace didactic, cookbook laboratories (see descriptions in [7]). As a result, it took us 5 years to finally offer all four of the new core courses in spring 2003 with a transition period that lasted for another 3 years before the first cohort of graduates had completed the new core curriculum. We had made major shifts in the reward system in our personnel document, re-defining scholarly activity to include research in teaching and learning. This paved the way for hiring biology faculty with interests in research on teaching and learning, and who could serve as resources within our department for faculty development.

In 2001, we applied for an NSF ASA grant, during the inception of the ASA program, to help us continue with faculty development initiated in the first NSF award. We proposed to create assessments relevant to biology, of what we felt were the most important attributes of a biologist: the ability to think critically and solve problems in biology. We had earlier utilized a more generic assessment of CT/PS, the California Critical Thinking Disposition Inventory (CCTDI) [4], but could not determine from the results what contribution our biology curriculum had made to the development of these skills in our majors. In addition, we proposed to create a database of faculty-contributed assessments of CT/PS as well as those aligned to the learning outcomes for our courses. This departmental database would also incorporate demographic information about our students so faculty could develop a profile of the students in their classes at the start of each term to facilitate establishing an appropriate learning environment. The new faculty we had hired, who are collaborators in the grant, were to conduct research on the impact of the learning environment on student achievements. These were ambitious goals and not all of them were achievable in the grant period, but major, long-lasting gains and dissemination of our processes to other departments and other CSU campuses have resulted from the support of both NSF grants.

COLLABORATIVELY DEVELOPING A WORKING DEFINITION FOR CT/PS IN BIOLOGY

In a faculty retreat, we brainstormed the attributes of CT/PS using an affinity process [5]. This process is commonly used for group brainstorming and consists of collecting information from each team member anonymously (we used statements written on Post-Its®), having teams categorize the individual contributions by their affinity to each other, and analyzing the resulting groupings. We used this method to generate our first version of a definition of CT/PS, which the department accepted in February 2000, prior to our application for support in the new NSF ASA program. Retrospectively, we had utilized the Delphi Method, in which experts convene to reach consensus about complex issues that are often a matter of opinion. A group convened by the American Philosophical Society used this method to develop the attributes of critical thinking that were later utilized to develop the CCTDI [6].

In 2002, with support from the ASA grant, we revised our CT/PS definition utilizing Bloom's revised Taxonomy [3] to conceptualize and organize our CT/PS definition. We also collected samples of assessments contributed by our faculty, both to refine our definition and to illustrate the elements of a definition for CT/PS in biology. Bloom's taxonomy is organized hierarchically into six levels of increasing cognitive difficulty [3]. Briefly, the levels from lowest to highest are: knowledge, understanding, application, analysis, synthesis, and evaluation, with verbs that describe the kinds of assessment or activities associated with each level. The higher levels correspond closely to the cognitive skills (interpretation, analysis, evaluation, inference, explanation, and meta-cognition) generated for the development of the CCTDI in the widely acclaimed *Delphi Report* [6]. We used these verbs to synthesize a Working Definition of Critical Thinking and Problem Solving in Biology (Table 1). A packet containing this definition and sample assessments contributed by faculty was distributed for faculty to use as a guide in creating assessments of CT/PS for their courses.

The process for utilizing our CT/PS definition for biology is to select and focus on several elements of the definition and to develop course-related assessments that require students to use the appropriate thinking skills described in selected elements. A rubric that would allow for an objective and consistent analysis of student work should also be developed along with the assessment. Such rubrics are developed in an iterative fashion and revised over time with use. We had hoped to house these assessments and rubrics developed by our faculty in a secure database so faculty could share in their use and make further refinements.

DEVELOPMENT OF THE CTLEA (CRITICAL THINKING LONGITUDINAL EMBEDDED ASSESSMENT) TEMPLATE

Using the process for developing CT/PS assessments described above, collaborators of the grant, Drs. Merri Lynn Casem, Bill Hoese, and Anne Houtman, developed and tested embedded assessments of three elements of our CT/PS definition in core and intermediate biology majors' courses and in the introductory non-majors' biology course. These assessments were utilized in a pre- and post-assessment design to track student achievements longitudinally. A template for developing a CTLEA was created. The three parts of the template are: 1) provide the background of a biological problem, theory, or hypothesis; 2) present data in the form of a table or graph relating to the background description; and 3) ask three questions that will a) require students to interpret data accurately, b) represent biological data in an accurate, visual format, and c) make appropriate inference and deductions. This assessment focuses on three of the elements of the CT/PS definition that we had developed and could be used in almost any biology course at any level. It focuses on basic skills in interpreting graphs and tables and is flexible for use in probably any course requiring visualization of data and their interpretation.

The rubrics for evaluating the assessments given in the three large biology courses are based on standardized set criteria specific to each question. Rubrics were developed in collaboration with the instructor(s) for a specific course. Student responses were scored on a scale of 0 to 5 where 5 represents a complete and correct answer and 0 was a wrong or irrelevant answer. Student responses that included some components of a correct answer received an intermediate score. This is the second version of the rubric, which is still being refined by iteratively comparing the scoring of the same set of student responses by different faculty and graduate students. An example of a CTLEA used in a cell biology and a microbiology course, and the rubric for the first question, are provided in Figure 1. Two other versions of the CTLEA were developed for other core and non-majors' biology courses and are being tested and refined.

RESEARCHING THE EFFECTS OF THE LEARNING ENVIRONMENT ON STUDENT ACHIEVEMENT

Figure 2 summarizes the overall research plan for assessing the learning environment on student achievement in our revised curriculum. One of the proposed applications of the newly developed CTLEA is to use it to improve the learning environment for students by assessing their CT/PS skills. The overall plan was to determine what faculty were doing in their classroom (Faculty Survey of Instructional Modes and a revised Peer Review Evaluation form), what students' perceptions were of their learning environment (SALGains or Student Assessment of Learning Gains online surveys [8]), and assess student achievements as they completed core courses (CTLEA) and in an exit examination (Educational Testing Service [ETS] Biology Major Field Test). Except for the nationally based ETS examination, we developed the faculty survey, and specific SALGains surveys for the core courses, and the CTLEA.

A SALGains survey developed by M. Casem was used to assess student perceptions about the inquiry-based laboratories in the core courses from fall 2002 to spring 2005 and the results published in Cell Biology Education [7].

SAMPLE PRELIMINARY DATA AND RESEARCH PLANS

We have collected preliminary data on the use of the CTLEA in several introductory core biology courses and in the non-major's introductory biology course. We share the results of using the CTLEA as pre- and post-tests, administered respectively on the first week of classes and at the end of the introductory cell biology courses in 6 classes taught over three semesters, with three different instructors. The graph in Figure 3 compares the combined student scores for each of the three questions in the CTLEA, where each student's scores at the beginning of the course and at the end of the course were paired. Data were eliminated if a complete pairing was not available. These data indicate that although the differences in the means between the pre-and post-test scores for each question were small, there was significant improvement in the first ($p = 1.982 \times 10^{-9}$) and third ($p = 4.302 \times 10^{-5}$) CTLEA questions. There was no significant improvement in the scores for the second question over the course of the semester. In all cases, the variances were larger for the post-tests, and significantly different for questions one and three ($p < .0001$). The results from question 2 suggest that this question may not be effectively discriminatory. A ceiling effect was attained, perhaps because most students are capable of graphing data at this stage of their education or the task does not truly require critical thinking since another graph similar to it is provided in the test. Perhaps if the data were not ordered ready for graphing, this question would be more discriminatory and a better test of the intended CT/PS element. This second question will be revised.

The plan for validating the CTLEA depicted in Figure 1 is to compare it to student scores in the California Critical Thinking Skills Test (CCTST). This test has been in wide use for many years and has been validated through testing students from our campus before and after completion of critical thinking courses that are part of the general education requirement [9]. We plan to have a cohort of 50 students take the CCTST, and their scores compared with those on the CTLEA. We hypothesize that if the CTLEA is a valid test of critical thinking, there will be good correlation between student scores in the CTLEA and those in the CCTST. Demographic information, course grades, and grade point averages in the major will also be collected from student participants for additional correlative studies.

We plan to use the CTLEA and develop additional CTLEAs that target other elements of our CT/PS definition for programmatic assessment of our curriculum. The same CTLEA presented in Figure 1 is being administered in an upper division microbiology course and an intermediate cell biology course to determine whether there is an improvement in scores from lower to upper division courses. Other versions of the CTLEA have been developed for the first core course, Evolution and Biodiversity, and in the non-major's Elements of Biology course. These too are in the process of being evaluated. In a recent faculty retreat, faculty collaborated to develop CTLEAs for each Concentration that may be utilized in an exit examination.

LESSONS LEARNED: WHAT WORKED AND WHAT DIDN'T WORK

To prevent others from taking pathways that may not be productive, we would like to share some of the activities we attempted that did not work. A major goal of the grant was to develop a database of CT/PS assessments in which faculty would share what worked in their classes. In our grant proposal, we had proposed to use the Assessment Wizard software, developed by the Educational Testing Service (ETS) in collaboration with Grant Wiggins to help K-12 teachers develop assessments aligned with standards using an interactive approach. The software also promised to offer a site to store the assessments that were developed so that they would be shareable with a national database of assessments aligned to standards. We found the software unsuited for our uses and it is no longer available at the ETS website. We encountered difficulties in finding database developers who could translate what we wanted into a web-accessible database that could be queried.

A positive result of our attempts to develop a database that required information from our campus' Admissions and Records and Institutional Research departments, was the opening of communication with a number of academic support personnel who welcomed being consulted early in the process of developing a departmental database for assessment. A surprising development that we had not anticipated was faculty resistance to sharing their assessments in a computer-based database. The issue was fear of having their assessment information compromised and we did not have sufficient expertise in server security to provide the necessary assurance. Moreover, the department elected not to divert resources for maintaining the database and the additional server that would house it. We are currently planning to utilize the campus' online course management system, Blackboard, as a secure, faculty- accessible-only warehouse for faculty-developed assessments. Since many campuses have such services, this may be a more tenable and cost-effective solution.

In planning for faculty development of assessment skills, we had proposed to utilize several processes that are also used in K-12, because there are few for higher education faculty. However, we found that some of the approaches utilized by K-12 teachers did not work for us. Looking at Student Work [10] is a formal protocol for educators to collaboratively examine samples of student work as a means to learn about assessing student work. The Cellular and Developmental Biology Concentration faculty tested several protocols developed by our facilitators and after several sessions, this approach was found to be ineffective for our use of time. Our most fruitful faculty development came during annual departmental retreats, which always included a faculty development component, most often by workshop presentations from national leaders in assessment approaches.

We have discovered that developing assessments of CT/PS and their rubrics, which are validated and reliable, requires expertise that is beyond most faculty in a biology department. The typical 3-year grant period is insufficient time for developing large-scale assessments that have been proven to be valid and reliable when faculty development is an additional major goal. We would have profited by having collaborations with psychometric experts to devise strategies for testing the assessments we developed and to conduct the appropriate statistical analyses of the resultant data.

On the positive side, our faculty have benefited greatly by the use of a facilitator, and the faculty development strategies we have employed throughout the support period to help faculty develop assessments of CT/PS. We feel we have developed a process for faculty to develop their own assessments of CT/PS utilizing the definition described in Table 1 and the sample assessments we have put together in a packet. Identifying a process for creating assessments of cognitive skills that are difficult to assess will be of great help to the many new faculty that are replacing retiring faculty. The departmental leadership and infrastructure are maintaining collaboration among faculty to support curricular development and to monitor the curriculum using the assessment plan we have developed (Figure 2). We have disseminated many of our curricular revision approaches to other biology departments in our state by conducting workshops with grant support from the California State University Institute for Teaching and Learning, and other departments within our college have followed our lead in revamping their curricula and teaching strategies by using a professional facilitator.

FUTURE PLANS

There are components of the overall research plan to assess our curriculum depicted in Figure 2 that we will continue to refine and test, such as the CTLEA, the faculty surveys, and SALGains to assess the student learning environment and student achievements. However, we are considering replacing the ETS Biology Major Field Test

with one that is shorter and aligned to assess the learning outcomes of each Concentration. We have developed an assessment plan to continuously monitor our curriculum at university and at programmatic levels, but have not implemented the plan fully. There is awareness in the department of the need for an assessment committee, and the current chair has plans to make it a regular, standing committee within the department. Thus, the infrastructure to continue the work supported by the NSF ASA grant is being built by the departmental leadership and the faculty group that has been collaborating on this grant project.

Table 1. Working Definition of Critical Thinking and Problem Solving in Biology

Students successfully completing a major in Biological Science at CSUF will demonstrate that they can use a range of skills in the acquisition and interpretation of biological knowledge. They will demonstrate in their thinking, writing, and verbal communications the command of key biological terms and concepts. Additionally, they will demonstrate skills in critical thinking and problem solving in biology.

Elements of the CT/PS definition	Associated Verbs	Model Questions	Bloom's Level
1. Interpret data accurately (e.g. in graphs, tables, images and text).	Distinguish, Explain, Observe, Interpret, Illustrate, Express, Demonstrate, Describe, Catalog, Summarize	• Interpret the graph (table)… • Explain your answer… • Illustrate the relationship shown… • Describe what would happen if … • Catalog your observations… • What part does not fit?	2 Understand, Comprehend, Interpret
2. Apply appropriate numerical computational analysis to solve biological problems.	Apply, Choose, Solve, Use, Judge, Organize, Calculate, Determine, Select	• Calculate what would result… • Judge the effects… • Use the Nernst equation… • Solve for the following…. • Apply statistical analysis to… • Select the correct equation…	3 Apply
3. Distinguish between relevant and irrelevant – or – valid and invalid information.	Analyze, Distinguish, Identify, Categorize, Select, Evaluate, Judge, Justify, Compare	• Select the critical variables... • Identify the errors… • Justify your answer… • Analyze the data set… • Categorize the information…	4 Analyze 6 Evaluate
4. Represent biological data in an accurate, visual format.	Graph, Draw, Construct, Diagram, Illustrate, Create	• Graph the following data… • Construct a concept map… • Diagram a flow chart showing… • Create a schematic drawing of…	4 Analyze 6 Evaluate
5. Make appropriate inferences and deductions from biological information.	Analyze, Categorize, Compare, Differentiate, Distinguish, Identify, Select, Survey, Predict	• Predict the likely function of… • Justify your answer… • Identify the assumptions… • Distinguish between… • Analyze the problem…	4 Analyze

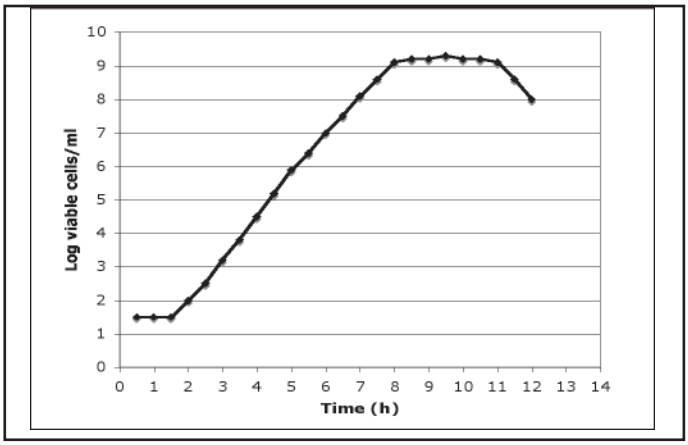
6. Make predictions from established hypotheses, theories or accepted models.	Apply, Explain, Organize, Select, Sketch, Solve, Use, Predict, Identify	• Predict what would happen if… • Explain how… • Sketch the relationship … • Identify the results of… • Select the most likely result…	3 Apply
7. Formulate a testable hypothesis with a coherent biological rationale.	Construct, Create, Formulate, Generate, Originate, Propose, Develop, Hypothesize	• Propose an alternative… • Explain how… • Generate a model to explain… • Formulate a hypothesis… • Construct an argument…	5 Synthesis
8. Devise tests for a hypothesis.	Construct, Design, Plan, Identify, Propose, Develop, Invent, Produce, Formulate	• How would you test… • Propose an alternative test… • Identify the variables… • Develop an experimental design • Plan a research protocol…	5 Synthesis
9. Evaluate evidence for and against different points of view or alternate hypotheses.	Judge, Evaluate, Critique, Defend, Compare, Assess, Identify	• Evaluate the alternate hypothesis… • Assess the societal or ethical issues… • Critique the study design… • Identify potential errors… • Discuss the merits… • Judge which is more valid…	6 Evaluate
10. Synthesize information and make connections to a principle, theme or concept.	Combine, Compose, Construct, Develop, Design, Formulate, Organize, Generate, Integrate, Review	• Describe the relationship of x to a principle or theme • Generate a model to… • Review the literature on… • Collect and analyze information related to… • Generate a concept map… • What are the implications of… • Integrate your observations of x with the concepts associated with y.	5 Synthesis

Verbs NOT typically associated with Critical Thinking and Problem Solving: Define, Label, List, Match, Name, Outline

Figure 1. Sample CTLEA (A) and rubric (B) for the first question utilized in the Introductory Cell Biology Course. Elements of the CT/PS definition that the question assesses are in brackets.

A. Sample CTLEA

Question 1: Interpret the data on the growth of a bacterial culture at 35°C as presented in the graph. Cells were inoculated into broth medium at time 0. Samples were removed at the various time points and plated on agar medium to measure viable (living) cell population. [Interpret data accurately]



Question 2: In the next experiment, equal amounts of cells were inoculated into ten culture tubes containing broth medium. The cells were then allowed to grow for 6 hours at various temperatures. After 6 hours, the number of viable cells was determined as described in question 1. Graph the results of this experiment. [Represent biological data in an accurate, visual format.]

Culture Tube #	Log_{10} viable cells/ml	Temperature (˚C)
1	0.3	5
2	0.5	10
3	1.2	15
4	1.5	20
5	2.1	25
6	2.9	30
7	3.1	35
8	1.5	40
9	0.06	45
10	0	50

Question 3: Predict what the shape of the growth curve would look like for a bacterial cell culture grown at 15°C. [Make appropriate inferences and deductions.]

B. Sample rubric for Question #1 of the CTLEA.
0 = A score of 0 indicates a blank answer (nothing is written).
1 = A score of 1 indicates that a student does not interpret any part of the question correctly. Nothing is correct but there is at least an effort made to answer the problem. A score of 1 also indicates an answer of strictly spurious information.
2 = A score of 2 indicates that a student interprets one of the phases of the graph correctly, but does not answer anything else that the problem asks for. A student may also have two out of four phases correct, but there are mistakes with units (time, log viable cells/ml) in their answer. The student may also have one of the phases interpreted correctly and a lot of other spurious information that does not have anything to do with the question.
3 = A score of 3 indicates that a student interprets about half of the graph correctly, or in other words explains two of the four phases correctly. A student may also have three out of four phases correct, but there are mistakes with units (time, log viable cells/ml) in their answer.
4 = A score of 4 indicates that a student has everything a 5 score would require in an answer, but only addresses three of the four phases. A student may also have everything correct that a score of 5 would require, but there are mistakes with units (time, log viable cells/ml) in the answer.
5 = A score of 5 indicates a correct and complete interpretation of the growth of the bacteria culture at 35°C as presented in the graph. A student must identify and explain the lag, log, stationary, and death phases of the bacterial growth curve. A student should also explain the time at which each phase is occurring and the number of log viable cells/ml that correspond with each phase.

Figure 2. Overall plan for researching the effects of learning environments on student achievements

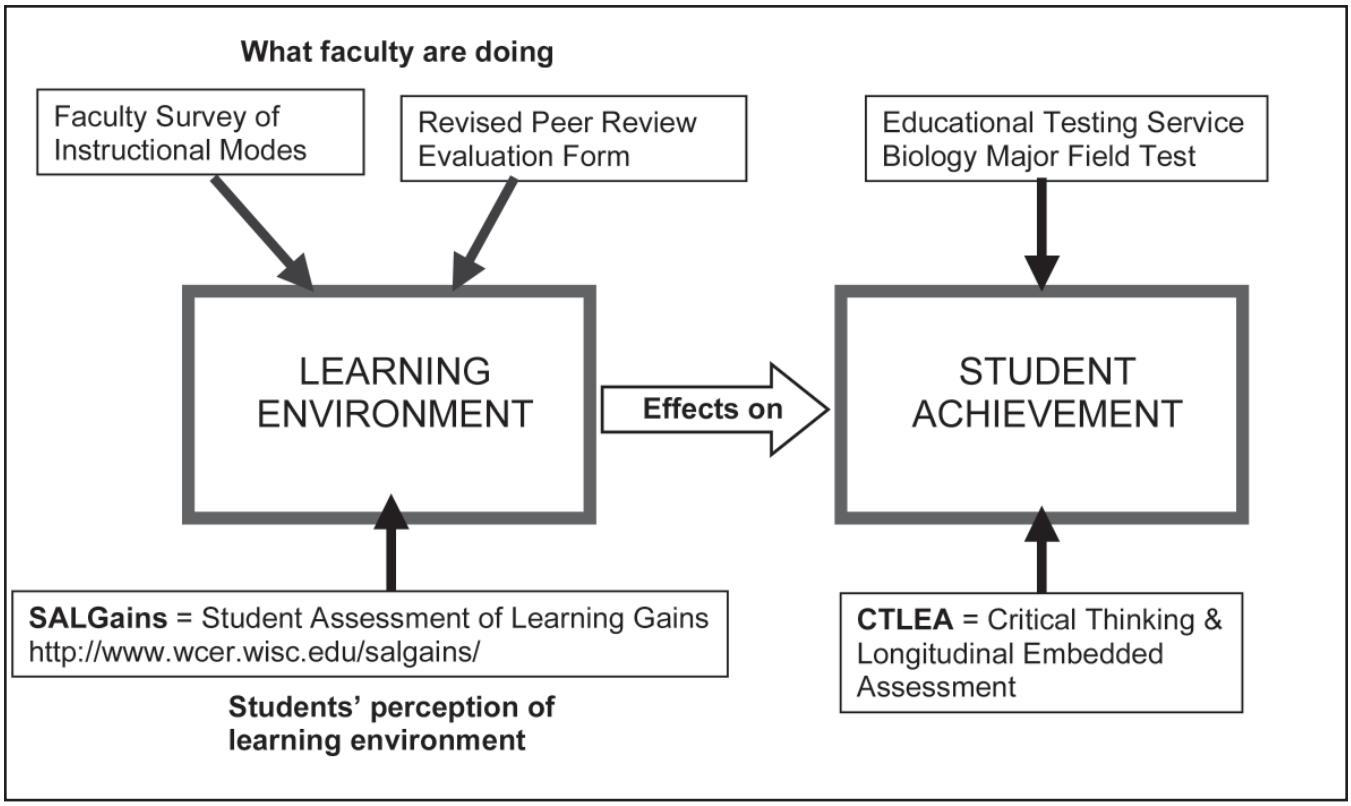
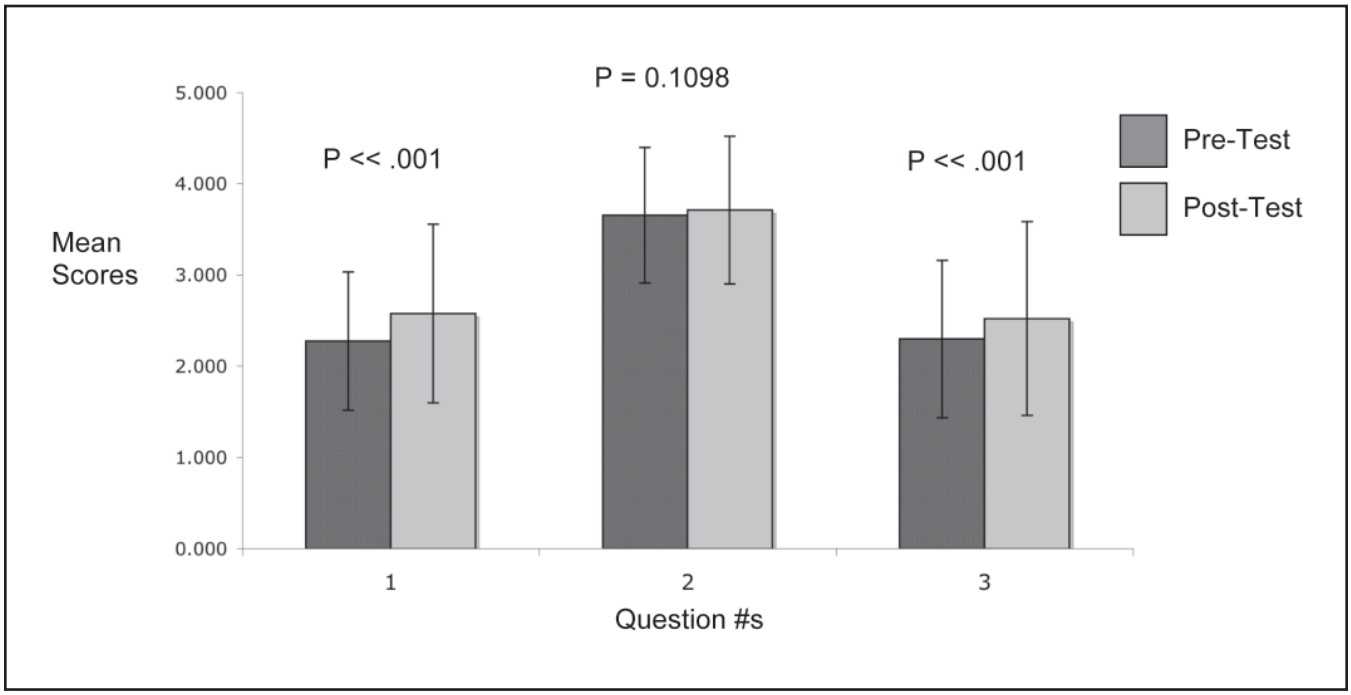


Figure 3. Preliminary data to test the CTLEA in Introductory Cellular Biology classes demonstrate that there is significant improvement in questions 1 and 3 of the assessment from the beginning of the semester to the end of the semester. Paired t-test, one-tailed P values, N= 453 students, in 6 classes. P = 1.982 x 10^{-9} for question #1, P = 4.302 x 10^{-5} for question #3. Difference between mean scores for pre-and post-tests for question 1 is 0.302, for question 2 is 0.057, and for question 3 is 0.225.



REFERENCES

[1] Ono, J.K., Dickson, K., Koch, R.A., Kandel J., Jones, C.E., Goodhue-McWilliams, K., and Banack, S., "Transformation of the Biology Major Core Sequence at Cal State Fullerton: Active Learning in Investigative, Problem-Solving Modules," CCLI DUE 9953108, 1999-2002. This grant established the collaborative processes utilized in the subsequent NSF ASA project.

[2] Wiggins, G., and McTighe, J., *Understanding by Design,* (Merrill Education/Prentice Hall, Inc., Upper Saddle River, N.J.), 1998.

[3] Bloom, B. S., (ed.) *Taxonomy of educational objectives: The classification of educational goals by a committee of college and university examiners*, (Longmans, New York), 1956; Anderson., L W., and Krathwohl, D. R., (eds.), *A Taxonomy for Learning, Teaching, and Assessing*. (Pearson Allyn & Bacon Publishers), 2000.

[4] Facione, P. A., and Facione, N. C., *California Critical Thinking Disposition Inventory* [CCTDI], (The California Academic Press, Millbrae, Calif.), 1992.

[5] Affinity Process definition and description: http://atwiki.assistivetech.net/index.php?title=Affinity_Diagram&printable=yes

[6] Facione, P.A., *The Delphi Report, Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction*, (The California Academic Press, Millbrae, Calif.), Executive Summary, 1990.

[7] Casem, M., "Student perspectives on curricular change: lessons from an undergraduate lower-division biology core," *Cell Biology Education*, vol. 5, pp. 65-75, 2006.

[8] Self Assessment of Learning (SAL) Gains, a free online survey to monitor student perceptions of class activities that helped them learn. SALGains provides a statistical report of the results and can be accessed at: http://www.wcer.wisc.edu/salgains/instructor

[9] Facione, P.A., "The California Critical Thinking Skills Test: College Level Technical Report #1: Experimental Validation and Content Validity," *ERIC Document ED* 327-549, (The California Academic Press, Millbrae, Calif.), 1990. http://www.insightassessment.com/pdf_files/TECH1.pdf

[10] "Looking at Student Work" is a formal procedure developed by faculty in K-12 using a protocol to examine samples of student work by teachers. These procedures may be viewed at: http://www.lasw.org/

Tracing Matter and Energy Through Biological and Geological Systems: Fundamental Problems of Students' Understanding

Christopher Wilson*, Charles Anderson, Merle Heidemann, Tammy Long, John Merrill, Brett Merritt, Gail Richmond, Duncan Sibley, Mark Urban-Lurain, and Joyce Parker - *Michigan State University*

INTRODUCTION

We have asked many students in a wide variety of undergraduate courses about human weight loss, often using the context of Jared, the Subway® Guy, who lost a lot of weight eating Subway sandwiches. However, regardless of the context, the course, or the wording of the question, there are a significant number of students whose explanations of the weight loss resemble the quotes in Figure 1.

Figure 1. Jared, the Subway® Guy

Where did the mass go?
"The fat was converted into useable energy and burned by muscle contraction for movement"
"The fat was also deposited out of his system through feces and excretion of sweat"

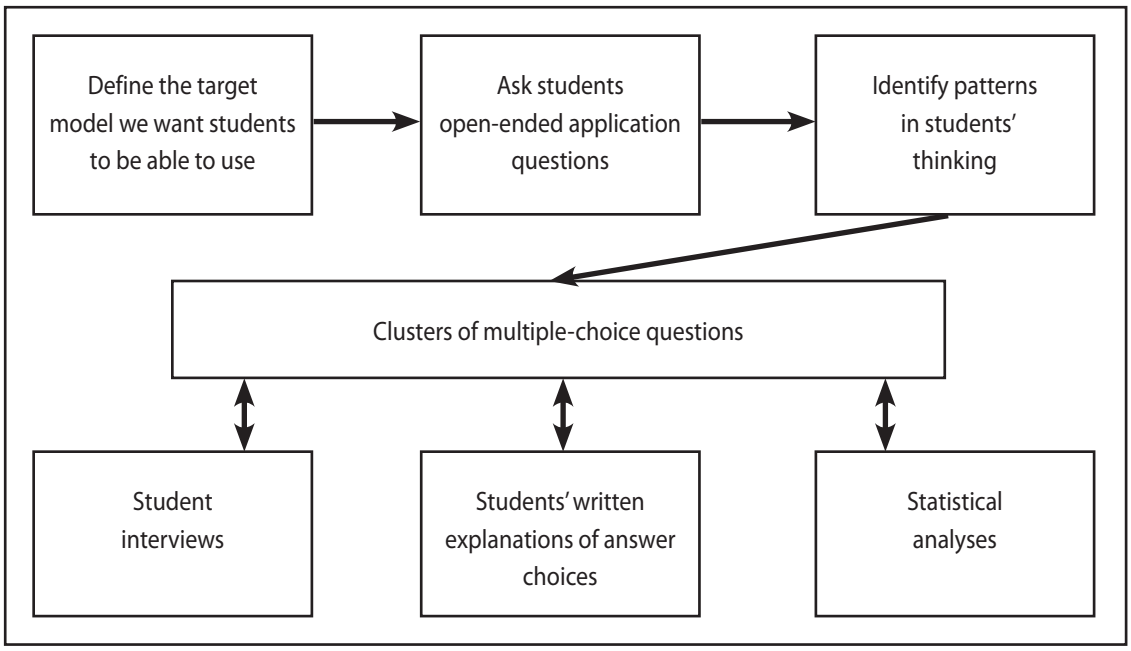
Students' incorrect answers to this question do not fit into a single category, as they might if they were based on a commonly held misconception. However, we would like to find patterns in students' problematic thinking and understand these well enough to guide effective instructional changes.

The goal of this project is to develop *Diagnostic Question Clusters* in biology and geology. We use the word "diagnostic" in the sense that we want the question clusters to identify patterns in student reasoning that underlie misconceptions or unscientific arguments. We define a cluster as a domain-specific set of questions designed to be used together. The clusters are distinct from Concept Inventories in that they are not focused on misconceptions *per se*, but rather on patterns in students' reasoning that underlie common misconceptions OR unscientific arguments.

METHODS

Figure 2 shows the method we use to develop diagnostic question clusters. We determine the model of the biological or geological system that we would like students to be able to apply. We ask students short-answer application questions that require the use of the designated model, and we look for patterns in their responses to these questions. Based on the patterns, we develop multiple choice items where the distractors are intended to represent common problematic ways of thinking. We test the robustness of our explanation of why students choose particular distractors by interviewing students, having them write about their answer choices, and by doing statistical analyses.

Figure 2. Method for developing diagnostic question clusters



RESULTS: COGNITION MODEL

We present our results in the format of the National Research Council's model for effective assessment [3]. In the report, *Knowing What Students Know: The Science and Design of Educational Assessment*, the National Research Council lays out three essential features of assessment: 1) a cognition model describing how we would like students to represent knowledge and develop competence in the domain; 2) observations of how students think revealed by assessment tasks; and 3) interpretations of students' thinking. We begin here with our cognition model, which has two components: a content framework that organizes and delimits the content as we want students to understand it, and a set of practices that students engage in when reasoning with the content framework.

We work both in biology and geology, so in order to have a general framework that works across these disciplines, we take a systems approach to the content. With this approach, we frame the content as systems with matter inputs, matter outputs, and energy transformations. Our content framework for cellular respiration is shown in Figure 3, and is organized around three levels or scales that represent a series of nested systems. We want students to be able to understand the process of cellular respiration as a series of enzymatic reactions that occur at the subcellular level; as an overall reaction that a cell undergoes to convert stored energy in the form of sugars to a form (ATP) that can be used to drive other reactions; and, finally, as part of a complex system whereby aerobic organisms use food to fuel processes. Each level is a different context for understanding cellular respiration involving different structures, ranging from molecules to organisms. The reader should note that this is an abbreviated version of the framework that focuses largely on carbon, so not all matter that the students need to trace (e.g. ADP) is represented.

Figure 3. Content Framework for the cellular respiration model.

	Tracing Matter	Tracing Energy	Context / Location
ORGANISMAL LEVEL			
Using Food for Energy	Food provides molecules that serve as fuel and building material for all organisms. Some of the matter in food leaves aerobic organisms in the form of carbon dioxide and water.	Organisms transform the chemical energy in food into usable chemical energy in energy management molecules.	All cells in all living organisms use food for fuel and building materials.
CELLULAR LEVEL			
Cellular Respiration	$C_6H_{12}O_6 + 6H_2O + 6O_2$ ➔ $6CO_2 + 12H_2O$	Some chemical energy in the C-C and C-H bonds in glucose ➔ chemical energy in ATP	Occurs in the cell cytoplasm, mitochondrial matrix, and mitochondrial membrane.
SUB-CELLULAR LEVEL			
Glycolysis	(1)6-carbon glucose ➔ (2)3-carbon pyruvate + (2)H_2O	Some chemical energy in the C-C and C-H bonds in glucose ➔ chemical energy in (2)ATP & (2)NADH	Occurs in the cytoplasm in all living cells.
Pyruvate oxidation (Acetyl CoA production)	(2)3-carbon pyruvate ➔ (2)2-carbon acetyl CoA + (2)CO_2	Some chemical energy in the C-C and C-H bonds in pyruvate ➔ chemical energy in (2)NADH	Occurs in the mitochondrial matrix in eukaryotes, in the cytoplasm in prokaryotes.
Kreb's Cycle	(2)2-carbon acetyl CoA ➔ (2)$2CO_2$	Some chemical energy in the C-C and C-H bonds in acetyl CoA ➔ chemical energy in (2)ATP, (2)3NADH & (2)$FADH_2$	Occurs in the mitochondrial matrix in eukaryotes, in the cytoplasm in prokaryotes.
Electron Transport Chain & Oxidative Phosphorylation	Electrons from NADH & $FADH_2$ +O2 ➔ H2O (No carbon skeleton)	Some chemical energy in (10)NADH & (2)$FADH_2$ ➔ chemical energy in (~34)ATP	Occurs in the inner mitochondrial membrane in eukaryotes, in the plasma membrane in prokaryotes.

When applying this content framework to different situations, there are four practices that students need to engage in: tracing matter; tracing energy; organizing systems; and segregating matter, energy, and systems. By ***tracing matter,*** we mean identifying the matter that is changing and the chemical identity of the molecules involved, describing the nature of the change, and in some cases, measuring the amounts of matter involved. The knowledge needed for this practice is represented by the first filled-in column in the content framework. The knowledge needed for the practice of ***tracing energy*** is represented in the next column. By tracing energy we mean identifying the forms of energy, how they are transformed and transferred, and in some cases, quantifying the amount of energy involved.

The third practice is ***organizing systems***. When students are asked an application question, they need to identify the level in which the problem is set, as well as in which level to look for explanations. For example, the weight loss question shown in Figure 1 is set at the organismal level, Jared, but the explanation of why he lost weight comes mainly from looking at what goes on at the cellular level: cellular respiration. In other words, we want students to make connections between the different levels moving vertically through the content framework. We also want students to move horizontally through the content framework by recognizing that the matter inputs and outputs of a process give us clues about energy, in that, for example, a reaction that oxidizes a carbon compound is likely to be exergonic (releasing energy).

While following matter can give us clues about energy transformations, and vice versa, students should know that tracing matter and tracing energy are not one and the same; we want them to ***segregate matter and energy***. For example, the student who responded to the question about Jared by saying that "the fat was converted into useable energy and burned" has not segregated changes in matter from changes in energy. Likewise, students who say that the fat has been transferred to ATP are mixing the idea that the chemical potential energy originally in the fat has been transferred to ATP with the practice of tracing matter. They are assuming the matter follows the energy transformations.

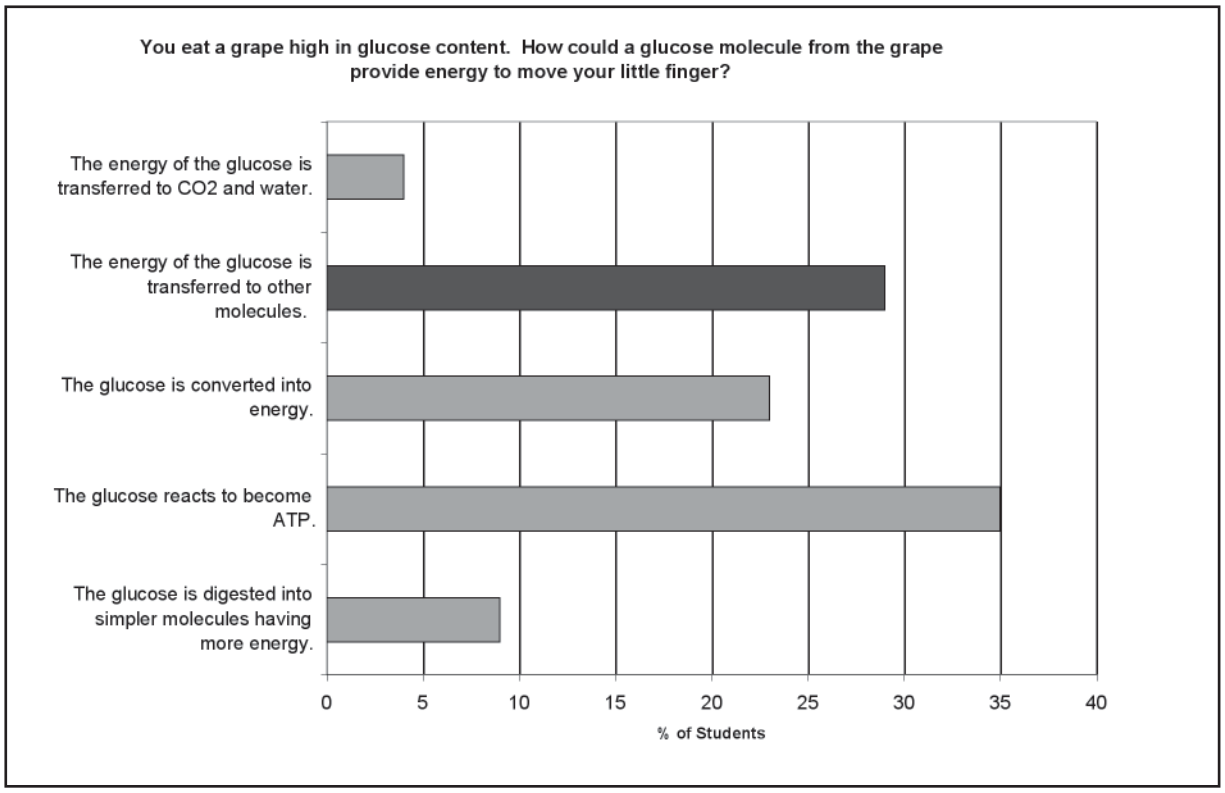
RESULTS: OBSERVATIONS

In the framework of the NRC Assessment Triangle, we monitor students' progress toward the cognition model with observations of students' responses to assessment items, in our case a *Diagnostic Question Cluster*. The questions in the cellular respiration cluster have gone through an iterative development process, in that we developed distractors based on patterns in students essay responses to the question stem. We are in the process of refining distractors based on their performance (frequency of selection, bias, confusing language, multiple interpretations, etc.) in large classes, interviews with students, and students' written explanations of why they picked particular choices and not others. Students' responses to three questions are shown in the next section. These come from an hourly exam in an introductory course in cellular biology with an enrollment of more than 400.

RESULTS: INTERPRETATION

The final component of the NRC Assessment Triangle is *Interpretation* – a way of making sense of the observations of student thinking. A meaningful interpretation requires that the assessment items and distractors be aligned with the cognition model including both the content framework and the practices. This alignment is discussed here with respect to Figure 4. One question from the cellular respiration cluster is shown along with a histogram of the proportion of students choosing each distractor. Each distractor represents an incorrect response that can be explained by students' failure to develop one of the four desired practices.

Figure 4. Example Assessment Item – Eating a grape



The second choice (The energy of the glucose is transferred to other molecules) is the correct answer. In order to reason their way to this answer, students would have to realize that while the question is posed at the organismal level – the movement of someone's little finger, the answer lies in understanding what is going on at the cellular level. A student must also recognize that the answer involves tracing energy through the process of cellular respiration. Thus, a student needs to understand the content in the first cells of the content framework and be able to trace chemical potential energy that is stored originally in the glucose and ends up in ATP.

Students who chose the fifth answer choice (The glucose is digested into simpler molecules having more energy) may not see the connection between the question posed at the organismal level and cellular respiration. Instead they look for an answer at the same level as the question, that is at the organismal level – digestion. These students need help developing their ability to *organize levels or systems*.

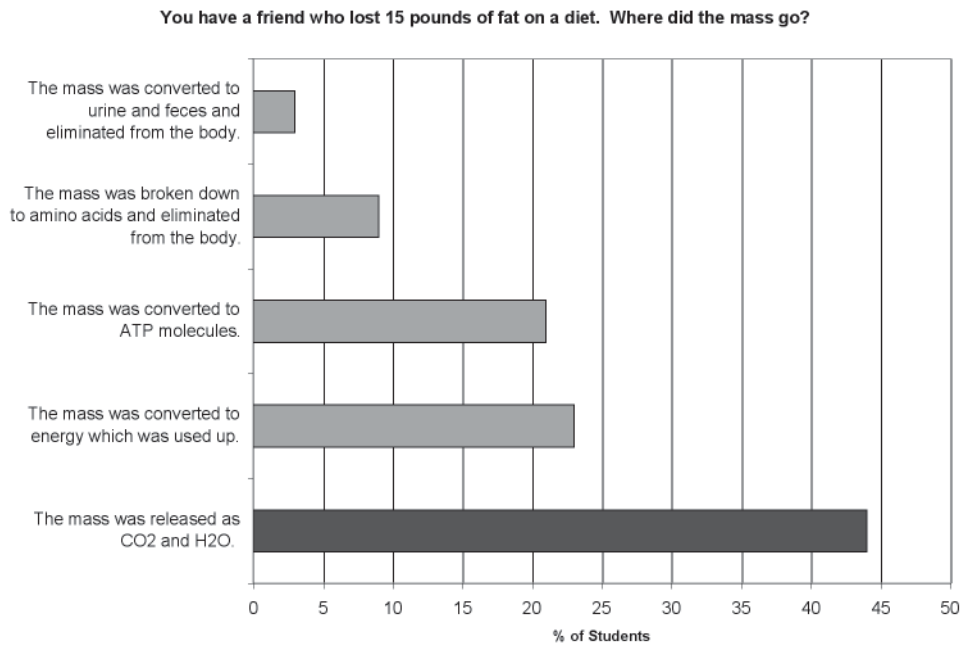
The fourth answer choice (The glucose reacts to become ATP) is based directly on student responses in interviews (see Figure 1) and essay questions. As described above, students who chose this response may be confusing the energy transfer from glucose to ATP with the changes in matter (chemical reactions) that result in glucose being broken down into carbon dioxide and water. Students choosing the first answer choice (The energy of the glucose is transferred to CO_2 and water) may be having similar problems, since they appear to be equating the matter conversion of glucose going to carbon dioxide and water with the energy conversions. These students need help *segregating changes in energy and matter*. Figure 5 shows two additional questions from the respiration cluster, along with histograms of students' responses. A *friend who lost weight* is a more formal version of the Jared problem. The correct answer is answer choice 5. Answer choices 3 and 4 (the mass was converted to energy and used up, and the mass was converted to ATP molecules) are plausible answers for students who do not segregate matter and energy conversions. An alternative interpretation is that students who choose the mass to ATP

conversion are not able or interested in tracing matter, since they have made a choice that does not account for the mass loss, and also has phosphorous and nitrogen atoms appearing without explanation. Students choosing this response may also be engaging in what we call procedural display, associating ATP in vague ways with cellular respiration. Answer choices 1 and 2 (the mass is converted into amino acids or urine and feces and eliminated form the body) are more often chosen when this item is used pre-instruction. These choices represent "street knowledge" about weight loss and do not require thinking about the problem at a different level from the one in which it was posed.

The distractors in *What happens to food*? are very similar to those of the other two example items. The correct choice is 1. Answer choice 4 (the food is digested into smaller sugars) is meant to appeal to students who are looking for an answer at the organismal level. However, there is some danger that the students who are not reading carefully simply see that it talks about the breakdown of glucose and choose it. The same is true of distractor 1 in the first example item. We are studying this by asking students to explain their answer choices. Answer choices 2 and 3 (atoms of the food are rearranged to form ATP and the food is turned into energy) both indicate failure to segregate matter and energy conversions. Another way to look at these responses is that students who choose them do not feel obligated to conserve matter. Answer choice 2 involves the unexplained appearance of phosphorous and nitrogen atoms, and answer choice 3 implies that the atoms of the food disappear.

Figure 5. Two more questions from the cellular respiration cluster

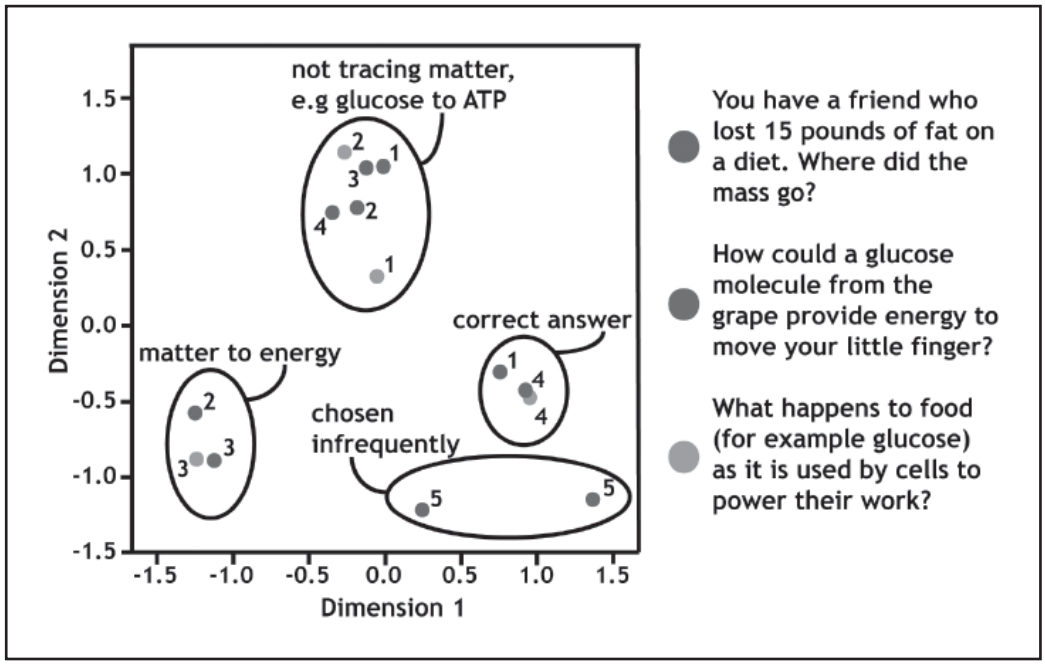
Sample Assessment Item – A friend who lost weight



We are currently testing the validity of these interpretations by analyzing student interviews, analyzing students' written explanations of their answer choices, and by doing statistical analyses of students' responses. Preliminary results from the statistical analyses are shown below.

One question we have is whether students who have not developed the desired four practices reason consistently across these questions. In order to look for patterns in students' responses, we did a multiple correspondence analysis [1,2]. This is similar to factor analysis but is appropriate for categorical data. Figure 6 shows the results of the multiple correspondence analysis assuming two independent variables. Each possible choice is indicated by a colored dot where the color identifies the question and the number indicates the specific distractor. Four clusters are apparent. The correct responses form a fairly tight cluster in the lower right portion of the field. This is not an unexpected result since students who understand one question are likely to understand the other two. It is more interesting to see that the choices that state that the matter (mass, glucose, food) was converted to energy also cluster (see the lower left part of the field), suggesting that students have consistent, incorrect ways of reasoning about respiration. A third cluster is evident in the upper middle of the field. These choices represent answers involving erroneous chemical reactions such as the formation of ATP from the atoms of glucose. Again, this may indicate that students have consistent but incorrect ways of reasoning about the material. A fourth loose cluster exists in the lower right corner of the field. The two responses in this cluster were infrequently chosen and are fairly widely separated, therefore we do not attach much significance to this group.

Figure 6. Results of multiple correspondence analysis (assuming two dimensions) of students' responses to the three assessment items.



The next steps in the statistical analyses are to try to understand the two latent variables represented by the axes and to see if this type of correlation occurs across less closely related questions. We are currently pursuing these lines of work. In particular, we are interested in understanding the horizontal axis, which might be a measure of students' ability or predilection to conserve matter.

CONCLUSIONS AND IMPLICATIONS

The focus of our work is a careful analysis of students' reasoning. Our content framework allows us to analyze informal student practices and compare them with the scientific practices that are the goals of our courses. The analyses of practices, the content framework, and assessment data suggest changes in instruction and provide

us with tools to evaluate their effects. The research is an iterative process in that no one component is developed first. Our initial frameworks suggest practices and content to focus on in the initial assessment; data from those assessments suggest revisions in the framework and changes in instructional models, which in turn provide new ideas about assessment, and so forth.

The work presented here is in the domain of cellular respiration. We are finding similar results in the closely related topic of photosynthesis. However, other work in our group indicates that this approach also has merit in other areas of biology, as well as in geoscience. For example, we can build a reasonable content framework for the water cycle focusing on the same practices presented here (tracing matter and energy). A framework in sedimentology will probably include the practice of keeping track of spatial and temporal dimensions. In biology, we assume that we will need to add tracing information (understanding signals and cues for events, following genetic material, etc) as a desired practice in order to organize all of the material usually taught.

We see two important advantages to this approach. First, this structure (content framework and associated practices), provides a framework for presenting a large amount of content to students using a few themes. This allows students to focus on a few themes for the duration of the course rather than a myriad of "new" topics every week. In other words, we hope that students will develop a general approach to biology or geoscience that they can enlarge as they progress through their studies. Second, this structure reveals patterns in what can sometimes appear to be random thinking on the part of students. Knowledge of the patterns in students' problematic thinking is the information needed by instructors to guide effective changes in teaching.

REFERENCES

[1] Greenacre, M. J., *Theory and application of correspondence analysis*, (Academic, Orlando, Fla.), 1984.

[2] Greenacre, M. J., and Blasius, J., *Correspondence analysis in the social sciences : recent developments and applications,* (Academic Press, London), 1994.

[3] National Research Council, *Knowing What Students Know: The Science and Design of Educational Assessment,* 2001. http://books.nap.edu/books/0309072727/html

ChemQuery: An Assessment System for Mapping Student Progress in Learning Chemistry

Kathleen Scalise* - *University of Oregon*
Jennifer Claesgens, Mark Wilson, Angelica Stacy - *University of California, Berkeley*

ABSTRACT

The recent U.S. report of the Spellings Commission on the Future of Higher Education called for measuring "meaningful student learning outcomes" in post-secondary education, including over time and between courses. This paper presents an assessment system, ChemQuery, for measuring growth over time in general chemistry understanding, which can be used within and between courses, both at the university level and for articulation between high school and college courses. The approach uses investigations of student answer patterns to build a framework of learning goals over time, called the *Perspectives of Chemists*, and calibrates the framework with item response modeling techniques. Pilot study data are presented and we report on the quality of the assessment evidence gathered through the approach. ChemQuery includes assessment questions, a scoring rubric, item exemplars, and a framework to describe the path of student understanding that emerges. Integral to this kind of criterion-referenced measurement is a focus on what is being measured: the intention of the assessment, its purpose, and the context in which it is going to be used. The construct allows us to intentionally acknowledge that what we teach, despite our best efforts, is not necessarily what students learn, and that the structure of the content that we teach does not necessarily determine how they learn. Moreover, the construct allows us to begin to narrate the development of understanding that occurs as students "learn" over the course of instruction. It can be used to guide curricular and pedagogical approaches, and to compare how students learn across contexts.

INTRODUCTION

When the final report of the Spellings Commission on the Future of Higher Education was released in September 2006, the call for higher education to focus on measuring and reporting "meaningful student learning outcomes" was a major charge [1]. The report's suggestion that some standardized tests might be used to collect assessment data was met with a call by the Association of American Colleges and Universities and others for better curriculum-embedded methods of assessment [2,3]. Ways to interpret and understand the building of understanding across instruction and over a student's course of learning have often been lacking in post-secondary education, where the focus of student assessment is usually within individual courses, and less between courses and over time. Types of growth modeling that allow a larger picture of learning outcomes to emerge bring their own challenges, including considerations of what should be measured and how it should be valued; technical concerns of how to best go about measuring growth; and issues of effectively collecting, sharing, and reporting data. In this paper, we describe one example of growth models in higher education outcomes assessment — ChemQuery, developed with NSF funding — that illustrates an approach useful for measurement within and between courses, and across subject matter areas.

ChemQuery is an assessment system that uses a framework of the key ideas in the discipline, called the *Perspectives of Chemists*, and criterion-referenced analysis using item response theory (IRT) to map student progress in general chemistry. It currently includes the *Perspectives* framework, a scale of student progression in understanding, an item bank of open-ended constructed response items, some computer-adaptive selected response items [4,5], a scoring rubric, and item exemplars. Analysis using item response models includes

generation of Wright maps, item fits, student progress reports, and reliability and validity evidence.
Integral to such criterion-reference measurement is a focus on what is being measured, which is referred to as the construct. The construct is the intention of the assessment, its purpose, and the context in which it is going to be used. This approach to research on student understanding allows us to use quantitative measures to test and refine hypotheses about how learning develops, and to relate that to qualitative evidence about student learning.

The *Perspectives* construct is built on the theoretical idea that the field of chemistry can be largely grouped into three core conceptions, or scientific models: matter, change, and energy. The purpose in framing the "big ideas" of chemistry is to organize the overarching ideas of the discipline while simultaneously constructing an instrument for measuring the values of these variables for individual students. In addition to this organization of topics, the ChemQuery construct further suggests that learning in each of these three areas, or "strands," is a rich and varied progression, from forms of naive to more complete and consistent understanding of explanatory models of chemistry, and that this progression in understanding represents a set of necessary and distinct areas of understanding enabling higher-order thinking in chemistry [6].

It may be that the identification of three, or some small discrete number, of "big idea" *Perspectives* could be a useful point of comparison for this type of assessment across courses in other STEM disciplines. We have been in discussions with physicists, biologists, and mathematicians on what such *Perspectives* might be in their fields, and are helping other STEM disciplines to explore some possible approaches. As an aside, it is interesting to note that while aspects of the Matter and Energy *Perspectives* have been identified by physicists as potentially crossing over to their fields fairly directly, the model of Change in our framework is perceived as very much the perspective of chemical and physical change generally described in chemistry objectives and standards. Change in physics might need to encompass both considerably more macroscopic change (for instance, positional change of large objects), and sub-atomic change for particles below the level of electrons, neutrons, and protons.

Developing the ChemQuery progress variables has required studying misconception research and theories of learning such as alternate conceptions and blending of models [7,8,9,10,11,12,13,14,15,16,17,18, 19]. This helps describe how students build understanding in chemistry. Implicit to the ChemQuery approach is the assumption that emphasizing deep understanding supports the learning of algorithms and chemical definitions [20,21], which is consistent with the Constructivist perspective that the learner builds his or her understanding. Thus, the purpose of the *Perspectives* construct is to describe a hierarchy of chemistry content that then *defines variables* to allow us to measure learning outcomes, *determines scales* for these variables, and *constructs instruments* for measuring the values of these variables for individual students.

The scores for a set of student responses and the questions are put on the same scale with item response models, also yielding fit, validity, and reliability evidence, matched to the *Perspectives* construct. Ultimately, the multidimensional construct allows mapping of individual student performance to reveal a picture of learning in the domain over time.

CHEMQUERY VARIABLES—MATTER, CHANGE, AND ENERGY

A fuller explanation of one *Perspectives* variable, Matter, is shown in Figure 2, and similar guidelines have been developed and tested for Change and Energy. While there are certainly other ways to divide the discipline of chemistry into overarching ideas, the usefulness of the approach is only realized when the details

for a chosen set are worked out and supported by empirical evidence. This set of *Perspectives* was chosen, an instrument constructed (a set of questions to assess student understanding), student responses gathered, and the components of the measurement system refined over several iterations, as evidence to examine and revise the model was collected.

Figure 1. ChemQuery Assessment System: *Perspectives of Chemists* on Matter

Level of Success	Big Ideas	Descriptions of Level	Item Exemplars
Generation13-15	Bonding models are used as a foundation for the generation of new knowledge (e.g., about living systems, the environment, and materials).	Students are becoming experts as they gain proficiency in generating new understanding of complex systems through the development of new instruments and new experiments.	a) Composition: What is the composition of complex systems? (e.g., cells, composites, computer microchips) b) Structure: What gives rise to the structure of complex systems? (e.g., skin, bones, plastics, fabrics, paints, food,) c) Properties: What is the nature of the interactions in complex systems that accounts for their properties? (e.g., between drug molecules and receptor sites, in ecosystems, between device components) d) Quantities: How can we determine the composition of complex systems? (e.g., biomolecules, nanocomposites)
Construction 10-12	The composition, structure, and properties of matter are explained by varying strengths of interactions between particles (electrons, nuclei, atoms, ions, molecules) and by the motions of these particles.	Students are able to reason using normative models of chemistry, and use these models to explain and analyze the phase, composition, and properties of matter. They are using accurate and appropriate chemistry models in their explanations, and understand the assumptions used to construct the models.	a) Composition: How can we account for composition? b) Structure: How can we account for 3-D structure? (e.g., crystal structure, formation of drops,) c) Properties: How can we account for variations in the properties of matter? (e.g., boiling point, viscosity, solubility, hardness, pH, etc.) d) Amount: What assumptions do we make when we measure the amount of matter? (e.g., non-ideal gas law, average mass)

Level of Success	Big Ideas	Descriptions of Level	Item Exemplars
Formulation **7-9**	The composition, structure, and properties of matter are related to how electrons are distributed among atoms.	Students are developing a more coherent understanding that matter is made of particles and the arrangements of these particles relate to the properties of matter. Their definitions are accurate, but understanding is not fully developed so that student reasoning is limited to causal instead of explanatory mechanisms. In their interpretations of new situations, students may over-generalize as they try to relate multiple ideas and construct formulas.	a) Composition: Why is the periodic table a roadmap for chemists (Why is it a "periodic" table)? How can we think about the arrangements of electrons in atoms (e.g., shells, orbitals)? How do the numbers of valence electrons relate to composition (e.g., transfer/share)? b) Structure: How can simple ideas about connections between atoms (bonds) and motions of atoms be used to explain the 3-D structure of matter (e.g., diamond is rigid, water flows, air is invisible)? c) Properties: How can matter be classified according to the types of bonds (e.g., ionic solids dissolve in water, covalent solids are hard, molecules tend to exist as liquids and gases)? d) Amount: How can one quantity of matter be related to another (e.g., mass/mole/number, ideal gas law, Beer's law)?
Recognition **4-6**	Matter is categorized and described by various types of subatomic particles, atoms, and molecules.	Students begin to explore the language and specific symbols used by chemists to describe matter. They relate numbers of electrons, protons, and neutrons to elements and mass, and the arrangements and motions of atoms to composition and phase. The ways of thinking about and classifying matter are limited to relating one idea to another at a simplistic level of understanding.	a) Composition: How is the periodic table used to understand atoms and elements? How can elements, compounds, and mixtures be classified by the letters and symbols used by chemists (e.g., $CuCl_2$ (s) is a blue solid, $CuCl_2$(aq) is a clear, blue solution)? b) Structure: How do the arrangements and motions of atoms differ in solids, liquids, and gases? c) Properties: How can the periodic table be used to predict properties? d) Amount: How do chemists keep track of quantities of particles? (e.g., number, mass, volume, pressure, mole)
Notions **1-3**	Matter has mass and takes up space.	Students articulate their ideas about matter, and use prior experiences, observations, logical reasoning, and knowledge to provide evidence for their ideas.	a) Composition: How is matter distinct from energy, thoughts, and feelings? b) Structure: How do solids, liquids, and gases differ from one another? c) Properties: How can you use properties to classify matter? d) Amount: How can you measure the amount of matter?

Perspectives of Chemists Matter Variable

Within each of the *Perspectives*, a scale to describe student understanding was proposed, as shown in the example in Figure 2. The levels within the proposed variables are constructed such that students give more complex and sophisticated responses as they develop, from describing their initial ideas in Level 1 (Notions), to relating the language of chemists to their view of the world in Level 2 (Recognition), to formulating connections between several ideas in Level 3 (Formulation), to fully developing models in Level 4 (Construction), to asking and researching new scientific questions in Level 5 (Generation).

As a specific example, consider the Matter variable. A sample question, the scoring guide, and examples of student work in Levels 1 and 2 are shown in Figure 2. In Level 1 (Notions), students can articulate their ideas about matter, and use prior experiences, observations, logical reasoning, and knowledge to provide evidence for their ideas. The focus is largely on macroscopic (not particulate) descriptions of matter, since students at this level rarely have particulate models to share. In Level 2 (Recognition), students begin to explore the language and specific symbols used by chemists to describe matter. The ways of thinking about and classifying matter are limited to relating one idea to another at a simplistic level of understanding, and include both particulate and macroscopic ideas. In Level 3 (Formulation), students are developing a more coherent understanding that matter is made of particles and the arrangements of these particles relate to the properties of matter. Their definitions are accurate, but understanding is not fully developed, so that student reasoning often is limited to causal instead of explanatory mechanisms. In their interpretations of new situations, students may over-generalize as they try to relate multiple ideas and construct formulas. Since the responses in the sample question were obtained from students near the beginning of high school chemistry, even the best students had not reached Level 4 (Construction), where they are expected to relate the arrangements of atoms with properties, for instance. However, we are currently collecting evidence from more advanced university-level students in these levels. In Level 4 (Construction), students are able to reason using normative models of chemistry, and use these models to explain and analyze the phase, composition, and properties of matter. They are using accurate and appropriate chemistry models in their explanations, and understand the assumptions used to construct the models. In Level 5 (Generation), students are becoming experts as they gain proficiency in generating new understanding of complex systems through the development of new instruments and new experiments.

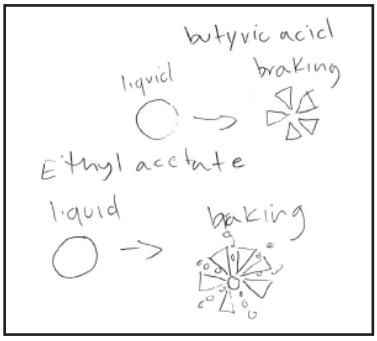
SAMPLE PROBLEM AND SCORING GUIDE

"You are given two liquids. One of the solutions is butyric acid, with a molecular formula of $C_4H_8O_2$. The other solution is ethyl acetate, with the molecular formula $C_4H_8O_2$. Both of the solutions have the same molecular formulas, but butyric acid smells bad and putrid while ethyl acetate smells good and sweet. Explain why you think these two solutions smell different."

Figure 2. Example item and scoring guide.

	0	Response: "I have absolutely no idea." Analysis: Response contains no information relevant to item.

Notions	1	Response: "Just because. That doesn't seem possible. How can they be different when they have the same molecular formula?" Analysis: Student makes one macroscopic observation by noting that the molecular formulas in the problem setup are the same.

	1	Response: "Using chemistry theories, I don't have the faintest idea, but using common knowledge I will say that the producers of the ethyl products add smell to them so that you can tell them apart." Response: "Just because they have the same molecular formula doesn't mean they are the same substance. Like different races of people: black people, white people. Maybe made of the same stuff but look different." Analysis: These students use ideas about phenomena they are familiar with from their experience, combined with logic/comparative skills to generate a reasonable answer, but do not employ molecular chemistry concepts.
	+1	Response: "Maybe the structure is the same but when it breaks into different little pieces and changes from liquid into gas they have a different structure in the center and have a different reaction with the air. (Shows drawing:) Analysis: This answer acknowledges that chemical principles or concepts can be used to explain phenomena. Attempts are made to employ chemical concepts based on a "perceived" but incorrect understanding of chemistry in the first example and solely employing chemical jargon in the second example. 

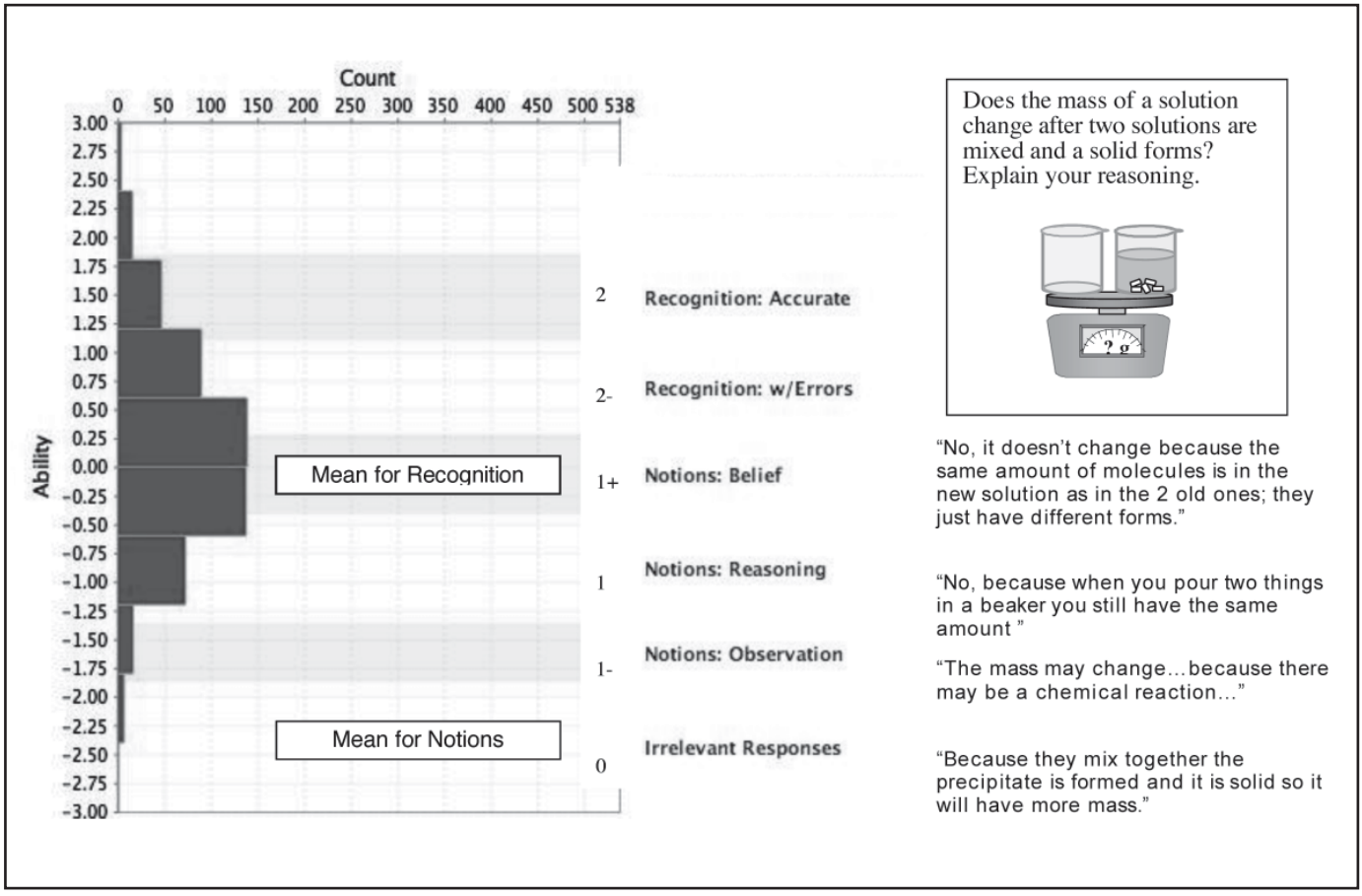
Recognition	2-	Response: "I think these two solutions smell different is because one chemical is an acid and most acids smell bad and putrid while the ethyl acetate smells good and sweet because its solution name ends with "ate" and that usually has a good sweet smell." Analysis: This response correctly cites evidence for the difference in smells between the two chemicals, appropriately using smell combinatorial patterns taught in class and chemical naming conventions, but does not explain the root cause as the difference in molecular structure between the two chemicals.
	2	Response: "They smell different b/c even though they have the same molecular formula, they have different structural formulas with different arrangements and patterns." Analysis: This response appropriately cites the principle that molecules with the same formula can have different structures, or arrangements of atoms within the structure described by the formula. However it shows an incomplete attempt to use such principles to describe the simple molecules given in the problem setup.
	+2	Response: (Begins with problem setup below, showing molecular formula of labeled butyric acid and same formula labeled ethyl acetate.) $C_9H_8O_2$ - butyric acid $C_4H_8O_2$ - ethyl acetate "The two molecules smell differently because the have different molecular structures. The butyric acid contains a carboxylic acid structure (which smells bad) and the ethyl acetate contains an ester (which smells good). We can tell which molecule will smell bad and which will smell good by studying the molecular structure and by looking at the names. Any 'ACID' ending name will smell bad and any '-ATE' ending name will smell good." Analysis: Response cites and appropriately uses the principle that molecules with the same formula can have different structures. Student correctly cites rule learned in class pertaining to smell patterns in relation to functional groups identified by chemical name, and uses this information to begin to explore simple molecules. However, student stops short of a Level Three response, which could be made by examining structure-property relationships through, for instance, presenting possible structural formulas for the two chemicals and explaining the bonding involved.

RESULTS

For our initial study, we focused on student understanding of matter. This study is unusual in that the same assessment was used to map student performance in chemistry across high school and university levels. The same open-ended questions as those given on the post-test to 418 high school students (academic year 2002-03) were administered to 116 first year university students at the UC Berkeley (June '02) after they had completed two semesters of college-level general chemistry. Since these were open-ended questions scored by human raters using scoring rubrics, or guides, as discussed above, it is important to consider the impact of rating on student scores. Inter-rater comparisons of scoring were found to be approximately ±0.15 on the (raw score) *Perspectives* scale of 1 to 5, meaning that having one rater rather than another would, on average, change a student's score by about one-sixth of a level on the 1 to 5 scale.

The student performance map for the Matter variable for one group of students is shown in Figure 4. A logit, or log odds, scale shows at the far left, ranging from -3 to 3, with the mean of the item difficulty estimates at 0. On the left side are noted the location of the respondents at post-test, in the shape of an on-the-side histogram. The levels of student response based on the Matter variable framework and examples of student answers at each of these levels for the given question shows in the text to the right.

Figure 3. Example of a Wright map for the Matter variable. Locations of students shows in the vertical histogram. Levels and examples of student answers at these levels for the given question shows in the text to the right.



Student fit across the instrument was also reasonably good, with only 17 of the 538 students, or only about 3.2 percent of students, with post-test scores outside a standard range . For all these students, the fit was too random, or in other words, the pattern of correct and incorrect responses was not consistent with the estimated difficulty of the items.

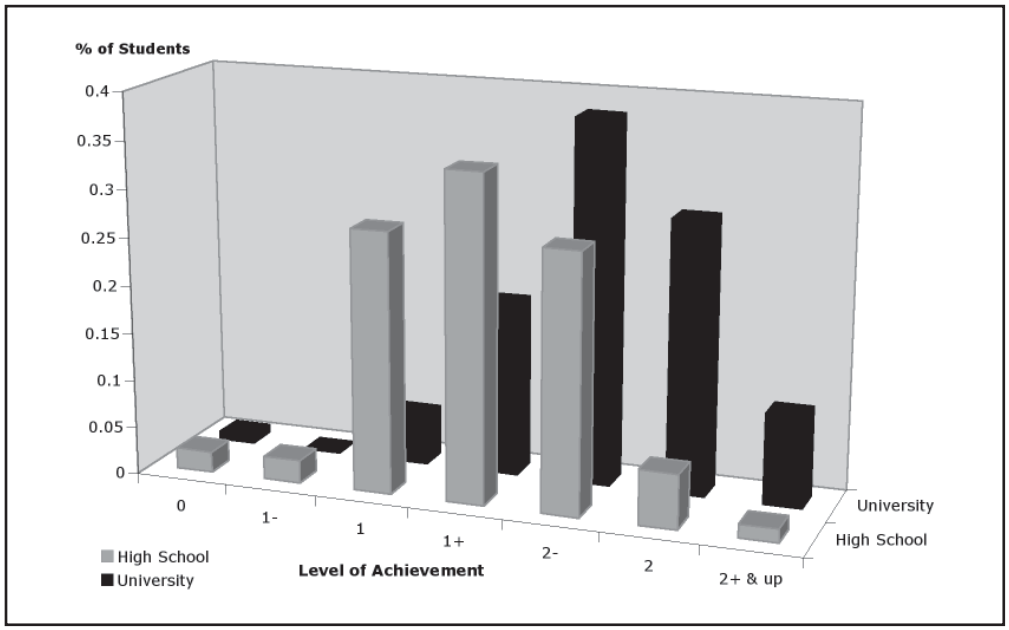
The average standard error for the student proficiency estimates (maximum likelihood estimation) on the post-test was .49 logits (.16 Standard Deviation) on the Matter variable and .50 logits (.27 SD) on the Change variable. EAP/PV reliability was .85 for the Matter variable, and .83 for the Change variable, with person separation reliability estimated at .82 and .80 respectively. Cronbach's alpha cannot be used for overall test reliability estimations in this case due to too much systematically planned missing data across the item bank.

The correlation on student proficiency estimates between the Matter and Change variables was .68, indicating a moderately strong tendency for the way in which students performed on one variable to be similar to how they performed on the other. However, the correlation was still substantially below 1, indicating that information was gained by measuring students on their understanding in both areas.

This pilot study focused primarily on the Matter variable, as the Change item bank was in preliminary development at the time. Subsequent work has been done in the Change area, and also on the third portion of the *Perspectives* framework, Energy, and these data are currently in analysis. The remainder of this paper discusses only the Matter results, pending more information on Change and Energy.

Item response model student proficiency estimates are categorized into the 5-point *Perspectives* scale [22], with cut scores between the categories specified as the mean of the item thresholds, as shown in Figure 4. High school students averaged 1.75 (0.18 SD) on Matter and 1.70 (0.18 SD) on Change. This can be compared to the average score of 2.3 (0.18 SD) out of 5 on both Matter and Change for the college freshman, who had completed college general chemistry as well as usually one year of high school chemistry. The distribution of students on the Matter variable can be seen in Figure 5.

Figure 4. Distribution of High school and University students on the Matter variable.



Interpreting these scores qualitatively, high school students in the sample population were found to be moving from a "notions" conception — based often on real-world knowledge rather than normative chemistry models —toward simple normative models of chemistry that begin to describe and explain matter at a particulate level. These students can articulate their ideas of matter, using prior experiences, observations, logical reasoning, and real-world knowledge to provide evidence for their ideas, but much of the evidence they bring to bear remains out of scope, off-topic, or "invented" relative to normative models of chemistry. Many students are beginning to relate numbers of electrons, protons, and neutrons to elements and mass, and the arrangements and motions of atoms to composition and phase. However, lower achieving students consistently generate answers focusing almost exclusively on macroscopic, and not particulate, descriptions of matter, while even for higher-achieving high school students, ways of thinking and classifying matter are limited to recognizing very basic and simple definitional concepts and representations.

After one year of high school and one year of college chemistry, UC Berkeley students in the sample population scored on the framework in the range of recognition of basic science models of matter, with some students beginning to move to a sound conceptual understanding of multi–relational interactions between atoms, and the generation of accurate causal mechanisms. Many of these students still overgeneralize as they try to relate multiple ideas and engage in problem solving, so students at this level would not be expected to effectively construct many models predicting composition, structure, properties, and amounts of matter, but many are beginning to grasp the multiple relationships needed to reason with these models.

We also have preliminary findings on 399 UC Berkeley students at the completion of their second-year organic chemistry coursework. The students sampled were in organic courses in the medical/biological science pathway and participated in a ChemQuery-designed assessment system called Smart Homework that draws on the same three-strand framework for estimating student understanding. Most had a prior year of general chemistry at the university level and a prior year of high school chemistry, although some combined the high school and university first year by completing AP chemistry, or by taking only one semester of general chemistry in addition to high school chemistry. By the end of second-year organic chemistry, the sampled students in this pathway measured about 20 percent in Recognition, 75 percent in Formulation, and about 5 percent in Construction. Generally, the spread of students over Perspective levels was much greater at the beginning of organic chemistry than at the end, where by completion of organic chemistry most students had progressed to Formulation and none remained in Notions. This is probably a combined effect of learning over instructional time and attrition of lower performing students. It should be noted that the instruments addressed chemistry across the breadth of the field, with only some items specifically in organic chemistry, and thus were not intended to measure progress in the subfield of organic chemistry but overall in the developing understanding and use of chemistry across topics. An organic-specific instrument that did not include content in other areas of chemistry might have found more students in Construction by the end of second-year chemistry. However, this remains to be seen as the ability of students to combine complex, novel ideas that the questions involved may relate to an overall maturation of reasoning skills in the field rather than simply domain knowledge in a recently learned subfield.

DISCUSSION, IMPLICATIONS, AND CONCLUSIONS

While working on this project, we have learned not only about our students, but also about our teaching and our discipline. By grouping the field of chemistry into three core conceptions — scientific models of matter, change, and energy — we were able to focus on the overarching ideas of the discipline while simultaneously

constructing a fairly robust assessment approach that could examine student learning not only within, but between courses. This allows us to obtain a larger perspective on the development of understanding in chemistry, over time and between populations of students. A first key conclusion then is that **the identification of a small number of "big idea" *Perspectives*, carefully developed, based both on expert knowledge and empirical evidence of student learning, might be useful for this type of assessment across courses in other STEM disciplines.**

In the process of examining learning patterns in this way, we are finding that it takes substantial time for students to achieve conceptual change in chemistry models. We are also finding that in the Notions level at the beginnings of that understanding, students often need to extensively explore and use the language of chemistry, and that types of "invented" chemistry reasoning are often apparent in their efforts. An example of "invented" chemistry includes student conceptions of solids and liquids. When students at the upper levels of Notions, who show a great inclination to use invented chemistry, are told that a particular liquid substance and another solid substance have been measured out in amounts such that they weigh the same, students in Notions often want to continue to insist that the solid is heavier. Students employing invented chemistry will often justify their instinctive reasoning with "invented" chemistry ideas drawn from something they have learned in class, for instance that if the solid is iron, the sample should weigh more "because it has a higher mass on the periodic table."

The so-called "invented" ideas are something of a mixed model between the beginnings of more correct and complete chemistry thinking and the kinds of prior knowledge and real-world experiential reasoning that students bring to the table. A second key conclusion then is that **the reasoning in the Notions level, though incorrectly answering the questions, does appear to bring value to the development of understanding, as students who reason with these models are significantly more likely to produce some correct answers on other questions and tasks than are students who attempt to employ only prior reasoning and do not introduce even incorrect attempts to incorporate the chemistry models introduced in instruction.**

Students also show the need for opportunities to use the language of chemistry, as they are being introduced to simple beginnings of models. A third key conclusion is that in order to "speak the language of chemistry," we have found that students **need substantial opportunity to explore chemistry words and symbols, before they become fully able to reason in meaningful ways with the symbol systems.** For instance, in exploring patterns in metals, students may know from real-world experience that silver, gold, and copper are all metals. But when asked to talk about what properties these three metals share, students may focus directly on the symbolic language rather than what they know about metals, citing for instance that Ag, Au and Cu all include either the vowel "A" or "U." Discussion and exploration of the symbolic language with others, including peers in their courses, will often allow students to delve beneath the symbols to connect to concepts they have some knowledge of regarding metals, such as ideas of luster and hardness, which students can readily connect to metals when English-language words such as gold and silver are used rather than the chemical symbols.

Recognizing that the chemical symbolism is a new language is important both for students and for chemistry instructors. Time to work on decoding the symbols and looking for patterns in the symbols gives students time to use and "speak" with the new language.

Our evidence to date shows that students enter high school chemistry in the lower areas of Level 1 (Notions), and that they are on average approaching Level 2 (Recognition) after a year of general chemistry in high school.

On average in our data sets, students were solidly exhibiting Level 2 reasoning following one year of university-level general chemistry, with some students able to reason in the lower regions of Level 3 (Formulation). Our most recent work is showing that some students at the end of organic chemistry are reaching Level 4 (Construction) reasoning. Level 5 (Generation) is not expected to be reached in most cases until graduate school.

In summary, often students are found to hold onto prior beliefs in chemistry and to develop models of reasoning that mix or blend preconceptions with scientific reasoning. A fourth key conclusion is that *understanding better the process and progress of learning for conceptual change in chemistry, and in other science disciplines, may help us to know what the blended models are that students develop, and what is helpful for bridging to new and more powerful understandings of science.* This paper shows some ways in which criterion-referenced assessments can help us to think about what students actually know and how to help them learn. By measuring across courses and considering the development of reasoning with specific scientific models over time, a clearer picture of both student progress and how people learn might be gained.

While our investigations involved chemistry, we believe there may be some commonality in other STEM disciplines, both at K-12 grade levels and across many university courses. Pointing out the existence in other STEM disciplines of the same issues considered in our four conclusions may be helpful for a larger dialogue about how to understand and support STEM student learning, in part with improved assessment approaches.

ACKNOWLEDGEMENTS

This material is based on work supported by the National Science Foundation under Grant No. DUE: 0125651. The authors thank Rebecca Krystyniak, Sheryl Mebane, Nathaniel Brown, and Karen Draney for their assistance with instrument and framework development, data collection, scoring, and assistance in discussions of student learning patterns.

REFERENCES

[1] U.S. Department of Education, *A Test of Leadership, Charting the Future of U.S. Higher Education*, A Report of the Commission Appointed by Secretary of Education Margaret Spellings, *http://www.ed.gov/about/bdscomm/list/hiedfuture/index.html*, 2006.

[2] American Association of State Colleges and Universities, "The Spellings Commission and the States," *Policy Matters*, vol. 3, no.9, pp. 1-4, 2006.

[3] Association of American Colleges and Universities, *Statement on Spellings Commission Report*, Board of Directors, AAC&U, http://www.aacu.org/About/statements/Spellings9_26_06.cfm, 2006.

[4] Scalise, K., *BEAR CAT*: *Toward a Theoretical Basis for Dynamically Driven Content in Computer-Mediated Environments*, dissertation, University of California, Berkeley, 2004.

[5] Scalise, K., Claesgens, J., Wilson, M., and Stacy, A., "Contrasting the expectations for student understanding of chemistry with levels achieved: a brief case-study of student nurses," *Chemistry Education Research and Practice, The Royal Society of Chemistry*, vol. 7, no.3, pp.170-184, 2006.

[6] Scalise, K., Assessment Instrument Report for "Living by Chemistry: Inquiry-Based Modules for High School," *Berkeley Evaluation and Assessment Research (BEAR) Report*, (University of California, Berkeley, Berkeley), 2001.

[7] Driver, R., Squires, A., Rushworth, P., and Wood-Robinson, V., *Making Sense of Secondary Science: Research into children's ideas*, (Routledge, New York), 1994.

[8] Gabel, D., and Samuel, K., "Understanding the Particulate Nature of Matter," *Journal of Chemical Education*, vol. 64, no.8, pp. 695-697, 1987.

[9] Hesse, J., and Anderson, C. W., "Students' Conceptions of Chemical Change," *Journal of Research in Science Teaching*, vol. 29, no.3, pp. 277-299, 1992.

[10] Krnel, D., Watson, R., and Glazar, S. A., "Survey of research related to the development of the concept of 'matter'," *International Journal of Science Education*, vol. 20, no.3, pp. 257-289, 1998.

[11] Mulford, D. R., and Robinson, W. R., "An Inventory for Alternate Conceptions among First-semester General Chemistry Students," *Journal of Chemical Education*, vol. 79, no.6, p. 739, 2002.

[12] National Research Council, *How Students Learn: Science in the Classroom/Committee on How People Learn, A Targeted Report for Teachers*, Donovan, M.S., and Bransford, J.D. (eds.), (National Academies Press, Washington, D.C.), 2005.

[13] National Research Council/National Academy of Sciences, *National Science Education Standards*, (National Academy Press, Washington, D.C.), 1996.

[14] Niaz, M., and Lawson, A., "Balancing chemical equations: The role of developmental level and mental capacity," *Journal of Research in Science Teaching*, vol. 22, no.1, pp. 41-51, 1985.

[15] Samarapungavan, A., and Robinson, W., "Implications of Cognitive Science Research for Models of the Science Learner," *Journal of Chemical Education*, vol. 78, no.8, online symposium accompanying print journal, 2001.

[16] Sawrey, B., "Concept Learning versus Problem Solving: Revisited," *Journal of Chemical Education*, vol. 67, no.3, pp. 253-254, 1990.

[17] Smith, C. L., Wiser, M., Anderson, C. W., and Krajcik, J., "Implications of Research on Children's Learning for Standards and Assessment: A Proposed Learning Progression for Matter and the Atomic Molecular Theory," *Measurement: Interdisciplinary Research and Perspectives*, vol. 4, nos. 1 & 2, pp. 1-98, 2006.

[18] Taber, K. S., "Chemistry lessons for universities? A review of constructivist ideas," *University Chemistry Education*, vol. 4, no.2, pp. 26-35, 2000.

[19] Talanquer, V., "Commonsense Chemistry: A Model for Understanding Students' Alternative Conceptions," *Journal of Chemical Education*, vol. 83, no.5, p. 811, 2006.

[20] Nurrenbern, S., "Piaget's Theory of Intellectual Development Revisited," *Journal of Chemical Education*, vol. 78, no.8, online symposium accompanying print journal, 2001.

[21] Nurrenbern, S., and Pickering, M., "Concept Learning versus Problem Solving: Is there a difference?," *Journal of Chemical Education*, vol. 64, no.6, pp. 508-510, 1987.

[22] Scalise, K., Claesgens, J., Wilson, M., and Stacy, A., "Mapping Student Learning Outcomes in Higher Education: A Case Study in Science, " in preparation.

Discovering Functional Role Specialization: The Effect on Undergraduate Student Learning of Engineering Project Teams

Jeannie Brown Leonard, Linda C. Schmidt, Janet A. Schmidt*, and Paige E. Smith - *University of Maryland*

OVERVIEW

This article presents an overview of our ongoing research supported by NSF's Assessing Student Achievement Program, which investigates the connection between student roles on team projects and learning engineering content. After briefly tracing the researchers' background and interest in this area of inquiry, we describe the research methods used, important findings, and finally the implications for 1) understanding student learning and, 2) based on this understanding, designing pedagogy to enhance student learning in engineering courses that use project teams.

BACKGROUND

This ASA-funded project grew out of a long-standing collaboration among Drs. Linda Schmidt, Janet Schmidt, and Paige Smith, as well as other colleagues, on the impact of team experiences on various learning outcomes in undergraduate engineering classes. Under the BESTEAMS (Building Engineering Student Teams and Management Systems) grant, this earlier work focused on providing engineering faculty with pedagogical tools related to teaching undergraduate students effective teamwork skills [1, 2, 3]. The BESTEAMS project created a range of curricular training modules that have been published as a book called *A Curriculum Guide for Engineering Faculty* [1].

RESEARCH ON FUNCTIONAL ROLES: RELATED LITERATURE

The use of project teams in engineering education is widespread. Engineering educators have attempted to teach students in a simulated "real world" environment by using industrial-type projects in the classroom. Interest in teams has also grown because of ABET's (2000) mandate on providing student learning outcomes that include the "ability to work on multidisciplinary teams" [4]. Creating high-quality learning experiences for students is difficult even when simulating engineering industry teams, because student teams cannot duplicate the diversity of experience both in team process and functional expertise of professional teams [5].

The influence of roles on team functioning is well documented in literature outside of engineering education, especially in business and organizational settings [6, 7]. Much of the focus on team research has been on the team member behaviors that contribute to a well-functioning team. This emphasis on team behaviors, commonly called team dynamics, has yielded important insights and suggests that communication, planning, being receptive to constructive criticism from teammates, being skilled at giving such criticism, positive attitudes, commitment, and flexibility are all important elements to strong team functioning. Additional research has noted that behaviors often come together into distinct positions with characteristic tasks and responsibilities, which are called roles [6]. Process roles include leader, facilitator, recorder, and organizer. In contrast to the roles that keep the project moving forward in terms of interpersonal dynamics, functional roles are crucial to the success of engineering projects because they are related to the technical demands of the task (e.g., designer, report writer, programmer). These task roles have been discussed in the more general team literature as "task work"[6, 7].

Other literature on teams identifies how roles and the parameters of the project interact. The business literature on teams makes a distinction between 'functional' and 'divisional' roles [8, 9]. The type of roles that are most effective to team functioning depend on the nature of the team environment. In this context, functional roles are narrow and specialized. Teams that distribute work based on functional roles are interdependent in that each team member is contributing a unique aspect to the end product and, therefore, all team members are essential to the efficient completion of the project. Divisional roles are more diffuse and general. There may be overlap in some tasks among team members and they may use some of the same skills, but they work more independently overseeing a specific territory or region. The divisional approach contributes to flexibility and allows for greater responsiveness to changing conditions. Teams that use functional roles tend to be very efficient, and are most effective in stable, predictable work environments. The task is completed, but individual team members' broad mastery of skills may be compromised.

RESEARCH QUESTIONS

A goal of this study was to understand the team experiences of undergraduate engineering students at different stages in the academic pipeline. The following research questions guided our inquiry:

1) What specific functional roles can be identified in engineering project teams?

2) Is there a connection between functional role performance and learning engineering content and skills?

3) Do upper-level students who have had a variety of team experiences gravitate toward roles (process or functional) that they have had in the past?

4) Do women and students from underrepresented racial/ethnic groups have different role-taking and team participation experiences compared to white men?

5) What insights do faculty have about the team project experiences in their courses?

METHOD

This study used focus groups to learn about student experiences on engineering project teams. By meeting in a group rather than conducting individual interviews, team members could react and respond to (both affirming and disagreeing with) their teammates or other focus group participants. Focus group facilitators also could ask follow-up questions to learn more about themes or ideas shared by the students. In all, we conducted 15 focus groups. The focus groups included intact teams from specific courses across the engineering curricula, including first-year students in an introductory, generalist engineering course as well as second-, third-, and fourth-year students in mechanical engineering. Senior students across all the engineering disciplines also were included (i.e., aerospace, civil, and electrical, in addition to mechanical). We also held focus groups with students and professionals who met specific criteria but were not necessarily on a team together. We were interested in comparing the experiences of women and students from underrepresented groups with the experiences of white men in their senior year. Table 1 summarizes the focus group composition.

All participants were volunteers and received a $20 gift card from the University bookstore (a national chain) in appreciation for their participation. Two investigators conducted each focus group: one facilitator experienced

in running focus groups and one engineer. The focus group protocol began with introductory questions about participants' previous experience on teams, followed by questions about team functioning, team roles (functional and process), and learning engineering content on teams. Questions were open-ended and neutral to avoid leading the participants toward a certain response. All focus groups were audio recorded and transcribed for analysis. A qualitative analysis software program, N6, helped manage the data and assisted the investigators with identifying patterns and themes in the data.

Table 1. Focus Group Composition

Course	Composition	Goal	Number
ENES 100: Introduction to Engineering Design	First-year students across engineering disciplines; intact teams	Establish baseline of team behaviors and knowledge of team roles	2
Senior Capstone	Seniors in capstone course for aerospace, electrical, mechanical, and civil engineering; intact teams	Examine advanced student knowledge of team issues and compare across disciplines	4
ENME 220: Mechanics and Materials ENME392: Statistical Methods for Product and Process Development	Sophomore and junior mechanical engineering majors; intact teams	Explore intermediate student knowledge of team issues within mechanical engineering, a major with extensive team experiences	2
Senior Capstone	Senior women	Explore whether role choice or participation on teams is influenced by gender	2
Senior Capstone	Senior minority students (from racial/ethnic groups underrepresented in engineering)	Explore whether role choice or participation on teams is influenced by race	2
Alumni & Industry Team	School of Engineering alumni currently working or retired from engineering	Compare industry teams with student teams	2

Additionally, we interviewed faculty members who teach the courses from which the focus groups were recruited.

CENTRAL FINDINGS

The most relevant findings with implications for student learning and engineering pedagogy are presented below.

Functional Roles

Students readily identified and named roles related to tasks or technical skills they were contributing or had contributed to team projects. First-year students identified the following functional roles: programmers, builders, web site developers, meeting organizers, photographers, and writers. Sophomore, junior, and senior students provided more engineering-based functional roles including: techies, data collectors, number crunchers, project managers, researchers, and component designers.

Learning Engineering Content Skills

We found consistent evidence that students take on the functional roles in which they feel most comfortable or competent. This tendency increases from year to year as students complete their engineering degree. In the first-year course, students were making the transition from high school style group projects and still expected the professors to provide all the information needed for mastering every aspect of the course.

- *"I feel like I am paying for my education so I want someone to teach me because I'm paying for it and I'm not really a hands-on person."*

On these novice teams, team members were rarely appreciated for what they could teach each other. Generally, students choose or are assigned project tasks that they like or have some previous experience doing. Most students in the introductory engineering design course correctly assume that the primary learning goal of the team project is learning to work on teams.

- *"It helps that a lot of our group had prior knowledge of things going into [the project]…we had someone who could already program, do the webpage, do the wiring."*
- *"I think I've learned more about being on a team than any kind of academic information so far."*

By the sophomore and junior years, the students recognize and accept that they can learn from their team mates.

- *"If you are a specialist, it can be good because, one, hopefully the rest of your group can take a little bit away from, you know, your increased knowledge of the subject."*

To complete a complex team project students have to perform many tasks independently. Again, students choose or are persuaded to do tasks they already do well for the sake of efficiency.

- *"Usually when you're on a team, you're like, well, what do you like to do? This is what has to be done. So, if you like to do that, just go do it so we can get it done."*

By their senior year capstone course, this 'divide and conquer' approach is well established. Various team members are assigned functional roles in which they are proficient and the team is confident they will be done correctly. Team members have become specialists.

- *"The roles fell into place….we found people's strengths and you know how people work."*

Thus, students rarely have the opportunity to learn new engineering skills through the team project.

Role Specialization

The study found that engineering students tended to gravitate toward the same functional roles over and over: the ones in which they specialize.

- *"We each know a different area that we can concentrate on. If I was doing it all by myself, I wouldn't know where to start."*
- *"It [being a specialist] is always going to be helpful."*

However, not all functional areas are needed for every project, thereby marginalizing some specialists (e.g., the website developer).

Experiences of Women and Underrepresented Students

Students report few limitations in their choice of team roles. Some, for whom English is not their first language, commented that they were rarely chosen to be the 'report writer.' However, others indicated that their attentiveness to grammar had recommended them for the editor role. Some African American and Latino students described minor, generalized tension or ambiguity around race, but the overwhelming finding was that if students do good work (perform their roles well), race was not an issue. Again, quality and efficiency in task completion was the ultimate goal.

The senior women who participated in women-only focus groups were asked to comment directly on role limitations based on gender. They did not voice any direct concerns. The consensus portrayed is that skilled female students can fully participate on the team if they assert themselves. However, various examples provided by the women directly challenged this perceived equity such as statements about not getting the chance to use power tools or being asked to write most of the reports. It is likely that women tuned out negative behaviors of men on their teams, again perhaps in service of project completion.

Faculty Perspectives

Faculty acknowledge that students bring different skill sets to the project without identifying separate functional roles taken on by students. Engineering faculty generally do not acknowledge functional roles in teaching or grading.

- Faculty who were interviewed about their view of student projects teams noted that they generally assumed that students would be dividing up the work to complete the team project, *"There's so much to do that they have to break it up."*

In some cases faculty would prescribe the steps the team needed to take to successfully complete the project, being particularly attentive to an even distribution of the workload:

- *"Typically, when I introduce the project, I talk about what they have to do and I generally outline the steps they have to go through in order to solve it."*
- *"You really do have a certain skill set to understand how to design the survey, get the information, go out and do that, and that takes a certain amount of time. So it's both skill set and effort…pretty good distribution of labor so you wouldn't have resentment."*

Faculty confess to not really knowing if individual students master the desired educational content or process knowledge driving the team assignment. Instead, faculty often rely on team member peer evaluations for individual assessment. Often only the project deliverables are graded, rather than individual contributions or mastery.

- *"I don't know how much they all learn of the project."*
- *"The project is due the last day of the class. And unless they see me next semester…they don't even know how they did, or they don't even get to read through their project other than the grade I put up on the website. So that's a problem."*

DISCUSSION

The data from this study indicate that functional roles are a construct embraced by undergraduate students and adopted as a strategy for completing complex projects. Functional roles promote efficiency and specialization

leads to even greater time-saving benefits. By dividing the responsibilities among team members with specialized skills, overall team functioning and the quality of the end-product increase. Exploiting specialization is a strategy used by engineering team members with implications for learning that are generally unrecognized by faculty. Professors usually create a 'predictable' environment that favors specialization and unintentionally reinforcing students' adoption of functionally-based, efficient task roles [10]. As a result, engineering educators appear to be short-changing their students by allowing them to become task specialists at the expense of broad learning of all aspects of the project.

This study identified increasing functional role specialization as students progressed through their education from freshmen to senior years. Students may have more than one specialization or the ability or preference to do certain tasks. Again, with task efficiency and the desire to prevent group dynamic problems, students reported, *"if someone else really wants to do the building, I agree and do something else."* Alternatively, some students reported wanting to learn material outside their determined team role and feeling frustrated that the team structure did not allow it. Fellow junior and senior team members were viewed more often as excellent sources of information and potential tutors. Students rarely reported peer-to-peer learning sessions, claiming a lack of time.

IMPLICATIONS

The research data suggest that the ways in which faculty design team projects may inadvertently promote functional role specialization, thus compromising individual learning. However, faculty designing engineering projects for their class might consider the following to reduce the negative consequences of specialization:

1) Create team assignments and role assignments that require a rotation of responsibilities either within a project or across projects in a sequence of required courses.

2) Include meaningful assessment of individual contributions to the team and individual mastery of project-related engineering content (e.g., homework and quizzes). Identifying clear expectations and learning goals for the team project experience, including individual learning outcomes, is the foundation for effective assessment.

3) Mitigate the influence of "too little time" as a barrier to learning new skills by emphasizing learning over the quality of the final product in the project goals. This shift in what is valued by the faculty member must be reflected in the grading policy.

While this study points out unanticipated learning restrictions as a result of specialization, it is also important to note that specialization is an efficient project completion strategy and a characteristic of engineering project teams in the world of work. With that in mind, perhaps specialization can be exploited in multidisciplinary teams that work together across several semesters. Alternatively, a longer time period for team projects may encourage more peer-to-peer learning from functional role specialists in a single disciplinary course.

Functional role specialization exists in engineering project teams. It can act as a detriment to comprehensive learning of engineering content. But, if managed correctly, functional role specialization can enrich the team learning environment.

REFERENCES

[1] Schmidt, L.C., Schmidt, J., Smith, P., Bigio, D., and Bayer-Contardo, J., *BESTEAMS (Building Engineering Student Teams Effectiveness and Management Systems): An Instructor Guide*, (College House Enterprises, LLC), February, 2005.

[2] Morgan, Jr., B.B., Glickman, A.S., Woodward, E.A., Blaises, A. and Salas, E., *Measurement of team behaviors in a navy environment (NTSC-TR-86-014),* (Naval Training Systems Center, Orlando, Fla.), 1986.

[3] Baker, D.P., and Salas, E., "Principles for measuring teamwork skills," *Human Factors*, vol. 34, no. 4, pp. 469-475, 1992.

[4] ABET Engineering Criteria, 2006-07. http://www.abet.org/forms.shtml

[5] Schmidt, L.C., "Engineering Teams: Individual or Group Sport?" *International Journal of Engineering Education*, vol. 22, no. 3, pp. 659-664, 2006.

[6] Gibbs, Focus Groups, *Social Research Update*, 19, Conducted for the Department of Social Medicine at Briston University, 1997.

[7] Rees, F., *Teamwork from start to finish*, (Jossey-Bass, San Francisco), 1997.

[8] Heany, D.F., "Prominent Barriers to Interfunctional Team Play," in *Cutthroat Teammates: Achieving Effective Teamwork Among Professionals*, (Dow Jones-Irwin), 1989.

[9] Hollenbeck, J. R., Moon, H., Ellis, A., West, B., Ilgen, D.R., Sheppard, L., Porter, C.O.L.H., and Wagner, III, J.A., "Structural contingency theory and individual differences: Examination of external and internal person-team fit," *Journal of Applied Psychology*, in press.

[10] Hackman, J.R., and Oldham, G., "Motivation through the design of work: Test of a theory," *Organizational Behavior and Human Performance*, vol. 16, pp. 250-279, 1976.

Assessing Pedagogical Content Knowledge Of Inquiry Science Teaching: Developing An Assessment Instrument To Support The Undergraduate Preparation Of Elementary Teachers To Teach Science As Inquiry

David Schuster*, William Cobern, Brooks Applegate, Renée Schwartz, Paul Vellom and Adriana Undreiu - *Western Michigan University*

ABSTRACT

There is a national commitment in the United States to the teaching of science as *inquiry* across the K-12 grades. Inquiry teaching of science reflects the investigative approach scientists use to discover and construct new knowledge. Nevertheless there remains debate about appropriate science instructional approaches, across a spectrum from 'direct' instruction to pure 'discovery' learning. Standards documents advocate a guided inquiry approach, often reflected in a learning cycle.

During their undergraduate preparation, pre-service teachers usually take a number of science content courses and a broad science teaching methods course. They are regularly assessed on their science *content* knowledge, but there is a critical need to complement this by also assessing *pedagogical content knowledge of inquiry science teaching*. That is, knowledge of topic teaching practices that reflect the inquiry nature of science. Currently this is rarely assessed because suitable instruments are not available.

In this project we are developing and validating such an assessment instrument, which we call POSITT, the Pedagogy of Science Inquiry Teaching Test. It consists of case-based objective items based on realistic vignettes of classroom teaching situations in K-8 science topics. An item poses a problem on the vignette and offers response options reflecting a spectrum of teaching orientations. Developed items are piloted twice on a large scale with pre-service teachers at diverse institutions, with analysis providing data for refinement and estimates of reliability. Suitably balanced item sets are compiled into POSITT instruments with grade range variants. POSITT is finally field validated by studying its predictive power with respect to classroom observations of actual teaching practice, taking into account possible covariant factors.

The instrument should be useful to teacher educators and researchers at any institution for assessing pedagogical content knowledge of inquiry science teaching (PCK-IST), as well as for evaluating and improving the effectiveness of instructional programs in this regard. Equally importantly, items can be used formatively to promote contextualized understanding of science inquiry pedagogy, for both pre- and in-service teachers. Examples of assessment items are provided.

INTRODUCTION, BACKGROUND, AND RATIONALE

Teaching science as inquiry

The United States has developed a national commitment to the teaching of science *as inquiry* across the K-12 grades [1,2,3,4]. *Inquiry teaching of science* refers to a curriculum approach and pedagogy that reflect the investigative approach and techniques that scientists use to discover and construct knowledge. Standards documents of the American Association for the Advancement of Science [1] and the National Research Council [2] advocate a scientific inquiry approach to teaching and learning. Inquiry processes of science range across aspects

like observing, posing questions, exploring phenomena, experimenting, interpreting data, seeking relationships, inferring from evidence, suggesting and testing hypotheses, developing concepts, laws and models, and communicating and explaining findings. 'Science' thus encompasses both *process* and *product* aspects, which are equally important. There are corresponding instructional approaches that reflect this, such as Karplus's learning-cycle approach [5].

The concept of teaching and learning science as inquiry is not new, but the origins of the modern day movement lie in the NSF-funded curriculum projects of the 1960s [6,7,8,9,10]. The effectiveness of those inquiry-based curricula was documented first by Shymansky et al. in 1983 [11] and again in 1990 [12], and by Secker and Lissitz [13] and Secker [14] using updated techniques.

Despite the emphasis on guided inquiry in national standards and state curricula, there is continuing debate (and dissent) about desirable and effective approaches to science instruction. There exists a range of approaches across a spectrum from 'direct' didactic presentation on the one hand to unguided 'discovery' learning on the other, each approach with its advocates. Teachers likewise have conscious or unconscious *orientations* toward one part of the instructional spectrum or another. There are both educational and political aspects to this debate.

There are sometimes different interpretations of what is meant by inquiry instruction, especially when usage is not made explicit. The approach advocated in standards documents of the NRC and AAAS is one of 'guided inquiry'. For example, science instruction based on the Karplus [15] learning cycle involves three main stages: exploration, concept development, and application. The teacher guides the process and intervenes as necessary to help students through the development sequence. This is certainly not synonymous with minimally-guided 'discovery'.

Pedagogical content knowledge

Teaching science is a demanding task, requiring teachers to understand not only the science *content* but also how to *translate the content and methods of science into analogous instructional practices.* Such ability is what Shulman called *pedagogical content knowledge* or PCK [16,17,18,19]. PCK is the knowledge of effective instructional practices pertinent to specific content areas. For science teaching, that emphatically includes understanding of *inquiry* as an approach to the subject [20,21,22].

Assessment during teacher preparation

During their undergraduate preparation, pre-service teachers normally take a number of science content courses and a broad science teaching methods course. Their knowledge of science *content* is regularly assessed in the science courses. The national move toward inquiry-based teaching means that an equally important knowledge component is *pedagogical content knowledge of inquiry science teaching;* that is, knowledge of teaching practices that specifically reflect the investigative nature of science. However, this crucial ability is rarely assessed directly or regularly during students' preparation. This contrast is an anomaly which likely arises because understanding of inquiry pedagogy, in specific topic contexts, is harder to assess than science content, and we do not currently have suitable instruments. Surely a necessary component of our national effort to improve science education must be the ability to effectively and efficiently assess teachers' pedagogical knowledge of inquiry science teaching. Without it, it is difficult to evaluate and improve undergraduate science teacher education programs, or to ensure competence of graduates in this regard. Currently, such an assessment tool is not available. Iris Weiss, a leading authority on science assessment, concurs: "I don't know of any such set of items, and I agree with you that this work is sorely needed!"

Undergraduate pre-service teachers do in fact have their teaching *practice* evaluated [23], but this is by classroom observation (e.g. [24]), toward the *end* of a program, and in limited measure. Although observational evaluation of practice is very important, as a practical matter only a few science lessons covering one or two topics can be evaluated in this labor-intensive way. Also, evaluations of pre-service elementary teachers tend to be done by teacher educators with little science background.

Desired assessment instrument

All the above considerations motivated this project – to develop and validate an assessment tool for testing pre-service teachers' pedagogical knowledge of inquiry science teaching.

For teachers, planning and implementing successful inquiry-based learning in the science classroom is a task which demands a *combination* of science content knowledge and inquiry pedagogy knowledge. Correspondingly, an assessment instrument must reflect this combination, and do so in the context of specific science topics and teaching situations. Thus, the main aims of this project are: to develop an objective, case-based assessment tool for testing pedagogical content knowledge of inquiry science teaching, to pilot and refine the assessment items, and to field-validate the resulting instrument against observed teaching practice. We dub this instrument POSITT, the Pedagogy of Science Inquiry Teaching Test. The final format will be sets of objective items based on realistic classroom scenarios and teaching issues encountered in practice.

Note that the instrument will not be based on broad statements about inquiry generally. There exist some tests of teachers' general beliefs about inquiry [25], but these serve a different purpose and are more distant from practice and content. We also note again that POSITT is *not* an observational instrument to evaluate inquiry teaching practice in the classroom. There are existing tools for this, such as the Reformed Teaching Observation Protocol (RTOP), and the Lesson Observation System used by the Science and Math Program Improvement (SAMPI) group. Rather, our instrument will serve other functions. It is a readily administered objective instrument for use during undergraduate instruction of prospective teachers, to both assess and promote understanding of inquiry science pedagogy. For our purposes, the role of observation of teaching practice will be to field validate the instrument's predictive validity.

Formative and summative uses

POSITT can have both summative and formative purposes. It is critical to ascertain how students moving through their preparation programs as teachers of science are developing the knowledge of inquiry that will form the basis for their practice. Equally importantly, feedback from formative assessment is known to be a singularly effective factor in learning [26]. In this regard the instrument could clearly support the professional development of both pre-and in-service teachers, in enhancing expertise in inquiry science teaching.

Target and level

Our focus in this project is on *undergraduate students preparing to be teachers of science.* Of particular interest is the inquiry pedagogical knowledge required for teaching science at the elementary (K–8) grades, because this sets the foundation for future learning. It is critical that children leave these grades with a good grounding in both the content and processes of science, as they progress to the more specialized study of the sciences at the secondary level. Nevertheless, the type of understanding required for good science inquiry teaching is surely not

grade-constrained, so the instrument could serve its purpose for other grade ranges too, or be extended in content coverage. In fact, we suspect the items might even be useful for introductory college instructors seeking to gain an appreciation of inquiry approaches and hence improve their science instruction.

THE ASSESSMENT TOOL – OUTLINE OF DEVELOPMENT, TESTING, AND VALIDATION STAGES

For all the reasons above, science teacher education desperately needs an assessment tool specific to the inquiry teaching of science, of broad scope and ease of use, validated by research, which can be readily used for both summative and formative purposes.

Scientifically based research and development

The U.S. Department of Education [27] has called for the use of scientifically based research and evidence as the foundation for education programs; and as *The Condition of Education 2002* states: "Reliable data are critical in guiding efforts to improve education in America" [28,29,30,31]. Accordingly, our project involves a multistage plan for test development followed by two rounds of piloting and revision, concluding with blinded field testing against classroom practice for validating the assessment tool.

Project outline

To that end, the project involves the following aspects:

- *Developing criteria for items regarding inquiry pedagogy*
- *Identifying various types or classes of items that are possible for this new type of assessment, characterizing by question type, formulation, response options, and item format*
- *Creating problem-based items involving realistic teaching scenarios across a range of representative K-8 elementary science topics*
- *Ensuring that items reflect diversity in race, gender, geographic, and economic contexts*
- *Establishing standardized scoring and administration directions for the instrument*
- *Piloting and revision of items on a large scale with about a thousand diverse undergraduate pre-service student teachers at collaborating institutions across the country*
- *Computing initial estimates of reliability*
- *Field validating the instrument by studying predictive power with respect to observed teaching in classrooms, for both pre- and in-service teachers*

Once fully developed and validated, this assessment tool, addressing a central area where there is currently a huge assessment gap, should be of value to every undergraduate science teacher program in the country. For the first time we will be able to assess science inquiry pedagogy knowledge as readily as we assess science content knowledge.

Design Considerations for the Assessment

It is well known that while objective-type tests are easy to administer and score, the creation and refinement of quality objective items with meaningful choice options is very demanding, especially items that measure understanding in realistic situations. Thus, the creative work of item construction forms a substantial and crucial part of this project. This is followed by careful piloting, revision, and validation. Here we describe the design considerations used to develop the individual items and final POSITT instrument.

View of science inquiry teaching and criteria for inquiry items

For any pedagogy-of-inquiry assessment tool to be effective, it must be based upon a widely accepted view of what appropriate inquiry teaching *is*, and how guided inquiry contrasts with other approaches across a spectrum from 'direct instruction' to open 'discovery'. We use the view of inquiry presented in two documents that have gained broad national acceptance, viz. *National Science Education Standards* [2] and *Science for all Americans* [1], and have developed a set of Inquiry-Item-Criteria to guide item development and evaluation. In abbreviated form, the criteria specify that inquiry–based science instruction proceeds such that learners: 1) engage in scientifically oriented questions and explorations, 2) give priority to evidence in addressing questions, 3) formulate explanations from investigation and evidence, 4) connect explanations to scientific knowledge, and 5) communicate and justify models and explanations.

Science content and level

The science content must be grade-appropriate, and, again, for this we look to the *National Science Education Standards* [2] and *Science for all Americans* [1]. Item scenarios are based on science content specified in standards for the K-8 grades. Note, however, that since we are testing science pedagogy knowledge, items should serve this purpose reasonably well at the secondary level also, within the content range covered.

Case-based teaching vignettes

As noted, planning and implementing successful inquiry-based learning in the science classroom is a task demanding a *combination* of science content knowledge and inquiry pedagogy knowledge – the latter not just in general terms but *as applied to specific topics and cases*. A feature of our POSITT tool is that the assessment items are posed not in generalities about inquiry, but instead are based on specific science topics in real teaching contexts.

The model we will use for item design is based on the curriculum technique of Problem-Based Learning (PBL), which has been widely used in medical education [32,33], and more recently adopted in science teacher education [19,34,35]. Case- or problem-based approaches present students with a practical problem, in our case in the form of a teaching *scenario* or *vignette* representing a realistic K-8 science-teaching situation. Thus, each item begins with a specific vignette followed by a question about it and a set of response choices. Responses might be possible evaluations of the teacher's actions so far, or alternative suggestions for what the teacher should do next, or ways of handling a particular event or question.

Examples of items based on teaching vignettes are provided in the Appendix. It will be seen that the examples reflect the kinds of instructional decisions that teachers have to make every day, both in lesson planning and on the spot in the classroom.

Basing a pedagogy-of-inquiry assessment on PBL and concrete cases has several advantages. Firstly, the assessment is realistic and, hence, more authentic, being built upon actual classroom occurrences. Secondly, it is an assessment that does not lapse into measurement of rote memory, or generalities about inquiry. Each item specifically requires either application or evaluation, in terms of Bloom's Taxonomy [36], in specific situations. Successful application and evaluation requires both *understanding* of inquiry and *how to use it* in real cases. Thirdly, because the assessment involves pedagogical approaches in real contexts, individual items are easily adapted for *formative* use with pre-service students. In order to fully grasp an area, students need the experience of *applying* knowledge to a range of problems, else their knowledge may be inert. Once a set of problem-based items on inquiry pedagogy is available, it can be used to help students develop a usable understanding of the general principles they are learning. Our items are essentially 'problems' involving alternative pedagogical approaches to a given teaching situation. Working through such problems with students operates as a scaffold for novices' current lack of schemas and serves as a basis for effective instruction based on active engagement with example cases.

Assessment formats

For its intended purposes, the assessment tool needs to be readily administered, easily scored, and reliable. Thus we have chosen an objective format with variations. For the development of the instrument, every item is first written in three different but related formats, viz.

1) **Multiple choice (MC) format:** *Stem with 4 or 5 response choices*

2) **Ranked Response (RR):** *Variation requiring ranking of the response choices [37,38]*

3) **Short-answer constructed response (CR):** *Students construct their own responses*

Because we are developing multiple items and item types to address each component of the Inquiry-Item-Criteria, we will be able to compare item difficulties and discrimination, which will provide evidence for the validity of the assessments. As such, the development of the assessment tool can be conceptualized in a similar framework as a multi-trait, multi-method (MTMM) type of approach advocated by Campbell & Fiske [39] (see also [40]). In our study, the 'traits' are Inquiry-Item-Criteria, and the 'methods' are the three item formats (see Appendix for the MTMM table). Although, in our application of the MTMM we will not be examining discriminate validity as originally conceptualized. The MTMM matrix approach to organizing reliability and validity information, particularly among the three different item formats, provides a logical and conceptual schema for organizing the psychometric information. Results on the Ranked Response (RR) and Constructed Response (CR) formats will also help refine the foils for the Multiple-Choice (MC) format. Note that having three formats initially will help develop and validate the items, but the final instrument will normally be used in the multiple-choice format. If instructors wish, they can also have students explain their answers in short constructed responses, especially when using items formatively.

ITEM DEVELOPMENT

The goal is to produce 30 pedagogy-of-inquiry assessment items for each of the three grade groups, for a total of 90 items, each in three formats, giving a collection of 270 individual items. Various POSITT instruments can be constructed from these; a test administered to a particular group will have about 30 appropriately selected items. Item development procedures are as follows.

Item writing group

The item writing group consists of four project investigators, a doctoral research associate, and five experienced schoolteachers representing three grade groups, K-2, 3-5, and 6-8. It is important to have teachers involved because they have direct knowledge of situations that teachers commonly encounter.

Item typology and writing guide

During the initial item creation stage we analyzed what main *types* of items are possible for scenario-based assessment of inquiry pedagogy, classifying with regard to question type, mode of formulation, responses, and item format. This typology then provided a guide for further item creation and refinement.

Detailed feedback and improvement

The development stage includes a detailed student feedback component for item improvement. We want to know *why* students are choosing or not choosing *each* of the responses, and how they are interpreting the question. Items are administered to a representative group of pre-service students who give written reasons for choosing or not choosing each option. This can be further probed by interviews. Project 2061 [41] uses a similar procedure for developing science *content* items, and states that the detailed feedback proves essential to the item development and modification process.

External review

We have an external panel of eight expert consultants to review and critique both the Inquiry-Item-Criteria and the POSITT items as they are developed. The panelists are nationally recognized experts on science teaching, science teacher education, and/or assessment. Each panel member rates items for science content, pedagogy, and appropriateness of response choices. These data are used to establish the content validity of the items, the item's construct link to the Inquiry-Item-Criteria, and to establish the best answers, especially important in the RR and CR item formats. Besides its role in item improvement, this process also assists in verifying the relevance of the criteria themselves.

Work so far

This is a new project currently in the item development and refinement stage. We have formulated criteria for inquiry items, and the writing team has thus far generated several dozen teaching scenarios, and written one or more items for each of these, ranging over a variety of types. Our immediate experience is that the task of creating vignettes, questions, and response choices for inquiry PCK items is innovative, unfamiliar, and far more demanding than producing conventional items on science *content*. Thus, to assist in item writing we are simultaneously devising typologies and guidance for the task.

Using a limited set of early items, one of the investigators has done preliminary pilots of this assessment concept in an elementary science methods course at Western Michigan University. This tryout of individual items lends support to the viability of the concept. There was a concern that students could sometimes guess the 'desired' answer to items. Experience with these early items suggests that this will not be a significant problem. There was often a bimodal test distribution; some methods students with good understanding of inquiry could readily identify in problem-based teaching vignettes, while others seemed not to be oriented toward the inquiry approach in contextual situations, and could not be guessed. To a fair extent, students seemed to either 'get it' or not. At one end of the instructional spectrum, the idea that a teacher's role is 'telling' is quite deeply entrenched,

hence students with this orientation, conscious or unconscious, are likely to choose a corresponding response. At the other end, those with a very laissez–faire interpretation of inquiry will tend to choose 'discovery' responses. To study this it will be useful to survey and interview selected participants, to probe their reasons for choosing or not choosing the various options. This will also feed into the improvement of items and options.

Examples of items
Some examples of items produced are provided in the Appendix.

ITEM PILOTING, ANALYSIS, AND REVISION

Piloting with pre-service students at collaborating institutions

After the writing and reviewing processes, items will be compiled into sets of test forms. Pilot tests of the instrument will then be conducted at 10 collaborating institutions across the country. The participants will be undergraduate students in elementary science methods courses who intend to be teachers. With several participating universities, we have access to between 500 and 800 participants in each of two semesters.

Members of our project team and our colleagues at the other universities teach these courses and supervise adjunct teachers. We have access to the courses through our collaborators, which allows us to gather additional information about the participants. After students have taken the POSITT, their instructor will go over the items with them as 'worked examples', and students will be able to discuss the science teaching aspects and raise questions. This uses the occasion of the research pilot as an instructional opportunity for students and teachers, and can feed back into item refinement.

Note that that there will be two stages of piloting, in successive semesters. The statistical analysis of Pilot 1, plus information from the associated discussions, will inform item revision, so that the Pilot 2 can be run with a modified and improved instrument.

It is possible that science content knowledge and/or science literacy generally might correlate with pedagogical knowledge of science inquiry. Therefore, during the second pilot we will also administer a Scientific Literacy Survey [42], based on *Science for All Americans* [1]. This will be the first test of the null hypothesis that science knowledge is not related to pedagogical knowledge of inquiry teaching. Alternatively or in addition, once POSITT is finalized we will administer it to new groups together with a science content test specifically targeting the topics in the teaching scenarios, as well as a science literacy test.

Analysis of data from each round of pilot studies
Following each of the pilot tests at participating institutions, analysis will focus on the following:

- *Reviewing and finalizing directions for test scoring and administration:*
  1) *Rubrics for Constructed Response (CR) items*
  2) *Rankings and ratings for Ranked Response (RR) items*
  3) *Scoring rules for full and partial credit in CR and RR item formats*

- *Initial reliability estimation*

- *Initial construction of the Multi-Trait Multi-Method (MTMM) matrix for examination of the three different item-formats' ability to assess each of the Inquiry-Item-Criteria*

- *Item Response Theory (IRT) to calibrate items and estimate examinee pedagogical knowledge of inquiry science teaching (PCK-IST)*

- *Differential Item Functioning (DIF) for possible bias in gender, ethnicity, and science concentration*

- *Initial estimates of criterion-related validity by correlation between performance on the new assessment instrument and academic performance*

For the second pilot, in addition to the above analyses, student performance on the revised POSITT will be correlated with responses on the *Scientific Literacy Survey* to establish further construct validity.

An important validity component of an educational assessment instrument is to establish content-related validity. Since POSITT is composed of different item formats and assesses different dimensions of inquiry science teaching, the MTMM matrix is used to obtain construct validity evidence by gathering and analyzing the correlations among the item sets. The correlation matrix is examined against the expected correlation pattern to see if there is convergent and divergent validity evidence.

Item revision and improvement

After analyzing each pilot of items, revisions and improvements will be made. The best items across the range of grades and desired abilities will then be elected and assembled into final test format, a Pedagogy Of Science Inquiry Teaching Test. Variants may be produced for various grade ranges or specific purposes.

FIELD VALIDATION STUDIES

Although a multi-stage procedure will have been used to develop, pilot, and refine the items and POSITT instrument, it is also important to validate them against observation of subsequent teaching practice.

Construct-related Evidence and Predictive Validity Evidence

The validation studies collect two primary aspects of validity evidence for POSITT, namely construct-related evidence and predictive validity evidence. The multiple development stages and piloting of POSITT will provide content-related validity evidence. Although we are working with undergraduate science education majors, the ultimate goal is that they become competent teachers of K-8 school science. Since our goal is to assess pedagogical content knowledge, an important aspect of validity will be to establish a relationship between POSITT performance and subsequent teaching practice. Therefore, we will test the hypothesis that *level of inquiry pedagogical content knowledge* is a predictor of the degree to which teachers will be good inquiry teachers of science. Note that good Inquiry PCK should be a *necessary*, but not *sufficient*, condition for teaching well by inquiry. If a teacher does *not* have sufficient understanding of inquiry pedagogy, we hypothesize that it will be almost impossible to implement good inquiry practice in the classroom. As a result, a low score on the POSITT should be a clear predictor of poor inquiry teaching, and we can see if a minimum scoring level can be set as a necessary condition. On the other hand, a good score on POSITT cannot on its own be a *sufficient* condition for good inquiry practice. Knowing the principles of inquiry science teaching does not guarantee that such teaching will occur – there are other factors

in a school environment as well as personal attributes that might work against it in practice. Thus, to investigate the 'positive' hypothesis we can look for correlation between POSITT results and inquiry teaching practice, while identifying and examining other variables that are likely to affect practice. Examination of literature and theory regarding inquiry, plus discussion with our panel, suggests several possible intervening variables, namely teacher science knowledge, science teaching self confidence, science teaching efficacy, science interest and attitude, and school environment for science teaching. These are listed as covariates in the table below, along with instruments for assessing these aspects.

<table>
<tr><th>Independent Variable</th><th>Covariables</th><th>Instrument</th><th>Dependent Variable</th></tr>
<tr><td rowspan="6">SCORE ON PEDAGOGY OF SCIENCE INQUIRY TEACHING TEST (POSITT)</td><td>A. Science Knowledge</td><td>Scientific Literacy Survey [42]</td><td rowspan="6">FIELD OBSERVATION ASSESSMENT OF TEACHING PRACTICE</td></tr>
<tr><td>B. Science Teaching Self-Confidence</td><td rowspan="2">Self-confidence and Efficacy Instrument [43]</td></tr>
<tr><td>C. Science Teaching Efficacy</td></tr>
<tr><td>D. Science Interest/Attitudes</td><td>Scientific Attitude Inventory [44]</td></tr>
<tr><td>E. School Environment for Science Teaching</td><td>Local Systemic Change Initiative [45]</td></tr>
<tr><td>F. Motivation/attitude toward hands-on and inquiry-based teaching</td><td>Attitude Toward Hands-on & Inquiry-based Teaching [46]</td></tr>
</table>

The instruments for covariables A and D will be administered along with POSITT to all students forming the undergraduate pool. Covariables B, C, E, and F are sensitive to classroom situations and thus will be administered only to the sample group (below) once students are in the field as student teachers. The data for covariables will be collected prior to the field observations. In-service subjects will take POSITT and covariant tests prior to classroom observation. Note that participants will be interviewed as part of the field observations of practice, yielding further data.

Field validation testing pools and samples

In field validation studies, a finalized POSITT instrument will be administered to a new pool of undergraduate students (over 600 in all, over two semesters). From this pool a sample of about 60 pre-service participants will be randomly drawn, tests of covariates will be administered, and two blinded classroom observation studies subsequently carried out for each participant, by the Science and Mathematics Program Improvement (SAMPI) group at Western Michigan University. This will provide predictive validity evidence.

Thus far we have talked of using POSITT during undergraduate instruction of *pre-service* students, and checking predictive validity by observing their subsequent teaching. It is recognized however that there may be other pressures influencing teachers, student teachers in particular, so that lessons may not always match intentions. Therefore we will also do field validation with a sample of about 20 experienced *in-service* teachers, selected from districts across the Southwest Michigan and Detroit areas. They will take the POSITT and covariate tests and will twice be observed teaching in blinded field observations by SAMPI. By including an in-service

component, the study also reflects the idea that POSITT items can also be used as an evaluative and professional development tool for in-service teachers.

Field evaluations of teacher practice

The Science and Mathematics Improvement (SAMPI) unit will conduct the classroom evaluations of teaching. The SAMPI observation tool is of established validity and reliability for evaluation of science teaching practice [24,47,48]. SAMPI will arrange for, and conduct, two observations per student, including interviews and lesson plan evaluation. The teaching evaluations will be blind; however, SAMPI will have the Inquiry-Item-Criteria so that they can ensure alignment with their classroom observation protocol.

SAMPI will not have access to any other data collected from these students. Based on analyzed data from two teaching observations, along with lesson plans and pre- and post-lesson interviews, SAMPI will categorize teachers as 'good', 'satisfactory,' or 'poor' users of inquiry science practice. SAMPI personnel will not have had any previous contact with our project, and our project team will not know the sample. Hence, this will be a *double-blind evaluation* of the participant teachers.

Once the classroom observations have been rated they will serve as the criterion in subsequent analyses for establishing the criterion-related validity of POSITT; for example, logistic regression (LR) and discriminate function analysis (DFA) will be examined. DFA will establish the ability of POSITT to correctly classify teachers as good, satisfactory, or poor users of inquiry teaching, and logistic regression can provide useful estimates (odds ratio) of the impact of POSITT on inquiry teaching practice after accounting for covariables.

CONCLUSION AND POTENTIAL IMPACTS

At the conclusion of the project, a case-based objective instrument for assessing pedagogical content knowledge of inquiry science teaching will have been developed, piloted, and finally field-validated by blinded testing against the 'gold standard' for pedagogy assessment, namely what teachers actually do in the classroom.

At this stage POSITT will be made widely available. Thus, teacher educators at *any* undergraduate institution will be able to assess the effectiveness of their instruction in regard to pedagogical content knowledge of inquiry science teaching, and be in an informed position to improve their science teacher education programs. They will be able to use the instrument both formatively and summatively during teacher preparation or professional development.

Even though the instrument is primarily designed for pre-service K-8 teachers, items of this nature should have the potential for much broader uses, e.g. in-service or at secondary level. It could also be useful for graduate students and teaching assistants in science education programs. Interestingly, it might even be of interest to faculty teaching regular undergraduate and graduate science, technology, engineering and mathematics (STEM) subjects at colleges and universities. Previous research indicates that science professors can be motivated to improve their own teaching practice by becoming aware of how science is best taught to young people [49]. This could be of importance given the recent efforts to recruit doctoral level science majors into school science teaching, e.g. the NSF program for Graduate Teaching Fellows in K-12 Education [50].

Finally, POSITT, as a tested and validated instrument, will also be of value to researchers investigating ways to improve the education and practice of science teachers.

ACKNOWLEDGEMENT

The project is supported by National Science Foundation grant DUE-0512596.

REFERENCES

[1] American Association for the Advancement of Science (AAAS), *Project 2061*: *Science for all Americans*, (American Association for the Advancement of Science, Washington, D.C.), 1990.

[2] National Research Council, *National Science Education Standards*, (National Academy Press, Washington, D.C.), 1996.

[3] National Research Council, Inquiry and the *National Science Education Standards: A guide for teaching and learning*, (National Academy Press, Washington, D.C.), 2000.

[4] National Research Council, Classroom Assessment and the *National Science Education Standards*, (National Academy Press, Washington, D.C.), 2001.

[5] Kennedy, M. M., *Defining optimal knowledge for teaching science and mathematics, Report No. RM10*, (National Institute for Science Education, Madison, Wis.), 1997. http://www.wcer.wisc.edu/nise/Publications/Research_Monographs/KENNEDY/KennedyALL.pdf

[6] Anderson, R. D., *Inquiry in the Everyday World of Schools* [Web Page], ENC Online, 2003. http://www.enc.org/topics/inquiry/context/document.shtm?input=FOC-000708-index.

[7] DeBoer, G. E., *A history of ideas in science education: Implications for practice*, (Teachers College Press, New York), 1991.

[8] Krajcik, J., Mamlok, R., and Hug, B., "Modern content and the enterprise of science: Science education in the twentieth century," in Corno, L., (ed.), *Education Across A Century: The Centennial Volume*, (University of Chicago Press, Chicago), 2001.

[9] Rudolph, J. L., *Scientists in the classroom: the cold war reconstruction of American science education,* (Palgave, New York), 2002.

[10] Schwab, J. J., *The teaching of science as enquiry: The Inglis Lectureship*, (Harvard University Press, Cambridge, Mass.), 1962.

[11] Shymansky, J. A., Kyle, W., and Alport, J., "The effects of new science curricula on student performance, *Journal of Research in Science Teaching*, vol. 20, no. 5, pp. 387-404, 1983.

[12] Shymansky, J. A., Hedges, L., and Woodworth, G., A reassessment of the effects of inquiry-based science curricula of the '60s on student performance, *Journal of Research in Science Teaching,* vol. 27, no. 2, pp. 127-144, 1990.

[13] Secker, C. V., and Lissitz, R. W., "Estimating the impact of instructional practices on student achievement in science," *Journal of Research in Science Teaching*, vol. 36, no. 10, pp. 1110-1126, 1999.

[14] Secker, C. V., "Effects of inquiry-based teacher practices on science excellence and equity," *Journal of Educational Research,* vol. 95, no. 3, pp. 151-160, 2002.

[15] Karplus, R., "Science teaching and the development of reasoning," *Journal of Research in Science Teaching,* vol. 14, no. 2, vol. 169-175, 1977.

[16] Shulman, L. S., "Those who understand: A conception of teacher knowledge," *American Educator,* vol. 10, no. 1, pp. 9-15, 1986.

[17] Shulman, L. S., "Knowledge and teaching: Foundations of the new reform," *Harvard Educational Review*, vol. 57, no. 1, pp. 1-22, 1987.

[18] Barnett, J., and Hodson, D., "Pedagogical context knowledge: Toward a fuller understanding of what good science teachers know," *Science Education*, vol. 85, no. 4, pp. 426-453, 2001.

[19] Dean, C. D., "Problem-based learning in teacher education," Annual Meeting of the American Educational Research Association, Montreal, 1999.

[20] Eick, C. J., "Inquiry, nature of science, and evolution: The need for a more complex pedagogical content knowledge in science teaching, *Electronic Journal of Science Education*, vol. 4, no. 3, 2000. Available at: http://unr.edu/homepage/crowther/ejse/ejsev4n3.html

[21] Lederman, N. G., "Do subject matter knowledge, pedagogical knowledge, and pedagogical content knowledge constitute the ideal gas law of science teaching?" *Journal of Science Teacher Education*, vol. 3, no. 1, pp. 16-20, 1992.

[22] Lowery, L. F., "The nature of inquiry," *Science Link*, vol. 13, no. 1, p. 3, 2002.

[23] Luft, J. A., "Assessing science teachers as they implement inquiry lessons: the extended inquiry observational rubric," *Science Educator*, vol. 8, no. 1, no. 9-18, 1999.

[24] Jenness, M., and Barley, Z. A., *Observing teaching practices in K-12 classrooms: Instruments and methods (Science - version B: for multiple classroom observations to improve programming)*, (Science and Mathematics Program Improvement (SAMPI), Kalamazoo, Mich.), 1999.

[25] Harwood, W.S., Hansen, J., Lotter, C., "Measuring teacher beliefs about inquiry: A blended qualitative/quantitative instrument," *Journal of Science Education & Technology*, vol. 15, no. 1, pp. 69-79, 2006.

[26] Black, P., Harrison, C., Lee, C.S., Marshall, B., and Wiliam, D., *Assessment for Learning - Putting it into practice,* (Open University Press, Maidenhead, UK), 2004.

[27] U.S. Department of Education., *Scientifically Based Research—U.S. Department of Education,* (U.S. Department of Education, Washington, D.C.), 2002.

[28] Livingston, A., (ed.), *The condition of education 2002*, (National Center for Education Statistics, Washington, D.C.), p. iii, 2002.

[29] Lesh, R. A., and Lovitts, B., "Research agendas: Identifying priority problems and developing useful theoretical perspectives," in Kelly, A.E., and Lesh, R.A. (eds.), *Handbook of research and design in mathematics and science education,* (Lawrence Erlbaum Associates, Mahwah, N.J.), 2000.

[30] Shavelson, R. J., and Towne, L., *Scientific research in education*, (National Academy Press, Washington, D.C.), 2002.

[31] Towne, L., and Hilton, M., *Implementing Randomized Field Trials in Education* Report of a Workshop, Committee on Research in Education, Center for Education Division of Behavioral and Social Sciences and Education, National Research Council, (The National Academy Press, Washington, D.C.), 2004.

[32] Albanese, M., and Mitchell, S., "Problem based learning: A review of literature on its outcomes and implementation issues," *Acad Med*, vol. 68, no. 1, pp. 52-81, 1993.

[33] Peterson, M., "Skills to enhance problem-based learning," *Med Educ Online*, vol. 2, no. 3, 1997. http://www.med-ed-online.org/f0000009.htm.

[34] Ngeow, K., and Kong, Y.S., *Learning to learn: Preparing teachers and students for problem-based learning,* ERIC, ED457524, 1994.

[35] Wang, H. C. A., Thompson, P., Shuler, C., and Harvey, L., "Problem-based learning approach for science teachers' professional development, Annual Meeting of the Association for the Education of Teachers in Science, Austin, Tex., 1999.

[36] Anderson, L. W., and Krathwohl, D. R., *A taxonomy for learning, teaching, and assessing: a revision of Bloom's Taxonomy of educational objective*s, (Longman, New York), 2001.

[37] Scriven, M., *Evaluation thesaurus*, (SAGE Publications, Newbury Park), 1991.

[38] Scriven, M., "The new science of evaluation," *Scandinavian Journal of Social Welfare,* vol. 7, no. 2, pp. 79-87, 1998.

[39] Campbell, D. T., and Fiske, D. W., "Convergent and discriminant validation by the multitrait-multimethod matrix," *Psychological Bulletin,* vol. 56, no. 2, pp. 81-105, 1959.

[40] Gulek, C., "Using Multiple Means of Inquiry To Gain Insight into Classrooms: A Multi-Trait Multi-Method Approach," Annual meeting of the American Educational Research Association, Montreal, Quebec, Canada, 1999.

[41] American Association for the Advancement of Science (AAAS), "Getting Assessment Right — Rigorous Criteria, Student Data Inform Project 2061's Assessment Design," *2061 Today*, vol. 17, no. 1, pp. 1, 4, 5, winter 2007.

[42] Laugksch, R. C., and Spargo, P. E., "Development of a pool of scientific literacy test-items based on selected AAAS literacy goals," *Science Education,* vol. 80, no. 2, pp. 121-143, 2000.

[43] Enochs, L. G., and Riggs, I. M., "Further development of an elementary science teaching efficacy belief instrument: a pre-service elementary scale," *School Science and Mathematics*, vol. 90, no. 8, pp. 694-706, 1990.

[44] Moore R. W., and Foy, R.L.H., "The scientific attitude inventory: A revision (SAI II)," *Journal of Research in Science Teaching*, vol. 34, no. 4, pp. 327-336, 1997.

[45] Horizon, Inc., 2003-04 *Data Collection Manual*, 2003. http://www.horizon-research.com/LSC/manual/

[46] Clettenberg, S., Johnson, D., Miller, C., and Rideout, K., *Attitude Toward Hands-on and Inquiry-based Teaching,* The University of Tulsa, School of Education, (Oklahoma Teacher Education Collaborative, Tulsa, Okla.), 2001.

[47] Barley, Z. A., and Jenness, M., "The Role of Evaluation in Systemic Change in Education," Annual Meeting of the American Educational Research Association, New Orleans, La., 1994.

[48] Jenness, M., *Michigan teacher perceptions of their classroom practice and preparation in science and mathematics: a report of findings from a survey of Michigan K-12 teachers receiving professional development services fully or partially funded by the Michigan Eisenhower Professional Development Program,* (Science and Mathematics Program Improvement (SAMPI), Kalamazoo, Mich.), 2001.

[49] Fedock, P. M., Zambo, R., and Cobern, W. W., "The professional development of college science professors as science teacher educators," *Science Education*, vol. 80, no. 1, pp. 5-19, 1996.

[50] Vines, V., "National program needed to draw postdoctoral scholars into K-12 education careers," 2002. http://www8.nationalacademies.org/onpinews/newsitem.aspx?RecordID=10129

APPENDIX: EXAMPLES OF ASSESSMENT ITEMS

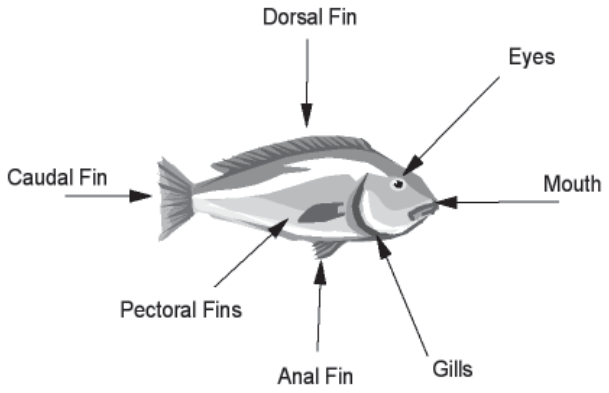
Each item begins with a realistic classroom teaching vignette on a particular science topic. This is turned into a problem by asking a question about pedagogy, with a set of alternative responses to choose from. There are various possible types of items, for example an evaluation of the lesson so far, suggestions for what the teacher should do next, alternative lesson designs and approaches, ways of handling questions or occurrences, etc.

Example 1. Starting to teach about form and function

FISH

Mr. Lowe is a 3rd grade teacher. Two of his eventual objectives are for students to learn at a simple level about the relationship between form and function.

He begins a specific lesson on fish by showing an overhead transparency of a fish, naming several parts, and labeling them as shown.



Which of the following is the best evaluation of the lesson so far?

A. This is a good lesson so far, because the teacher is clearly and systematically introducing the vocabulary that the children will need for further studies of fish.

B. This is a good lesson so far, because by learning the names of the fish parts, the students are more engaged and will ask appropriate questions about their function.

C. This lesson is not off to a good start, because it begins with the teacher giving the children information about fish, before any attempt to develop a sense of questioning or investigation on the part of the students.

D. The lesson is not off to a good start, simply because it begins with the teacher doing the talking, which is never a good idea.

E. This lesson is not off to a good start, because the students are not doing anything "hands-on." There should always be real fish for students to observe, so they would connect the lesson to the real world.

COMMENTS ON EXAMPLE ITEM 1

Of the options, "C" is the desired response according to the inquiry pedagogy criteria. "C" suggests the teacher should engage students through questioning about what they notice and know about fish. The teacher should guide students to describe the various fish parts and ask students to pose questions about what the parts do for the fish. As it is described, the lesson does not necessarily engage the students' thinking.

Responses "A" and "B" align only with knowledge-level objectives, whereby the intent is for students to know the vocabulary. Knowing formal names of body parts is not a necessary criterion for associating form and function and the approach is not inquiry.

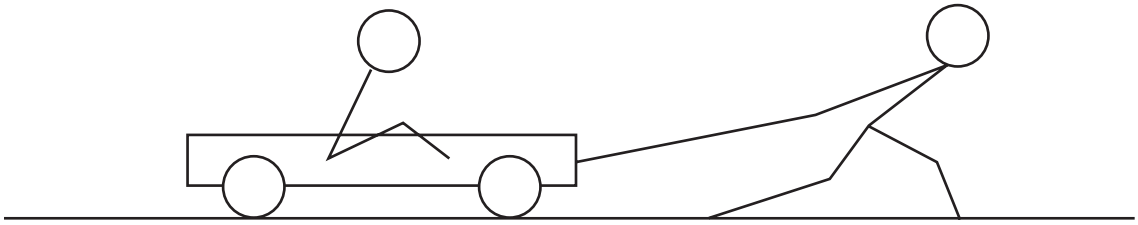
Response "D" suggests that a good inquiry lesson is never teacher-centered. A teacher-centered portion of a lesson can be inquiry-oriented by engaging students through modeling of investigative activities. Teachers can provide students with questions, data, and explanations; all the while discussing the reasoning processes that lead to justification for claims.

Response "E" is not the best because it suggests a good lesson must always be "hands-on." Hands-on does not ensure inquiry, nor does it ensure students will connect the lesson to the real world. The teacher could engage students through questioning and other scenarios familiar to students, without needing to have students observe real fish – though this would be ideal.

Example 2. Teaching approaches for force and motion

FORCE & MOTION

A useful activity for teaching force and motion is to have one student sit in a trolley with little friction while another can pull it along by exerting force on the handle.



The goal is that students gain a conceptual understanding of the relationship between motion and force, viz. that an applied force will cause an object to change its motion, i.e. speed up or slow down. (Newton's second law).

Five teachers have five different lesson plans for using this activity to teach the relationship between force and motion. Which plan below would be best?

A. Mr. Adams starts by writing a heading on the board: 'Newton's Second Law of Motion', and dictates the law (in conceptual terms) for students to write down. He then explains the law and illustrates it with a diagram of a trolley being pulled. At any stage he gives students the opportunity to ask questions. Finally he has students verify the law experimentally by checking what happens to a trolley when a person pulls it with a constant force.

B. Ms. Burke first has students explore what happens to the trolley when a steady force is applied to it, and asks them to describe the kind of motion that results. She elicits the focus question of how force and motion might be related, then asks for suggestions for a 'law' that would describe their observations. Having put forward a proposed law (or laws), students then test it by making predictions in various situations and trying out. They finally write their own statements of the law they have generated.

C. Mr. Campos gives students freedom to try out anything they wish with the trolleys, intending that they should be drawn in to the hands-on activity and discover on their own the relation between force and motion. He does not impose structure or tell students what to do, but is available for discussion, in which he does not give 'answers' to questions but instead asks questions in return. At the end of the session he does not provide the 'correct' law, since the point is for students to discover their own.

D. Ms. Davis, as a prelude to Newton's second law of motion, defines the term acceleration and has students write it down. She then explains the concept carefully with examples. Thereafter she presents Newton's second law in the form 'acceleration is proportional to net force'. Students then verify the law by doing the hands-on trolley activity.

E. Mr. Estrada feels that the textbook treats force and motion clearly and correctly. Thus he has several students in succession read paragraphs aloud from the book, and encourages students to ask if they don't understand something. He then demonstrates the law for the whole class with the trolley activity and two students assisting, to verify the textbook statement.

Note that this item may be most suited to formative use because of the length of its options.

COMMENTS ON EXAMPLE ITEM 2

Only options B and C represent inquiry approaches, but C is essentially unguided discovery. B addresses all of our inquiry pedagogy criteria, while the unstructured nature of option C makes it hard to know which criteria might be attained in a class. The other options A, D, and E present the conclusions of science first, then explain and confirm them, the antithesis of inquiry and investigation.

A. This approach is completely non-inquiry, though organized and methodical. The lesson is a rhetoric of 'conclusions first', to paraphrase Schwab. Experiments are seen as confirmatory not investigative.

B. A good inquiry approach, generating questions, ideas, and concepts from exploration. Students propose a possible law from evidence and test it. Guided inquiry and investigation, appropriately structured, as advocated by standards.

C. Unstructured and unguided discovery for the most part. It is unlikely that students will be able to make sense of the activities or reach the desired learning outcomes. Pure discovery is not advocated.

D. Presents conclusions first, again the antithesis of inquiry. Moreover, difficult concepts (acceleration) are introduced and formally defined in a way that is unnecessary at this level and will likely interfere at this stage with developing the desired conceptual understanding.

E. This is a dreary passive class activity, though the teacher may be seeking to avoid 'teacher talking' to some extent. Approach is non-inquiry, little engaged. Experiments seen as confirming book knowledge rather than generating knowledge.

EXAMPLE 3. Anomalous results in a classroom investigation on earthworms

Earthworm investigation
Ms. Lefevre's third grade class has been doing a long investigation activity with earthworms. Besides teaching her students about the basic needs of earthworms, Ms. Lefevre also wants to develop their skills of observing, investigating, recording, and seeking patterns.

Several groups had been making observations and taking data over some time, and she brought the class together around the data chart, so that they could all look for patterns in their observations. She wanted her students to rely on evidence to develop knowledge. During this analysis, a student pointed out that data collected by one group seemed to contradict that of another group.

What should Ms. Lefevre do in this situation?

A. Tell the students which of the two sets of data is correct and cross out the other data, so that none of the students get wrong ideas about earthworms.

B. Ask the students to suggest ways to resolve the issue, valuing any response that relied on evidence, e.g. re-examining recorded data or comparing procedures, repeating, or taking more observations.

C. Ask everyone to look at the two data sets and to pick the one they think is right. Then have a show-of-hands vote to see which one should stay and which should be crossed off. This would ensure that the data that remained reflected the majority view.

D. Tell the students that since there was conflicting data and it wasn't clear which was right, she would look it up and get back to them the next time. Then move on to look at other aspects of the observations.

E. Ask the students to read through the topic resources again to see if they can find information that will resolve the dispute.

COMMENTS ON EXAMPLE ITEM 3

The desired response is B. This response most closely mirrors what scientists do when variations occur in data. They first recheck and rethink their observations, looking for sources of error. Then they often make new observations under more closely prescribed conditions. In this way, they hope to gather enough data to see clear patterns.

Items A, D, and E essentially sideline the classroom inquiry to refer to an outside source, a poor choice when evidence or procedure is available to resolve the dispute.

Item C involves voting, which discounts certain data based on reasons other than the data itself. In science inquiry, all data is important initially, and data can only be discounted when error in procedure, observation or recording can be identified. Otherwise the data counts, even if it seems not to fit or illustrate a clear pattern.

Calibrated Peer Review™ in Math and Science: Research and Practice

Nancy Simpson* - *Texas A&M University*
Arlene Russell - *University of California, Los Angeles*
Adalet Baris Gunersel and Jean Layne - *Texas A&M University*

ABSTRACT

Questions related to assessment of student learning, faculty development, and writing-intensive courses were the catalyst for the NSF-funded project *Writing for Assessment and Learning in the Natural and Mathematical Sciences (WALS)*. The project has included the use of Calibrated Peer Review, an online tool that provides instruction in both writing and critical review. This paper presents project activities, studies and findings, and plans for future work. Project activities included consisted of faculty workshops and retreats. Studies focused on faculty's educational beliefs and practices, CPR effectiveness and student outcomes, and student perceptions of CPR. Among the products of WALS are the 94 CPR assignments created by faculty participants and a national CPR Symposium held at Texas A&M University in June 2007.

BACKGROUND

Three sets of questions merged to become the catalyst for *Writing for Assessment and Learning in the Natural and Mathematical Sciences (WALS),* a project funded by NSF in May 2003. The first set of questions was related to assessment of student learning. If we want to develop courses, teaching methods, and learning tools that promote conceptual understanding, critical thinking, and communication skills, how will we know if these courses, methods, and tools work? How can we generate artifacts of student learning that give us clues to the way they do or do not understand the content we are trying to teach?

The second set of questions was related to faculty development. What kinds of opportunities and experiences lead to real and sustained change in the way faculty design and conduct their classes? What will inspire faculty to engage in scholarly inquiry into their students' learning?

The third set of questions was more pragmatic in nature and had to do with a new university requirement for writing-intensive courses in the major. What support do faculty need for creating and delivering writing-intensive courses? How can we recruit, train, and pay for teaching assistants to assist with the grading?

The catalyst for pulling these three strands of thought together was the introduction of Calibrated Peer ReviewTM (CPR) to our campus. CPR (http://cpr.molsci.ucla.edu) is an online tool that provides instruction in both writing and critical review. Because the program is web-delivered, computer-managed, and uses peer review, instructors can use CPR with any class size; students can do their assignments wherever they have web access; and administrators do not need to provide additional teaching resources for grading. These factors made CPR at least a partial answer to the need for supporting faculty as they developed and implemented writing-intensive courses. A belief that CPR would also help us address our assessment and faculty development questions led us to propose—and then implement—WALS.

Because CPR is a relatively new tool, this paper begins with a brief description of CPR in order to provide context for the reader. The remainder of the paper describes project activities and products, illustrates some of the research findings, and discusses plans for future work.

CALIBRATED PEER REVIEW

All phases of a CPR assignment are computer managed. The instructor creates the component parts ahead of time, determines the start and stop time of each phase, and determines the weight that will be given to each phase in the computation of the student's grade.

A CPR assignment consists of a text entry phase, a calibration phase, and a review phase. In the text entry phase, students log on to CPR, read the assignment instructions and guiding questions, write the assignment, and enter their text into the CPR program. In the calibration phase, students review three sample responses to the assignment using an instructor-designed rubric. Their reviews are compared to the instructors' reviews of the assignments and they receive feedback on their review. Through this process, students learn what criteria to use in reviewing their peers' work. In the review phase, students review the work of three of their peers, and then their own text, using the same instructor-designed rubric that was used during the calibration phase.

PROJECT ACTIVITIES

Workshops

For faculty, the most time-consuming aspect of CPR is the authoring of the assignments. Each assignment requires that the instructor:

1) Determine learning objectives for the assignment
2) Select resource material for the student to use in completing the assignment
3) Write instructions for the assignment, including guiding questions and a writing prompt
4) Write three sample essays in response to the assignment
5) Create a rubric in the form of "calibration questions"
6) Create an answer key by responding to each of the calibration questions for each of the three essays

Since CPR was developed under the Molecular Science Project, a UCLA-based NSF chemistry initiative, most of the existing assignments in 2003 were for chemistry. To encourage the authoring of assignments in other science disciplines, we held summer workshops for faculty from physics, biology, and mathematics at Texas A&M University, Texas A&M Corpus Christi, Texas A&M Kingsville, and Texas Woman's University. During these workshops, most participants were able to complete a draft of at least one assignment, and developed the skill and understanding to create others on their own.

Faculty Retreats

In order to promote faculty inquiry into their students' learning we created opportunities for the faculty in our project to come together to share stories—successes, challenges, and frustrations. These opportunities ranged from short, lunch-hour discussions to an all-day retreat. We discussed sections from *How People Learn* [1] and generated lists of potential research questions related to student learning for which CPR might provide data. During the all-day retreat, we introduced Anderson and Krathwohl's revision of Bloom's taxonomy [2] (the

knowledge dimension) as a framework for analysis of student work. Faculty worked in disciplinary groups to address the question, "In order to complete this assignment successfully, what factual, conceptual, procedural and/or metacognitive knowledge do they need? " They then examined samples of student work to find evidence that students either had or did not have this knowledge.

RESEARCH

This section provides information about three studies we have conducted with faculty participants. The first study investigated faculty's educational beliefs and practices; the second study investigated CPR effectiveness and student outcomes with repeated use of CPR; and the third study investigated student perceptions of CPR. While thorough discussion of methods and results of these studies is beyond the scope of this paper, we present brief summaries of each below.

Faculty Educational Beliefs and Practices

To gain an understanding of our faculty participants' educational beliefs and practices, we conducted stimulated-recall interviews. This method involved videotaping a class session and using segments of the videotape to stimulate reflection. Interview questions included, What were your hoped-for learning outcomes for this class? Why did you choose that particular example? What were your thoughts when that student asked that question? What do you think the students were thinking during this segment? and Do you think your students understood that concept? We transcribed the interviews and analyzed the resulting data, looking for patterns and themes in responses.

In general, faculty responses to the first question centered on content coverage. Faculty participants were certain about the importance of their content, but many had difficulty articulating what they wanted students to be able to do with that content. They also expressed uncertainty about whether or not students were "getting it" and in some cases puzzlement about why students struggled. The faculty varied in their assessment of CPR as a tool to help them with these questions. The complete results of this study are reported elsewhere [3] and provided baseline information about faculty teaching beliefs and practices. As we approach the end of the project, we are conducting a second set of interviews to learn what has changed over the three years of the project.

Student Outcomes

Among the many research questions raised by faculty participating in this project, a prominent one pertains to the effectiveness of CPR. This issue is complex because the variables are many. The nature of the assignments, the level of the students, and the strategies for implementation all affect student outcomes. We began to investigate CPR's effectiveness by looking at one professor's results at a time. A summary of a study involving one biology professor and his students is given below:

The instructor gave the same four assignments to a total of 83 students over two semesters. The assignments called for the students to write an abstract based on one or more scientific articles. The concepts covered increased in difficulty from assignment #1 to assignment #4, and assignment #4 involved more than one article. We separated the students into three groups based on their performance on the first assignment. Figures 1 and 2 show the progress of the three groups as measured by 1) text rating (an indication of content understanding and writing quality), and 2) reviewer competency index (RCI) (an indication of their ability to evaluate peers' work based on the instructor-designed rubric). As the graphs indicate, there was a general trend of improvement for all students

and the difference among low, medium, and high groups decreased for both text rating and RCI. The instructor attributed the drop in RCI on the fourth assignment to the difficulty of that assignment. These results are reported in Gunersel, Simpson, Aufderheide, and Wang [4], and mirror those reported by Gerdeman, Russell, and Worden [5].

Figure 1. Group progress as measered by text rating

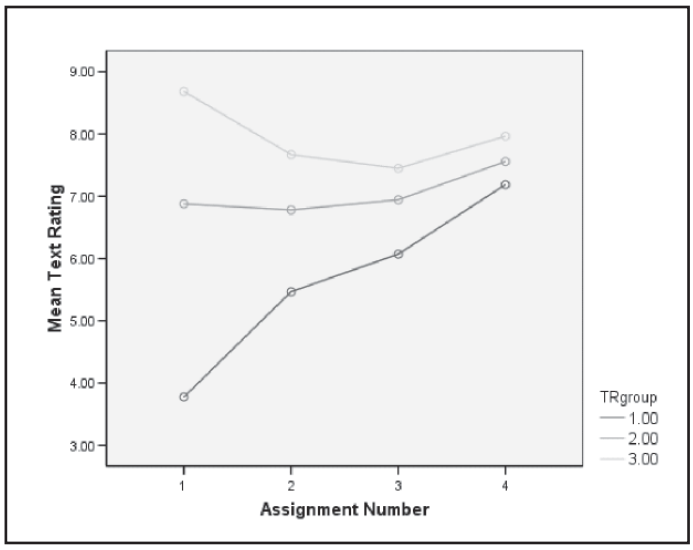
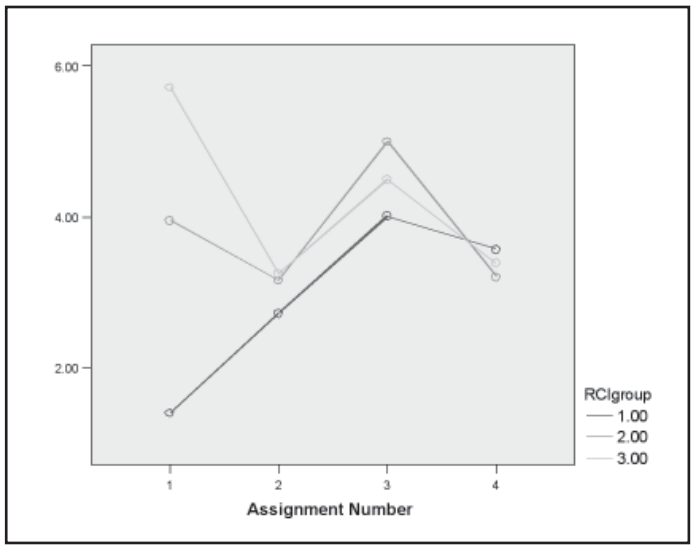


Figure 2. Group progress as measured by reviewer competency index (RCI)



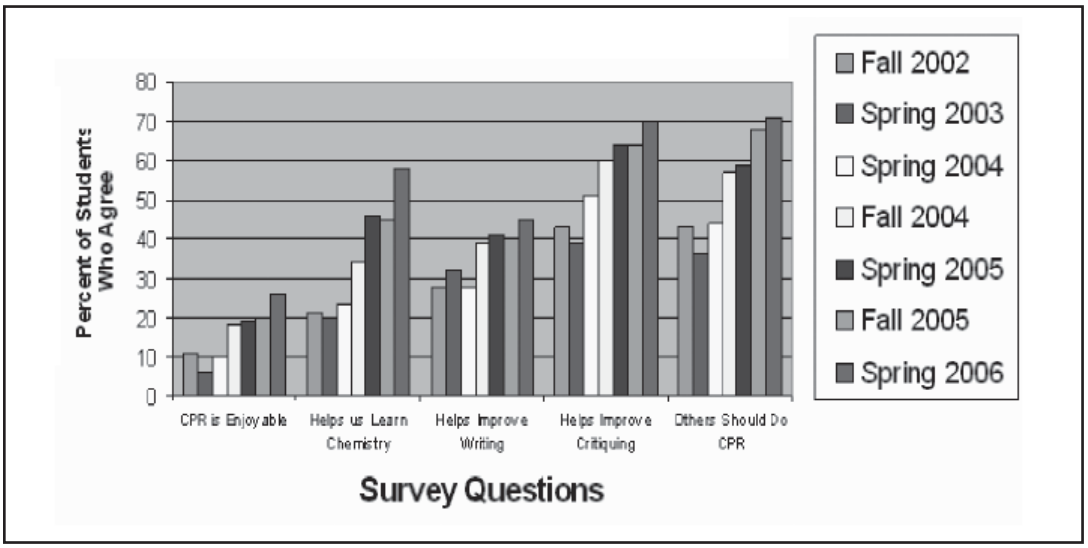
Student Perceptions of CPR

Another study focused on student perceptions of CPR. A chemistry instructor used Student Assessment of Learning Gains (SALG), a web-based program for collecting students' perceptions of their learning in a particular course. For seven semesters, she asked students to agree or disagree with five statements and to explain their response:

1) CPR is enjoyable
2) CPR helps me learn chemistry
3) CPR helps improve writing
4) CPR helps improve critiquing skill
5) Future students should use CPR

Each semester, the instructor used the SALG data to modify her implementation of CPR. As Figure 3 indicates, students became more positive in their perceptions of CPR as the instructor became more knowledgeable and skilled in the way she introduced and motivated the use of CPR. Our study, reported in Keeney-Kennicutt, Gunersel, and Simpson [6], focused on the nature of student initial resistance to CPR and on changes in implementation that led to a more positive response. We discovered two primary sources of student resistance: a belief that writing and peer review do not belong in a chemistry class and a distrust of peers' ability to fairly critique. We learned that successful implementation of CPR included 1) explicit explanation of the importance to their future career of competence in writing and evaluating and 2) frequent assurance of instructor availability and willingness to support the peer review process.

Figure 3. Student Assessment of Learning Gains (SALG) results on how students view CPR.



PRODUCTS, LESSONS LEARNED, AND FUTURE WORK

To date, our project participants have authored a total of 94 assignments (29 in Biology, 30 in Mathematics, and 35 in Physics). Early in the project, one of the departments at TAMU applied a strict interpretation of Family Educational Rights and Privacy Act (FERPA) regulations to CPR use and did not want student CPR data outside the university firewall. Because UCLA and TAMU were collaborators on this grant, TAMU was able to obtain a copy of CPR on its own server, allowing the project to continue uninterrupted. While this has worked well for TAMU in many ways, a negative consequence has been that TAMU faculty do not have access to the hundreds of CPR assignments on the UCLA server and that the 94 TAMU faculty-generated assignments are not available to CPR users. Work is underway, supported by another NSF grant, to separate the student portion and the faculty authoring portion of CPR, thereby alleviating this problem. In the meantime, we will be making annotated CPR assignments from all of our faculty participants available via the KEEP Toolkit hosted by Carnegie's Knowledge Media Lab (http://www.cfkeep.org/static/index.html). One area for future work involves identifying characteristics of effective assignments. As a beginning point, we have drafted a rubric that may be used to guide the development and evaluation of CPR assignments. This rubric is shown in Appendix 1.

We have learned, both through the study described above and through our work with other faculty participants, that successful implementation of CPR requires explicit attention to motivation and potential sources of resistance. The faculty who have persisted in using CPR are those who are convinced of its usefulness to student learning. They are undeterred by student resistance but rather find ways to help their students see that CPR helps them learn concepts and develop competence in writing and reviewing.

At the beginning of the project, we expected that our participants would create and implement CPR assignments, and plan and carry out research projects, all within their first year of participation. In reality, the creation and implementation of assignments proved sufficient challenge for more than one year, and keeping faculty engaged in the research part after their year of participation proved difficult. However, among the many strengths of the CPR program is that all of the data—both quantitative and qualitative—are stored, ready for

analysis when the researcher is ready. For example, one of our mathematics professors wrote many assignments that asked students to write mathematical proofs. While the text ratings, reviewer competence indices, and other scores were used immediately as a contribution to course grades, the student writing itself may be analyzed at any time to gain increased understanding of why and in what ways mathematics students struggle with proof writing.

The stimulated-recall interviews gave us insight into faculty beliefs and practices at the outset of the project, but there are additional questions: Have there been any changes in faculty beliefs and practices since the beginning of the project? If so, to what might these changes be attributed? Are there identifiable differences between those who reported success with CPR and those who did not? Are there differences in kinds of assignments authored and methods of implementation? What, if anything, do these differences tell us about future curricular or teaching development efforts? In order to answer these questions, we are conducting semi-structured interviews with faculty participants. We will continue to work individually with the faculty who have participated in this project, both to improve implementation of CPR and to use the data generated to answer their questions about their students' learning.

The CPR Symposium

The culminating activity of WALS was a CPR Symposium, June 17–19, 2007 (http://cte.tamu.edu/cpr/cpr.html). The Symposium was open to all who are interested in CPR and was organized around three themes:

Effective CPR Assignments. What characterizes a CPR assignment that works? What have been your most successful assignments? What kinds of learning do your assignments seek to facilitate? How have your assignments evolved as a result of your experience? What challenges have you encountered in the creation of CPR assignments?

Successful Implementation of CPR. What strategies do you use to implement CPR in your courses? What challenges have you encountered and how have you overcome them? How do you help students accept their peers' review of their work as legitimate? How do you help students learn to use the calibration questions to help them recognize quality work?

CPR Research. What research have you conducted regarding effectiveness of CPR for improving writing, critical thinking, and/or conceptual understanding? What research have you conducted regarding student acceptance of CPR? What are your methods? What are your findings?

While the Symposium will mark the official conclusion of the WALS project, the work—teaching, faculty development, and research on learning—will continue.

REFERENCES

[1] Bransford, J. D., Donovan, M. S., and Pellegrino, J. W., *How people learn: Bridging research and practice*, (National Academy Press, Washington, D.C.), 2002.

[2] Anderson, L. W., and Krathwohl, D., *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives* (abridged ed.), (Addison Wesley Longman, Inc., New York), 2001.

[3] Simpson, N., Layne, J., Gunersel, A. B., Godkin, B., and Froyd, J., "Faculty development through student learning initiatives: Lessons learned, in Robertson, D.R., and Nilson, L.B., (eds.), To Improve the Academy, Vol. 25, pp. 109-122, (Anker Pub. Co., Inc, Bolton, Mass.), 2007.

[4] Gunersel, A. B., Simpson, N. J., Aufderheide, K., and Wang, L., "Effectiveness of Calibrated Peer Review™ for improving writing and critical thinking skills in biology undergraduate students," under review.

[5] Gerdeman, R. D., Russell, A. R., and Worden, K. J. (2007), "Web-based student writing and reviewing in a large biology lecture course," *Journal of College Science Teaching*, March/ April, pp. 46-52, 2007.

[6] Keeney-Kennicutt, W. L., Gunersel, A. B., and Simpson, N. J., "Overcoming student resistance to a teaching innovation" under review.

APPENDIX 1

CPR Assignment Assessment Rubric

Components	Excellent	Acceptable	Needs additional work
Learning Goals**	Goals are clearly written using action verbs that convey to students what they will learn (know and be able to do) by successfully completing the assignment.	Goals read like instructions for completion of the assignment and few action verbs are used to articulate what students will learn (know and be able to do) by successfully completing the assignment.	Gives overview of what students will write but not learning goals for the assignment.
Source Materials/ Resources*	Sources are readily available to students. All links work properly. The number of sources is appropriate. Descriptions of source materials make it clear what they are and why they are included with the assignment.	All source materials are readily available. Source material descriptions indicate what the sources are but do not include why they are required.	Documents are not in an easily accessible format. Links are broken or missing. Descriptions of source materials do not make clear what they are or why they have been selected for the assignment.
Instructions**	Provide a roadmap for completion of the assignment. For example: describing the context of the problem, directing the order in which to view source material, reminding students of class discussion pertaining to assignment, etc.	Leave the students uncertainty about one or more aspects of the assignment. For example: lack of context, no direction about how to examine source material, vague performance expectations.	Instructions are non-specific, leaving the students to make assumptions regarding the context of the assignment and how to complete it.
Guiding Questions**	Call student attention to significant points to address and questions to think about while writing and serve as a self-assessment guide for students after they have written.	Serve as a template that tells students what to write about, but does not promote reflection for writing.	Do not provide students with guidance for responding to the assignment or are not present.
Assignment Title*	Notes discipline. Gives students a good idea of the topic and purpose of the assignment.	Notes discipline. Gives students an idea of the topic of the assignment, but not of the purpose.	Does not note discipline. Is disconnected from the assignment or misleading about the assignment.

Text Entry Directions* (Writing Prompt)	Concisely describes the students' role, their audience, the format, and the topic of the assignment.	Lengthy and either role, audience, format, or topic is missing.	Extremely lengthy and two or more of the following components are missing – role, audience, format, topic.
Calibration Essays*	The three essays are of distinct levels of quality. Medium and low quality essays include important learning mistakes.	Two essays are not distinct from each other in quality. Medium and low quality essays include some of the important learning mistakes.	Essays do not differ significantly and emphasize cosmetic errors over content specific mistakes.
Calibration Questions*	Include both content and style questions. Include enough information that students not already informed about the topic will be able to competently review.	Include both content and style questions. At least half of the calibration questions have enough information to promote competent review by students not already knowledgeable about the topic.	Most of the questions fail to include enough information to support competent review by students not already knowledgeable about the topic.
Calibration Answer Key/ Quality of Feedback*	Most of the answer keys provide feedback so that students know the "what" and "why."	Half of the answer keys provide feedback.	No feedback is provided.
Assignment Analysis Summary - Overall Alignment	All components: goals, guiding questions, writing prompt, exemplary essay, and calibration questions are very well aligned.	Only one of these elements is out of alignment with the others.	Two or more elements are out of alignment.

*Essential CPR Assignment Component required by the CPR software.

**Not an essential CPR component – but strongly recommended.

Developing Assessments That Help Students Learn to Develop Open-ended Designs

Judith E. Sims-Knight*, Richard Upchurch, Paul Fortier - *University of Massachusetts Dartmouth*

OVERVIEW

This project developed a kit of tools to assess students' understanding of and ability to use design process. This paper will describe three of them—the Design Process Knowledge Test, a concept map construction task, and a design process simulation task. These tools were designed to 1) be easy to administer in a course, 2) provide pointers toward improvement for both instructors and students, and 3) provide a summary score for instructors who wish to grade them.

THE PROBLEM

Project Kaleidoscope's (PKAL) Report on Reports II [1] calls for educational changes that help students become innovative practitioners and lifelong learners. Although research has demonstrated that senior undergraduates are better than freshmen in components of analysis (e.g., [2,3]), their design skills are still inadequate (e.g., [4,5,6]). For example, Eckerdal et al. [4] studied the detailed designs of 150 senior computer science students from 21 institutions in four countries and found that 62 percent of the so-called designs had no design information, and only 3percent of the designs could be called complete. Furthermore, those who had complete designs had only average grades, suggesting that the curricula are rewarding students for something other than developing design expertise.

The Accreditation Board for Engineering and Technology (ABET), the accrediting body for engineering, requires that engineering programs institute a continuous improvement process based on assessment of eleven student learning outcomes, including (c) an ability to design a system, component, or process to meet desired needs and (i) a recognition of the need for, and an ability to engage in, lifelong learning. The Computing Sciences Accreditation Board (CSAB) has comparable requirements. Preliminary data on changes in engineering education over the last ten years [7] has shown that engineering programs are increasing their use of student assessment (93 percent of program chairs report some or significant increase). Nonetheless, their use of the data for continuous improvement of the curriculum is much lower: only 55 percent reported that half or more of their faculty supported the use of assessment in continuous improvement, and only 35 percent reported that they used such data for continuous improvement.

We perceive that a major reason for the slow progress is that few assessments provide pointers about how to change curricula to improve learning outcomes. Knowing that students produce poor designs does not tell how or why they are failing. The quality model promotes continuous improvement because practitioners assess the process and then change the process in ways that improve the product (e.g., [8]).

A second problem with many of the assessments of student learning outcomes is that they are not given in the context of specific courses. It has been our experience in assessing design that students are often less motivated to do their best if they are asked to engage in a design task that gives them no credit. In addition, embedding such assessments in courses provides opportunities to complete several improvement loops within one semester.

THE PROPOSED SOLUTION

The goal of this project was to create assessments of design process that can be used for continuous improvement in undergraduate courses. Such assessments must be easily administered, easily scored, and provide information that can be used in continuous improvement. In addition, they must be authentic enough that both students and instructors believe their validity.

We believe that the same assessments can be used by instructors to improve their instruction and by students to improve both their design process skills and their ability to improve through self-reflection. Assessments of students' design process knowledge and skill are the basis for continuous improvement for both students and teachers. The results of such assessments tell instructors and students what aspects of the design process students understand and what aspects are missing or misunderstood (i.e., pointers toward improvement). Instructors can use the results to improve their instruction and students can use them to engage in reflective practice to figure out how to improve their design process. The changes are then implemented and students' learning is again assessed, beginning the cycle again. We feel it is important to teach students continuous improvement because that is the basis on which they can continue to learn and improve the way they learn throughout their lives.

NATURE OF DESIGN PROCESS

To use process assessment as a basis for continuous improvement, both instructors and students have to learn to focus on the students' design processes. Discovering the nature of design process is extremely difficult for several reasons. First, experts' knowledge is largely intuitive and subconscious, so they cannot describe their process. Second, experienced designers do not all share a common design process. Third, textbooks often present a prescriptive analysis of the design process, which sometimes is rooted in research on effective practices, but often is based on opinion. For example, many textbooks present the waterfall model of design [10], which presents the phases of design in a lockstep fashion (requirements, then design, then implementation, then verification, then maintenance). It assumes, for example, that requirements are set, stable, and fully evolved before analysis begins. The reality of design is that change is unavoidable and must therefore be explicitly accommodated in the life cycle.

There is now a fairly sizeable body of research in which experienced designers are asked to think aloud while designing, which will be described below. This research has resulted in a variety of claims about what designers actually do, but very little analysis of whether what experienced designers do is sound practice. The many examples of massive failures in large software products (e.g., The Standish Group reports a 74 percent project failure rate with 28 percent of projects being cancelled, [10]) suggest that standard practice may not be an adequate model on which to base curricular change. A few studies have investigated exceptional designers [11,12,13,14], but a solid picture has not yet emerged from such research. A final problem is that when experienced designers design in familiar areas, their process is quite different from when they design in unfamiliar areas (e.g., [15,16]).

Our search for a conceptualization of design process upon which to build instructional interventions was based on the following principles. First, we argue that student designers are always faced with unfamiliar content and so should learn the design process that works best for expert designers when designing unfamiliar products. In addition, expert designers in familiar contexts tend to skip steps that are important to learn. Second, we focused on the practices of extraordinary designers when possible. Third, when we had to rely on the research on

experienced, but not necessarily exceptional designers, we selected design principles for which there was research confirming their effectiveness. Fourth, we wanted our principles to permit innovative design, which occurs primarily during conceptual design.

From this large, diverse body of work, we developed a number of principles of design process skill that seemed both solid and important for students to learn. First, expert designers take time to understand the problem [13,17,18]. For example, Sonnentag found that excellent designers, in contrast to moderate performers, stopped after first reading each line of a requirements document and sketched out scenarios related to that requirement [13]. Goel [19] found that approximately 25 percent of statements (range 18 to 30 percent) in an open-ended design task were problem structuring. Most of these were at the beginning, but they also occurred later in design. Such practices, including taking the time to develop structured requirements, resulted in a 55 percent decrease in defects, a 50 percent reduction in total estimation error due to increased understanding of the features, and a 45 percent decrease in user-reported deficiencies, according to a case study of one software company [20].

The second principle is to consider alternative solutions, or at least to maintain flexibility in considering a solution, until the problem is well thought out. This is admittedly not always the way designers design. A number of studies show that designers tend to latch onto a solution early in the design process and to explore the problem in terms of that specific solution. This is particularly true of expert designers with extensive experience in the problem domain in which they are working [16,21,22]. Even when substantial difficulties are encountered, designers tend to make patches rather than changing solutions [22,23]. This trend has also been found with senior students [24]. Such strategies seem to be part of what Simon called "satisficing" strategies, which take less cognitive space than considering alternatives.

Although this is common practice, it is not optimal. Generating alternative conceptual designs from the beginning [18,25] results in better designs than considering either too few or too many. It can be effective to consider one design at the beginning so long as one is willing to consider alternatives at the first sign of trouble [26]. Outstanding designers tend to use one of these strategies [27,28]. In general, they have an ability to maintain openness and to tolerate ambiguity until the situation is fully explored.

A third characteristic of effective design process is early and repeated reviews. Sonnentag [13] found that high-performing designers "more often evaluated their design solution and started early with these evaluations." They also were much more likely to solicit feedback from team leaders and from coworkers. Other studies found that experienced designers are likely to review and reflect on design in terms of requirements [29,30,31]. Research supports this principle. Peer reviews catch over half the defects in requirements, design, and code [32]. Catching defects early is also cost-effective. Shull et al. report that finding and fixing a severe software problem after delivery is more expensive than finding and fixing it during the requirements and design phase—often 100 times more expensive for severe problems and twice as expensive for nonsevere defects.

Fourth, a popular prescriptive principle is top-down design, in which designers systematically consider components at a global level, and then refine all components to similar levels, etc. Experienced designers have been found to use top-down design, but they also deviate from it by exploring one component to a detailed level, called opportunistic design [25,31,33,34]. Although opportunistic behaviors sometimes represent a breakdown in top-down design or result from an incomplete exploration of the problem [34], they can be an intelligent

exploration of the feasibility of part of the design [33], and therefore a way in which designers provide self-feedback. A related advantage of top-down design is that it facilitates the examination of interrelations among components [35].

DESIGN PROCESS ASSESSMENTS

We developed five assessments of design process. Three of them assess students' declarative knowledge, that is, what they know about the design process. The other two assess how students use design process in action. We also developed a rubric for assessing design products. Each of the assessments was created in an iterative process. The team, which included an electrical engineer (Nixon Pendergrass), a mechanical engineer (Tesfay Meressi), a computer engineer (Paul Fortier), a computer scientist (Richard Upchurch), and a psychologist (Judith Sims-Knight), jointly developed the components of the assessment. For each assessment we developed a draft, gave it to engineering and computer science students, analyzed the results, and used the feedback to improve the assessment.

For each assessment we made sure that it met our basic requirements: 1) it could be assessed successfully within a course, 2) it could be analyzed to give both overall scores (so it could be graded), and 3) it would provide pointers toward improvement that would help both instructors and students.

We assessed reliability by the criterion of internal consistency, using Cronbach's alpha. We also collected preliminary evidence regarding construct validity in two ways. First, if we believe that engineering curricula provide an opportunity for students to improve their design process knowledge and skill, performance on these measures should improve as students progress through the curriculum. Second, scores should be correlated with project scores that include specific scores for design process as well as for design product, as was available in software engineering. Other tests of the assessments were included when relevant.

Below, we describe the three most promising assessments, two of declarative knowledge and one of applied skill. Information on all of them is available at www.cogsci.umassd.edu/design_process_assessments.

Design Process Knowledge Test

The Design Process Knowledge Test (DPKT) consists of 31 multiple-choice questions that assess whether students understand the design process principles described above. It was developed in three iterations. The third version was given to 10 freshmen in an integrated year engineering class, 11 sophomores in a circuits class, and 56 juniors and seniors in three courses (software engineering, senior design course for electrical and computer engineering, and a senior computer engineering course). It exhibited high reliability, with Cronbach's alpha = .89. It also successfully discriminated between students who understood design process and those who did not—16 of the 31 questions had Discrimination Indices (DIs) greater than .40 and 6 had DIs in the 30s.

The DPKT also showed some construct validity. Because we had far more seniors than students in the other classes, we randomly eliminated 28 seniors so we could test whether there was improvement in performance as a function of year. An analysis of variance revealed that performance differed as a function of year, $F(3, 40) = 21.2$, $p < .001$. Engineering freshmen scored significantly lower than sophomores, juniors, and seniors, and the three upper classes did not differ by Tukey HSD post hoc test at alpha = .05 (see Table 1). To assess whether the random

elimination of seniors affected the results, we ran a t-test comparing performance of seniors who were in the analysis vs. those who had been eliminated. The difference was not significant, $t(35) < 1$, $p = .82$.

Table 1. Performance on DPKT as a Function of Year

				95% Confidence Interval	
	n	**Mean**	**Standard Deviation**	**Lower Bound**	**Upper Bound**
Freshmen	10	.11	.04	.08	.14
Sophomores	11	.47	.14	.38	.57
Juniors	10	.54	.15	.43	.64
Seniors	14	.56	.19	.45	.67

We also have preliminary evidence that the DPKT is tapping knowledge related to design. Performance correlated with project scores in a semester-long project in software engineering, $r(48) = .32$, $p = .03$, but not with the nonproject score in that class, $r(48) = .16$, $p = .27$. This is evidence for both convergent and discriminant validity that the DPKT is assessing knowledge related to design and not just intelligence or academic motivation.

The DPKT provides two kinds of pointers for improvement. First, the questions can be grouped into the categories described earlier. Categories on which a class scores below 80 percent can be targeted for improvement. Second, some of the questions provide false alternatives that reveal particular misconceptions. For example, only 29 percent of students knew that the best time to identify communication between components is "global design". A common alternative choice was "detailed design". This suggests that instructors need to demonstrate why waiting until detailed design is counterproductive. A further indication that the DPKT is appropriate in this context is that seniors on average were correct on only 56 per cent of the questions; there is plenty of room for improvement.

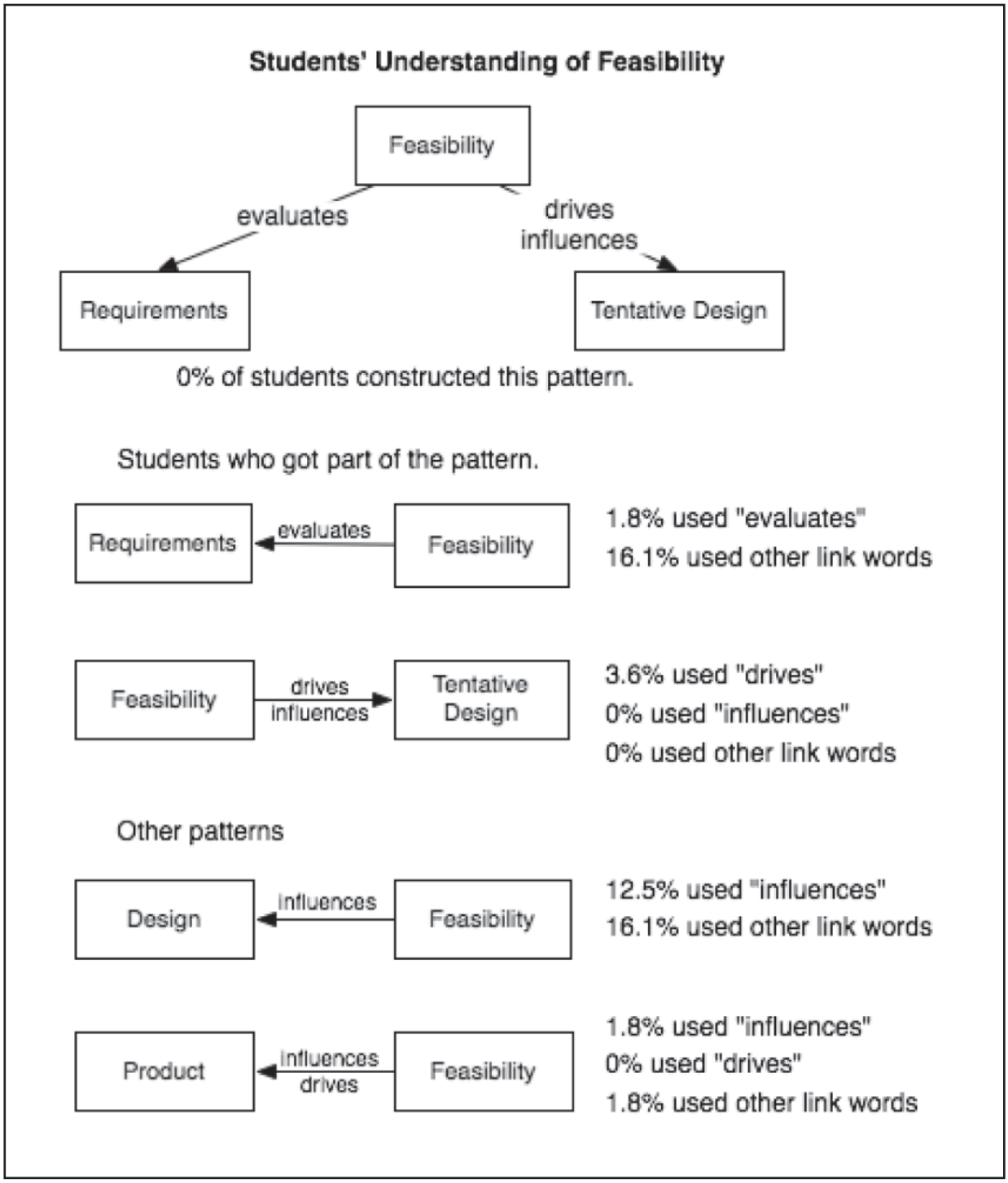
Concept Map Task

People with effective declarative knowledge not only understand constructs, they also know how such constructs are interrelated. Concept maps, broadly conceived, are network diagrams in which concepts (nouns) are nodes and the relationships between concepts (verbs) are the links (cf. [36], for uses in engineering). Concept maps have been successfully used as assessment devices [37,38,39] and as pedagogical devices [40,41,42], but generally are used to map factual knowledge rather than process knowledge.

After several iterations, we arrived at the following procedure. Students completed a short training, which involved filling in blanks and extending a concept map of a computer game. The training also included instructions on using IHMC Concept Map Software (CMAP) [43], a free concept map generation tool from the University of South Florida. The students then opened a CMAP that had ten nodes and a list of six verbs. They were instructed to move the nodes around the workspace and to create links using only the six verbs provided.

We analyzed the maps in two ways: 1) by comparing propositions (node-link-node) to those of an expert map and 2) by overall patterns. The correct propositions were derived from an expert map agreed upon by the participating faculty. The number of correct propositions can be added to derive a grade for each student and the performance across a class can be used for continuous improvement. Figure 1 shows that no one in the class understood the relationships among feasibility, requirements, and tentative design. In fact, fewer than 20 percent

Figure 1. Part of a Continuous Improvement Report Based on Concept Map Propositions for a Software Engineering Class



were correct on any part of the pattern (i.e., feasibility evaluates requirements, feasibility drives or influences tentative design), and 28 percent linked feasibility with design, which is much too late in the design process. This result tells the instructor that she needs to address these issues.

In another analysis we tested the hypothesis that students in a senior design course should have better maps than students in a sophomore course, and whether this varied by major. Because the variance of performance on

propositional scoring was restricted by low performance , we used a related measure, number of links, in a 2-by-2 analysis of variance. The independent variables were course (sophomore course vs. senior design course), and major (electrical or computer engineering). The main effect of major was highly significant, $F(1, 37) = 9.24$, $p = .004$, and the interaction was of borderline significance, $F(1, 37) = 4.28$, $p = .05$, respectively. This means that computer science majors made more links in their maps and that perhaps the increased experience affected the two majors differently (see Table 2).

Table 2. Descriptive Statistics for Number of Links on Concept Map

	Sophomore Course			Senior Course		
	Means	**Standard Deviations**	**n**	**Means**	**Standard Deviations**	**n**
Computer Engineers	12.5	1.52	6	13.06	1.70	18
Electrical Engineers	11.8	1.60	11	10.37	.52	8

The second way the concept maps were analyzed was by categorizing the overall patterns, based on Hart's system [44]. Like Hart, we had a cat's cradle, linear, web (see Figures 2a, b, and c, respectively), and a branching pattern. We developed our own coding system because Hart did not provide one. Two raters independently rated 82 concept maps and agreed on 96 percent of them. Reliabilities for individual categories ranged from 91 to 100 percent agreement.

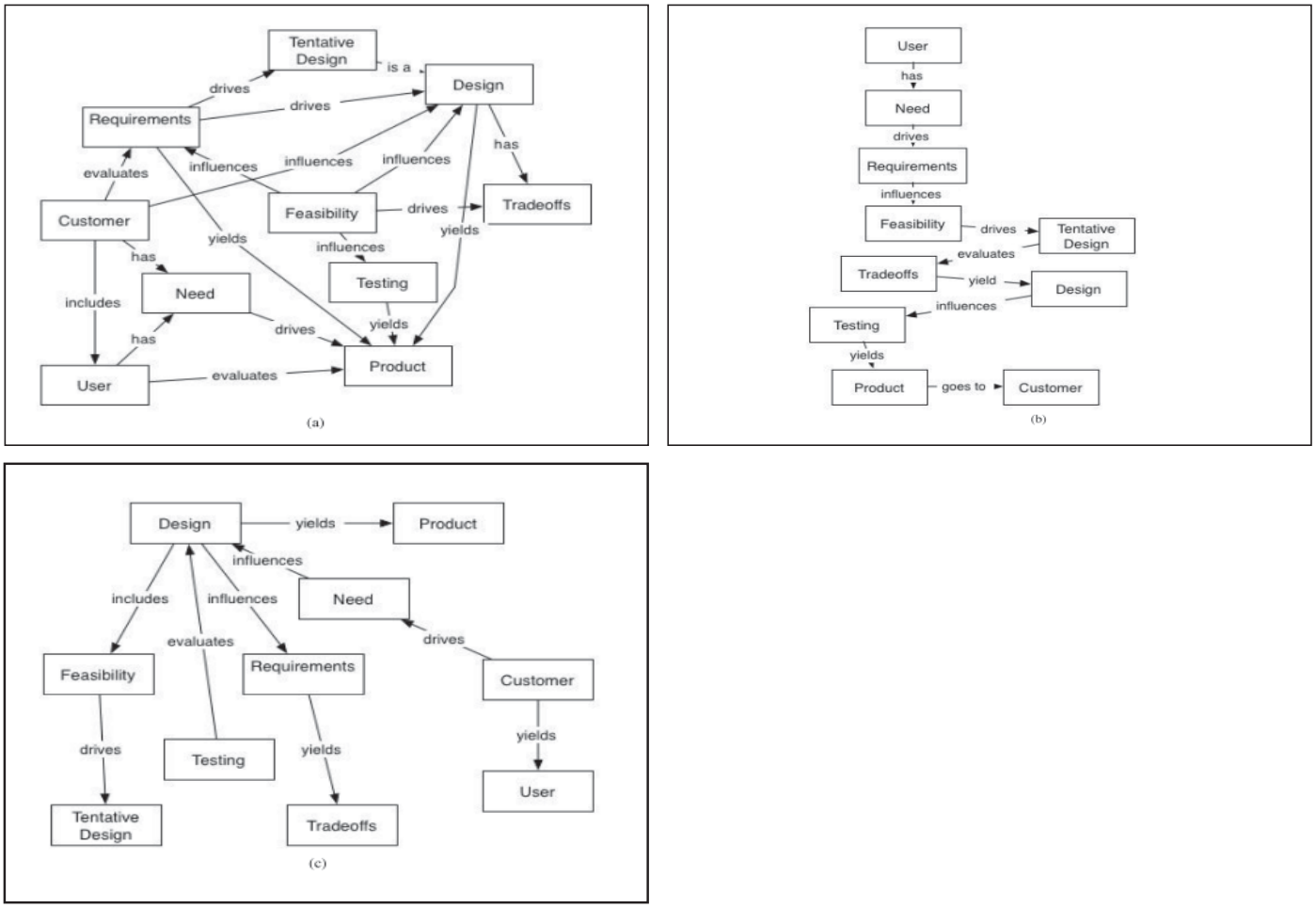
The patterns differed across majors, $c2\ (4, n = 42) = 10.76$, $p = .03$ (see Table 3), but not across course level, $c2\ (4, n = 41) = 3.0$, $p = .56$. Computer engineers predominantly created cat's cradles, whereas electrical engineering students were much more likely to create linear and web maps. The cat's cradle pattern (Figure 2a) suggests a representation of the design process as a highly integrated network with interconnections among the parts (although the student in the example clearly did not understand the role of the tentative design). Linear patterns (Figure 2b) reflect a waterfall model. Web patterns (Figure 2c) reflect a conception of the design process in which one phase of design is central.

Table 3. Proportion of Concept Map Patterns as a Function of Major

	Strictly Linear	**Linear with Branches**	**Web**	**Cats Cradle**	**Branching**
Computer Engineers	.04	.21	.21	.50	.04
Electrical Engineers	.28	.06	.39	.17	.11

These analyses should not be taken as the basis for any strong conclusions about differences between engineering majors, but rather as preliminary evidence that a concept map construction task can detect differences between groups. That the contrasting analysis techniques yield similar findings is quite reassuring, because they are based on different conceptualizations about concept maps. Propositional scoring is predicated on the belief that experts' maps are better than novices, whereas categorical coding is based on the belief that various mental models of the domain lead to different patterns of maps.

Figure 2. Examples of Student Concept Maps. (a) Cat's cradle, a network pattern that indicates the student understands the iterative nature of the design process, (b) A linear pattern reminiscent of the waterfall model, with no iterations, and (c) A web pattern, in which one concept (design, in this case) is conceived as a central process



Simulation Task

Students do not exhibit much process awareness while creating designs, even when asked explicitly to include process information or to reflect on their process after completion of a design task [45,46]. This lack of awareness has also been found in physics and other cognitive tasks [47,48,49]. A simulation task, in contrast, might provide a context in which students can exhibit their process skill because it can remove the cognitive demands of creating the design and so allow students to focus on process issues. This also reduces the administration time, but retains the authenticity of a situated cognition task. Furthermore, simulations have been shown to be effective learning devices (see [50,51,52]) in the closely related domain of software management).

In the simulation, respondents are advising a design team composed of four members, one of which is team leader. A fifth character, the manager, intrudes on occasion. The respondent "watches" the team create a design and is asked to give "advice" to the team at points when a design principle would be in play. For example, the respondent is asked what the team should do after they have completed the requirements (correct answer: review them in terms of feasibility and trade-offs). Later the respondent is given the opportunity to advise the team to consider several alternatives. If s/he does not choose the alternative designs path, later the team runs into

trouble with their first-chosen design, and the respondent is given an opportunity to suggest that the team should reconsider that design and consider alternatives.

The task is presented via the web. The format of the simulation design follows a standard sequence (see Figure 3 for the opening sequence). Respondents are told about what the team has done and presented the dilemma of what to do next. They first answer the question in narrative format. Then the program presents possible courses of action that the "team" has generated, and respondents have to choose among them. The number of alternatives varies from 2 to 5. Frequently there is more than one acceptable answer, which may be equally good or not. When respondents choose a response, they are asked to elaborate on their answer. If their choice is poor, the virtual team considers that response and eventually the simulation asks the respondent to consider the remaining alternatives.

Figure 3. Example of simulation task

Simulation:
You have been asked to advise a software team as they develop a design. There are four team members--Jill, Michael, David, and Pedro. Pedro is the team leader and supervising him is Tom.

Here is the Problem Statement: . . .

What should the team do first?

Student:
They should figure out a way to design it.

Simulation:
Here are some suggestions various team members have made. Choose the one you think is the best idea.
- ○ Meet with the customer.
- ○ Elaborate the problem statement.
- ○ Choose a design.

The student chooses "Choose a design" and is given a score of -1.

Simulation:
The team discusses potential designs, but they soon decide they do not understand the problem statement. What do you think they should do?
- ○ Meet with the customer.
- ○ Elaborate the problem statement.

The student chooses "Meet with the customer" and is given a score of +2.

Simulation:
How should the team prepare for the meeting with the customers?

Student:
They should develop something to show the customers how the program might work.

The multiple choice alternatives are all allocated a specific number of points ranging from -1 to +4. The -1 score is assigned to alternatives that will lead to negative outcomes and 0 to alternatives that will neither hurt nor help. The scoring scheme gives the highest score to students who choose the best alternative the first time they are given the choice, and a lower score to those who figure it out on the second or third pass.

The numeric scores may be added to create an overall score. In addition, all students' answers to each individual question, both open-ended and multiple-choice, can be analyzed to provide pointers toward improvement.

We gave a software design prototype to 8 students in a junior-level computer science course. The results demonstrated that the task was at the right level for the students, which contrasted favorably to various ways of assessing design process during design tasks, which were too difficult even for seniors in design courses [39]. On average, 4.6 of 8 students answered the questions correctly (SD = 2.3) and even on the hardest question one student answered correctly.

The simulation task also provided pointers toward improvement. On the questions demonstrated in Figure 3, only one student had the misconception that one should start designing before exploring the problem. In contrast, when asked to respond to a team member's suggestion of a design, four students chose to proceed with that design and four others chose to explore alternatives. In both these examples, the students replied to the follow-up questions in the same way they had the first. This suggests that these errors are heartfelt beliefs rather than random responses.

EFFECTIVE ADMINISTRATION

The best assessments will not work unless students are convinced that it is a meaningful activity. The traditional way education makes activities "meaningful" is to grade them. All our assessments can be graded. While not denigrating the power of grading, we recommend that students take assessments more seriously if they see the relevance to the course and to their learning. Thus, we recommend that these assessments be used in courses in which learning design process is a priority and that use the results of the assessments for continuous improvement, either for the course or for individuals. Instructors share all or portions of the class results with the students. This is, of course, easy to do if instructors are analyzing the results for continuous improvement.

SUMMARY AND IMPLICATIONS

Design process is fiendishly difficult to teach, because it involves complex, implicit, intuitive skills that are not well understood, let alone translatable to students. To teach such complex skills requires innovative pedagogical techniques. One promising strategy is assessment-based continuous improvement. With such a technique instructors can identify where and how students are failing, design instruction that might help them succeed, and then assess whether the new intervention results in higher success rates.

Teaching students to use the same continuous improvement loop has the potential to help them in two ways. First, of course, they can reflect on how they go about designing and figure out more effective ways. Second, they learn to attend to the processes by which they work and change those processes. This is the essence of lifelong learning.

To even test whether these claims are true, one must have process assessments that form the core of the continuous improvement loop. In this project we developed a suite of assessments of design process to use in such a context. The goal of this project was to develop course-based assessments of design process knowledge and skill that could anchor continuous improvement loops, both for faculty's instruction and for students' reflective practice. The three most successful assessments were:

1) Design Process Knowledge Test (DPKT), a multiple-choice test of what students know about design process (declarative knowledge),

2) Concept map construction task, in which students draw a network diagram to represent the relationships among design process concepts (structure of declarative knowledge), and

3) Design process simulation task, in which the respondent acts as "a fly on the wall" to offer "advice" to a virtual design team as the latter creates a design (design process procedural knowledge).

All of these assessments can be given in a course context, either as homework, or as class or laboratory exercises. All assessments are structured so that they can provide information about what ideas students grasp and what they do not grasp. As we developed the assessments, we developed a straightforward procedure for each, we tested their psychometric properties where appropriate, and we collected some preliminary data about their utility.

This project also revealed that providing semester-long or year-long project courses is not sufficient to develop good design skills. We gave these assessments to three years worth of students in software engineering (one semester) and in the senior electrical and computing engineering project course (two semesters). Students consistently showed little use of or knowledge of design processes, and quite low quality of designs. The best performance we got was their scores on the declarative knowledge test, on which the seniors got just over half the questions correct. This is consistent with the study cited earlier [4] of students from 21 institutions in four countries that found that 62 percent of the so-called designs had no design information and that only 3 percent of the designs could be called complete.

The next step in this research is to embed these assessments in courses that focus on improving design process skills through continuous improvement. The assessments will provide instructors feedback about what the class understands and fails to understand, which can serve as a basis on which they can change instruction. Results of the same assessments can be fed back to the students to help them reflect upon their knowledge and skill and take steps to improve. This will permit an evaluation of whether such practice helps students improve in the effective use of design processes, in creating better designs, and in engaging in reflective practice.

Uses Beyond Engineering

Assessment-based continuous improvement is being mandated across the United States by the accrediting agencies of higher education. Departments are instituting assessment of learning outcomes, but currently few such assessments give pointers toward improvement, and even fewer analyze the process by which students learn. The quality model is predicated on assessment of the process, not the product (e.g., [8]). It is not sufficient to have curricula increase their emphasis on skill development. To make assessment-based continuous improvement work, it is necessary to assess the processes by which students work and then to change their processes based on the outcomes of the assessment.

This project provides a model of how to develop process-based assessments. In addition, the assessments themselves can be adapted to any discipline that includes designing, which narrowly speaking includes all science and engineering disciplines. Broadly speaking it includes disciplines as disparate as visual design and management. Even humanities can consider design processes in the writing of term papers. Indeed, the Psychology Department at University of Massachusetts Dartmouth has already adapted the simulation task as an assessment of understanding of evaluation of research.

For learning objectives involving declarative knowledge, a concept map task might tap the understanding of relations among constructs in ways traditional multiple-choice tests cannot, and with much less faculty work than essay exams entail. Although we are far from the first or only people to use concept maps in education, we have developed a training task and evaluation technique that can be easily learned and applied by others. Even the design process knowledge test may inspire faculty to include assessment of process knowledge in addition to factual knowledge.

The promise of future contributions to other disciplines is even greater. If we can show that these assessments can be used effectively to help instructors to improve their instruction and help students to improve their learning, it would serve as a model for other disciplines to use process assessments in continuous improvement.

REFERENCES

[1] Project Kaleidoscope, *Transforming America's Scientific and Technological Infrastructure: Recommendations for Urgent Action: Report on Reports II,* (Author, Washington, D.C.), 2006.

[2] Atman, C.J., and Bursic, K.M., "Teaching engineering design: Does reading a textbook make a difference?" *Research in Engineering Design*, vol. 8, pp. 240-250, 1996.

[3] Fincher, S., Petre, M., Tenenberg, J., Blaha, K., Bouvier, D., Chen, T-Y., Chinn, D., Cooper, S., Eckerdal, A., Johnson, H., McCartney, R., Monge, A., Mostrom, J. E., Powers, K., Ratcliffe, M., Robins, A., Sanders, D., Schwartzman, L., Simon, B., Stoker, C., Tew, A. E., and VanDeGrift, T., *Cause for alarm?: A multi-national, multi-institutional study of student-generated software designs*, Technical Report 16-04, (Computing Laboratory, University of Kent, Canterbury), 2004.

[4] Eckerdal, A., McCartney, R., Mostrom, J. E., Ratcliffe, M., and Zander, C., "Can graduating students design software systems?" *SIGCSE-06*, March 1-6. Houston, Tex., 2006.

[5] Ratcliff, B., and Siddiqi, J. I. A., "An empirical investigation into problem decomposition strategies used in program design," *International Journal of Man-Machine Studies,* vol. 22, pp. 77-90, 1985.

[6] Sims-Knight, J., Upchurch, R., and Fortier, P., "A simulation task to assess students' design process skill, *Proceedings of the Frontiers in Education Conference*, October, 2005.

[7] Prados, J. W., Peterson, G. D., and Lattuca, L. R., "Quality assurance of engineering education through accreditation: The impact of Engineering Criteria 2000 and its global influence," *Journal of Engineering Education*, vol. 94, pp. 165-183, 2005.

[8] Repenning, N. P., and Sterman, J. D., "Getting quality the old-fashioned way: Self-confirming attributions in the dynamics of process improvement," 1997. web.mit.edu/jsterman/www/

[9] Royce, W., "Managing the development of large software systems," *Proceedings of IEEE WESCON*, vol. 26, pp. 1-9, 1970.

[10] The Standish Group, *The CHAOS Report,* (The Standish Group International), 2003. www.standishgroup.com

[11] Cross, N., "The expertise of exceptional designers," in Cross, N., and Edmonds, E., (eds.), *Expertise in design*, (Creativity and Cognition Press, University of Technology, Sydney, Australia), 2003.

[12] Lawson, B., *Design in Mind*, (Butterworth-Heinemann, Oxford), 1994.

[13] Sonnentag, S., "Using and gaining experience in professional software development," in Salas, E., and Klein, G., (eds.), *Linking Expertise and Naturalistic Decision Making,* (Erlbaum, Mahwah, N.J.), 2001.

[14] Turley, R. T., and Bieman, J. M., "Competencies of exceptional and non-exceptional software engineers," *Journal of Systems and Software*, vol. 28, no. 1, pp. 19-38, 1995.

[15] Adelson, B., and Soloway, E., "The role of domain experience in software design," *IEEE Transactions on Software Engineering,* vol. 11, pp. 1351-1360, 1985.

[16] Lloyd, P., and Scott, P., "Discovering the design problem," *Design Studies*, vol. 15, pp. 125-140, 1994.

[17] Batra, D., and Davis, J. G., "Conceptual data modeling in database design: Similarities and differences between expert and novice designers," *International Journal of Man-Machine Studies,* vol. 37, pp. 83-101, 1992.

[18] Fricke, G., "Successful individual approaches in engineering design," *Research in Engineering Design*, vol. 8, pp. 151-165, 1996.

[19] Goel, V., *Sketches of Thought*, (MIT Press, Cambridge, Mass.), 1995.

[20] Damian, D., and Chisan, J., "An empirical study of the complex relationships between requirements, engineering processes and other processes that lead to payoffs in productivity, quality, and risk management," *IEEE Transactions on Software Engineering,* vol. 32, pp. 433-453, 2006.

[21] Lloyd, P., and Scott, P., "Difference in similarity: interpreting the architectural design process," *Planning and Design,* vol. 22, pp. 385-406, 1995.

[22] Ullman, D. G., Dietterich, T. G., and Stauffer, L. A., "A model of the mechanical design process based on empirical data," *Artificial Intelligence in Engineering Design and Manufacturing*, vol. 2, pp. 33-52, 1998.

[23] Rowe, P., *Design Thinking*, (MIT Press, Cambridge, Mass.), 1987.

[24] Ball, L. J., Evans, J. St. B. T., and Dennis, L., "Cognitive processes in engineering design: a longitudinal study," *Ergonomics*, vol. 37, pp. 1753-1786, 1994.

[25] Gunther, J., and Ehrlenspiel, K., "Comparing designers from practice and designers with systematic design education," *Design Studies*, vol. 20, pp. 439-451, 1999.

[26] Visser, W., "More or less following a plan during design: Opportunistic deviations in specification," *International Journal of Man-Machine Studies,* vol. 33, pp. 247-278, 1990.

[27] Lawson, B., *Design in Mind*, (Butterworth-Heinemann, Oxford), 1994.

[28] Cross, N., "Expertise in design: An overview," *Design Studies,* vol. 25, pp. 427-441, 2004.

[29] Mullins, C.A., Atman, C.J., and Shuman, L.J., "Freshman engineers' performance when solving design problems," *IEEE Transactions on Education,* vol. 42, no. 4, pp. 281-287, 1999.

[30] Guindon, R., "Knowledge exploited by experts during software system design," *International Journal of Man-Machine Studies*, vol. 33, pp. 279-304, 1990.

[31] Valkenburg, R, and Dorst, K., "The reflective practice of design teams," *Design Studies*, vol. 19, pp. 249-271, 1998.

[32] Shull, F., Basili, V., Boehm, B. A., Brown, W., Costa, P., Lindvall, M., Port, D., Rus, I., Tesoriero, R., and Zelkowitz, M., "What we have learned about fighting defects," *Proceedings of the Eighth IEEE Symposium on Software Metrics METRICS'02*, Ottawa, Canada, 2002.

[33] Ball, L.J. and Omerod, T.C., "Structured and opportunistic processing in design: a critical discussion," *International Journal of Human-Computer Studies*, vol. 43, pp. 131-151, 1995.

[34] Guindon, R., "Designing the design process: Exploiting opportunistic thoughts," *Human Computer Interaction,* vol. 5, pp. 305-344, 1990.

[35] Biswas, G., Goldman, S., Fisher, D., Bhuva, B., and Glewwe, G., "Assessing design activity in complex CMOS circuit design," in Nichols, S.F., Chipman, and Brennan, R.L., (eds), *Cognitively Diagnostic Assessment*, (Erlbaum, Hillsdale, N.J.), pp.167-188, 1995.

[36] Turns, J., Atman, C. J., and Adams, R., "Concept maps for engineering education: A cognitively motivated tool supporting varied assessment functions," *IEEE Transactions on Education*, vol. 43, pp. 164-173, 2000.

[37] McClure, J. R., "Concept mapping and acquisition on cognitive skills: Concept maps as a tool to study skill acquisition", *American Educational Research Association*, Montreal, Canada, April 1999.

[38] Herl, H. E., Baker, E. L., and Niemi, D., "Construct validation of an approach to modeling cognitive structure of U.S. history knowledge," *Journal of Educational Research*, vol. 89, pp. 206-218, 1996.

[39] Coleman, E., "Using explanatory knowledge during collaborative problem solving in science," *Journal of the Learning Sciences*, vol. 7, pp. 387-427, 1998.

[40] Jones, M. G., Carter, G., and Rua, M. J., "Children's concepts: Tools for transforming science teachers' knowledge," *Science Education*, vol. 83, pp. 545-557, 1999.

[41] Walker, J. M. T., Cordray, D. S., King, P. H., and Fries, R. C., "Expert and student conceptions of the design process: Developmental differences with implications for educators," *International Journal of Engineering Education*, vol. 21, pp. 467-479, 2005.

[42] West, D. C., Pomeroy, J. R., Park, J. K., Gerstenberger, E. A., and Sandoval, J., "Critical thinking in graduate medical education: A role for concept mapping assessment?" *Journal of the American Medical Association*, vol. 284, pp. 1105-1110, 2000.

[43] Institute for Human and Machine Cognition, University of West Florida. http://cmap.ihmc.us/

[44] Hart, I., "Visualising structural knowledge," Fifth International Conference for Computers in Education, Beijing, October, 1998.

[45] Sims-Knight, J., Upchurch, R., and Fortier, P., "Assessing students' knowledge of design process in a design task," *Proceedings of the Frontier In Education Conference*, 2005.

[46] Smith, R. P., and Tjandra, P., "Experimental observation of iteration in engineering design," *Research in Engineering Design*, vol. 10, pp. 107-117, 1998.

[47] Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., and Glaser, R., "Self-explanations: How students study and use examples in learning to solve problems," *Cognitive Science*, vol. 13, pp. 145-182, 1989.

[48] Chi, M. T. H., de Leeuw, N., Chiu, M., and LeVancher, C., "Eliciting self-explanations improves understanding," *Cognitive Science*, vol. 18, pp. 439-447, 1994.

[49] Berardi-Coletta, B., Buyer, L. S., Dominowski, R. L., and Rellinger, E. R., "Metacognition and problem solving: A process-oriented approach," *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 21, pp. 205-223, 1995.

[50] Pfahl, D., Laitenberger, O., Dorsch, J., and Ruhe, G., "An externally replicated experiment for evaluating the learning effectiveness of using simulations in software project management education," *Empirical Software Engineering*, vol. 8, pp. 367-395, 2003.

[51] Collofello, J., Rus, I., Ramkrishnan, V., and Dooley, K., "Using software simulation to assist software development organizations in making 'good enough quality' decisions," Summer Computer Simulation Conference, SCSC98, 1998.

[52] Navarro, E. O., and van der Hoek, A., "Design and evaluation of educational software process simulation environment and associated model," *Proceedings of the 18th Conference on Software Engineering Education and Training*, CSEET05, 2005.

Project CAT: Assessing Critical Thinking Skills

Barry Stein*, Ada Haynes, Michael Redding - *Tennessee Technological University*

OVERVIEW

This chapter reviews the development of the CAT (Critical thinking Assessment Test) instrument, a new interdisciplinary assessment tool for evaluating students' critical thinking skills. The CAT instrument is designed to directly involve faculty in the assessment of student strengths and weaknesses and thereby support the implementation of pedagogical improvements within and across disciplines. An NSF award allowed Tennessee Technological University (TTU) to partner with six other institutions across the United States (University of Texas, University of Colorado, University of Washington, University of Hawaii, University of Southern Maine, and Howard University) to evaluate and refine the CAT instrument. The CAT instrument has high face validity when evaluated by a broad spectrum of faculty across the United States in STEM and non-STEM disciplines, good criterion validity when compared to other instruments that measure critical thinking and intellectual performance, good reliability, and good construct validity using expert evaluation in the area of learning sciences. The broader impacts and potential benefits for improving undergraduate education are discussed.

IMPORTANCE OF CRITICAL THINKING

Most employers and educational institutions recognize the importance of effective critical thinking skills. For example, in 1990, the U.S. Department of Education stated as a goal that "the proportion of college graduates who demonstrate an advanced ability to think critically, communicate effectively, and solve problems will increase substantially" [1]. This goal became part of the "Goals 2000: Educate America Act" passed by Congress [2]. Many educators have also argued for the importance of preparing people to think critically [3,4,5,6,7,8,9,10]. In the National Assessment of Educational Progress (NAEP), "results suggest that although basic skills have their place in pedagogy, critical thinking skills are essential" [11]. Similarly, a report from the American Association of Universities indicated critical thinking and problem solving skills were essential for college success [12]. According to Derek Bok, president of Harvard University, national studies have found that more than 90 percent of faculty members in the United States consider critical thinking the most important goal of an undergraduate education [13].

Increasingly, the importance of critical thinking/problem solving skills in the workplace is also being recognized. For example, Halpern [14] argues, "virtually every business or industry position that involves responsibility and action in the face of uncertainty would benefit if the people filling that position obtained a high level of the ability to think critically" (see also [15]). A TTU survey of employers revealed that skills typically associated with critical thinking represented four out of the top five skills considered most important [16]. Similarly, a survey of New Jersey employers by the John J. Heldrich Center at Rutgers also found that skills typically associated with critical thinking represented three out of the top six skills considered most important. Those same employers also considered less than half of the graduates of two-year programs prepared for critical thinking, and only 56 percent of the graduates of four-year programs prepared for critical thinking [17]. A recent CNN poll of employers also found that critical thinking is one of the top five skills employers felt both critical to their businesses and most important in potential job candidates [18].

Although the importance of critical thinking is widely recognized, there is a gap between that recognition and the reality of what is taught and learned in the classroom. In order to bridge this gap, valid and reliable assessment tools are needed to measure these higher order thinking skills. With increasing pressure for accountability in education, "What gets measured gets taught... We must measure what we value or it won't get taught" [19].

DEVELOPMENT PROCESS

Our initial efforts to identify and then develop an effective assessment tool for critical thinking were stimulated by a state-wide "Performance Funding Initiative" to pilot test instruments designed to assess critical thinking beginning in 2000. TTU approached this task with the understanding that assessment would ultimately need to be linked to improvement initiatives at some point in the future.

There are three important characteristics of assessments that can foster genuine efforts to improve the quality of student learning:

1) The assessment must be an authentic and valid measure of progress toward the underlying goal.

2) The assessment should promote the motivation from within to change - the more engaged faculty are in the evaluation process the better.

3) The assessment should evaluate skills that are considered important within the framework of contemporary learning sciences.

Assessment approaches that violate these criteria can actually be detrimental to the quality of student learning or at best point out weaknesses that will never be addressed.

Although we explored a variety of assessment tools and pilot tested several existing instruments, none of the available tools satisfied all of our criteria. Many instruments assessed very narrow definitions of critical thinking (i.e., logical reasoning/reading comprehension) and/or did not sufficiently involve faculty in the evaluation of student performance. Consequently, we assembled an interdisciplinary team of faculty with an expert in learning sciences to identify a core set of skills associated with critical thinking across disciplines and then developed questions to assess those skills. TTU spent three years refining the questions and accompanying scoring guide for this test, which involved mostly short answer essay-type questions. These efforts were guided by the following principles:

1) Identify critical thinking skills across disciplines that faculty genuinely believe underlie critical thinking.

2) Develop an instrument that involves faculty and students in activities that reveal weaknesses and encourages quality improvement initiatives.

3) Develop a reliable instrument that students find intrinsically interesting.

4) Develop an instrument based upon contemporary theory in learning sciences.

Once the instrument was developed at our university, NSF funding supported collaboration with six other institutions across the country (University of Texas, University of Colorado, University of Washington, University of Hawaii, University of Southern Maine, and Howard University) to administer the test, conduct scoring workshops, evaluate student performance, and gather faculty input to evaluate and refine the instrument. In addition, NSF funding allowed us to work with external consultants in learning science and education to refine the instrument and collect information to evaluate validity and reliability.

GENERAL FEATURES OF THE CAT INSTRUMENT

The CAT instrument is currently administered in paper form, although there are plans to develop an electronic version. The test consists of 15 questions, with the majority requiring short-answer essay responses to evaluate the 12 skill areas listed in Table 1. Student responses to these essay questions provide a better understanding of students' thought processes and their ability to think critically and creatively when confronted with real-world problems than would be provided by multiple-choice questions. Essay assessments are purported to 1) have higher construct validity; 2) exhibit less racial bias; 3) foster more faculty involvement; and 4) flexibly assess a wider range of skills than multiple-choice questions [20].

Another feature of the CAT instrument is that it provides numerous opportunities for students to learn during the assessment, a process known as "dynamic assessment" [21,22,23,24, 25]. Dynamic assessment techniques are needed to measure the extent to which people can understand new information and apply that information to a novel situation. The CAT instrument uses dynamic assessment together with problems that are intrinsically interesting to students and representative of real-world problems. These features contribute to the validity of the test and students' motivation to perform well on the test regardless of their discipline. Anecdotal reports from a variety of institutions indicate that students do find the test interesting and engaging. The latter observation is particularly important for institutions that are trying to accurately assess students' performance in situations where test performance does not directly impact a student's course grade.

The CAT instrument can be administered in one hour and most students complete the test in about 45 minutes. A detailed scoring guide has been developed and refined to help faculty evaluate student responses.

FINDINGS RELATED TO PROJECT CAT

A major challenge in developing any instrument designed to evaluate critical thinking is to find agreement among faculty across disciplines and institutions about what skills underlie critical thinking. These skill areas must also link to current theory in learning sciences and cognition for construct validity. The skill areas assessed by the CAT instrument correspond to the higher order cognitive skills in Bloom's Taxonomy (comprehension, application, analysis, synthesis, and evaluation) [26].

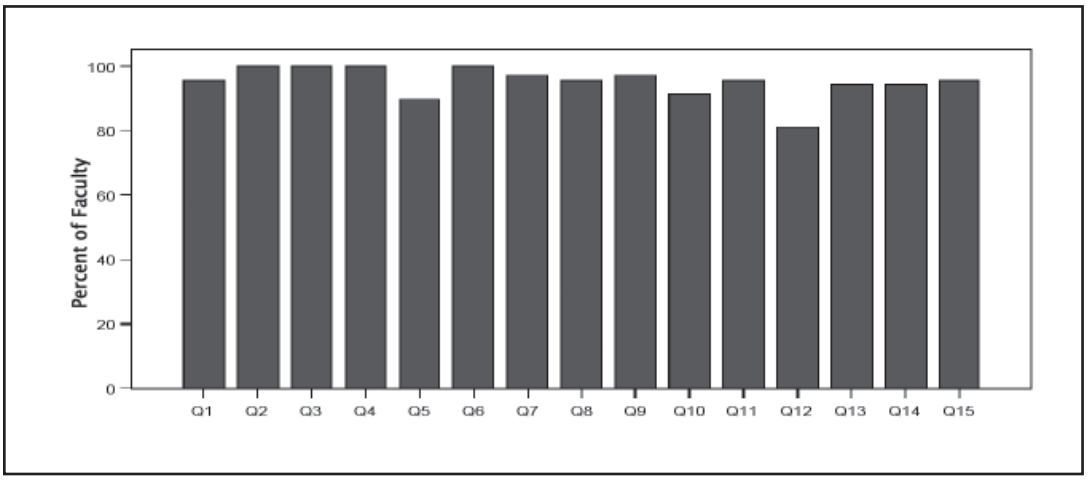
Faculty participants in scoring workshops at each participating university were asked to indicate which of the skill areas targeted by the CAT Instrument (see Table 1) they considered to be important components of critical thinking. The findings indicate that all of the skill areas targeted by the CAT instrument were generally perceived as important components of critical thinking by most faculty who participated in the scoring workshops. The area with least agreement (79.4 percent) concerned using mathematical skills to solve a complex real-world problem.

Table 1. Skill Areas Assessed by the CAT Instrument

Separate factual information from inferences that might be used to interpret those facts.
Identify inappropriate conclusions.
Understand the limitations of correlational data.
Identify evidence that might support or contradict a hypothesis.
Identify new information that is needed to draw conclusions.
Separate relevant from irrelevant information when solving a problem.
Learn and understand complex relationships in an unfamiliar domain.
Interpret numerical relationships in graphs and separate those relationships from inferences.
Use mathematical skills in the context of solving a larger real-world problem.
Analyze and integrate information from separate sources to solve a complex problem.
Recognize how new information might change the solution to a problem.
Communicate critical analyses and problem solutions effectively.

The faculty who participated in the scoring workshops were also asked to evaluate the face validity of each question contained in the CAT instrument. Most faculty felt that the questions included on the CAT instrument were valid measures of critical thinking (see Figure 1). The question with the lowest overall support (question 12 = 81 percent) involved using a mathematical calculation to help solve a complex real-world problem. The percent of faculty who considered this question a valid measure has risen as a result of improvements in the question. STEM faculty rate the face validity of questions slightly higher than non-STEM faculty in every case, although the difference between STEM and non-STEM faculty was only significant ($p < .05$) on question 12. These findings provide strong evidence for the face validity of the test. The external consultant and evaluators have also positively evaluated the construct validity of the instrument.

Figure 1. Percent of Faculty Indicating Each Question Measures a Valid Component of Critical Thinking



Criterion validity for a test of this type is rather difficult to establish since there are no clearly accepted measures that could be used as a standard for comparison. The approach we have taken is to look for reasonable but moderate correlations with other (more narrow) measures of critical thinking and other general measures of academic performance. The rationale underlying these comparisons is that the critical thinking skills measured by the CAT instrument should correlate at a moderate level with other measures of critical thinking and academic

performance. The findings support these aims, with the highest correlation between the CAT and the California Critical Thinking Skills Test (CCTST), indicating that only about 42 percent of the variability in the CAT instrument is explained by the CCTST (see Table 2).

Table 2. CAT Correlations with other Performance Measures

	ACT	SAT	Academic Profile	Grade Point Average	CCTST
CAT	0.599*	0.527*	0.558*	0.345*	0.645*

* correlations significant, $p < .01$

The National Survey of Student Engagement (NSSE) is a widely used instrument designed to assess the types of activities students are engaged in as well as their perceptions of institutional emphasis and the institution's contribution to their learning [27]. The NSSE was administered together with the CAT instrument to a stratified random sample of seniors at TTU. Misty Cecil, one of our doctoral students working on the project, examined the relationship between relevant NSSE questions and student performance on the CAT instrument. Five items on the NSSE were significant predictors of performance on the CAT instrument (multiple $R = .49$, $p < .01$). These five items are listed in the table below. The negative relationship between CAT performance and the extent to which students felt that their college courses emphasized rote retention is particularly important and supports both the criterion validity and the construct validity of the CAT instrument [28].

Table 3. NSSE Questions Related to CAT Performance

NSSE Question	Beta Coefficient
(2a) Memorizing facts, ideas, or methods from your courses and readings so you can repeat them in pretty much the same form (negative relationship)	-.341 **
(3b) Number of books read on your own (not assigned) for personal enjoyment or academic enrichment	.277 **
(11e) Thinking critically and analytically	
(11m) Solving complex real-world problems	.244 **
(7h) Culminating Senior Experience (thesis, capstone course, project, comprehensive exam, etc.)	.231 *

* Significant at .01 level; ** Significant at .001 level

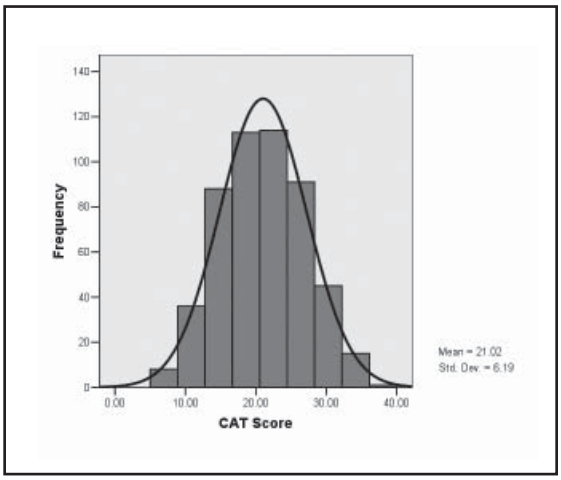
Several other key measures are reported in Table 4 below. Scoring reliability and internal consistency (calculated using Cronbach's alpha) are quite good for a test of this type and have increased as a result of test refinement. In addition, our latest measure of test-retest reliability has risen as a result of instrument improvements. The CAT instrument has also been found to be sensitive enough to assess changes between freshmen and seniors and to reveal the effects of a single course that includes components designed to improve critical thinking [29].

Table 4. Other CAT Statistical Findings

Scoring Reliability = 0.82
Internal Consistency = 0.695
Test–Retest Reliability > 0.80

Figure 2 shows the distribution of student scores (raw) on the CAT (version 4) instrument against the normal curve. Scores ranged from a low of 6 to a high of 36.3. There was no evidence of a floor effect or a ceiling effect (lowest possible score = 0, highest possible score = 40).

Figure 2. Distribution of Student Scores



Although more extensive analyses of any possible ethnic/racial/gender bias in the CAT instrument are currently being conducted, the preliminary analysis of available data is quite encouraging. A multiple regression analysis revealed that once the effects of entering SAT score and GPA, and whether English was the primary language (evaluated during year 2 testing) were taken into account, neither gender, race, nor ethnic background were significant predictors of overall CAT performance.

In addition to the quantitative survey data discussed above, qualitative data were collected from the local testing coordinators and the faculty scorers. Overall, the comments from both these groups were overwhelmingly positive. Many of the faculty members who participated in the CAT scoring workshops also exhibited increased interest in exploring methods for improving their students' critical thinking skills.

EXPLORATORY STUDY: PREDICTIVE VALIDITY

At the suggestion of our external consultant, John Bransford, from the LIFE Center at the University of Washington, Seattle, we began to examine ways in which the CAT instrument might be combined with training to improve students' critical thinking skills. While the benefits of such work go far beyond the goals of the current grant, we began to explore this issue as an additional means to help establish the validity of the CAT instrument. Specifically, the predictive validity of the test could be supported if students could be trained to improve performance on the CAT instrument, and if this training improved their performance on other relevant critical thinking tasks.

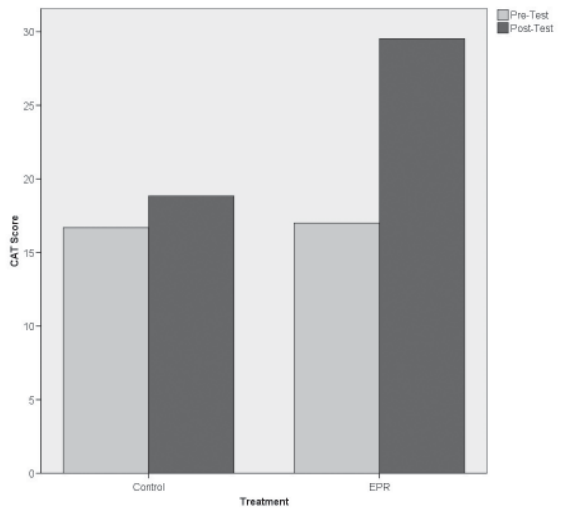
One existing methodology that seemed to hold promise for facilitating training using the CAT instrument was Calibrated Peer Review™ (CPR). Developed at UCLA with NSF support, the CPR system involves a computer network that provides opportunities for students to learn how to evaluate essay-writing assignments through a process of calibrating their evaluations to expert evaluations of the same essays. These exercises afford numerous opportunities to learn and critically evaluate ideas [30].

The constraints surrounding the use of the CAT instrument and the opportunities for implementing special training necessitated our modifying the CPR process and adapting it to this new situation. Training was conducted in group sessions with numerous opportunities for formative feedback. In the pilot study, Theresa Ennis, one of our doctoral students working on the project, trained a small group of students to score the CAT instrument using a procedure we call Enhanced Peer Review (EPR). The training used the same detailed scoring guide that had been developed for faculty scoring workshops. Students were initially given a small sample of test responses that reflected a range of scores for each question on the CAT instrument. During training students calibrated their evaluations of tests with those of the faculty graders. The training afforded numerous opportunities to explore the

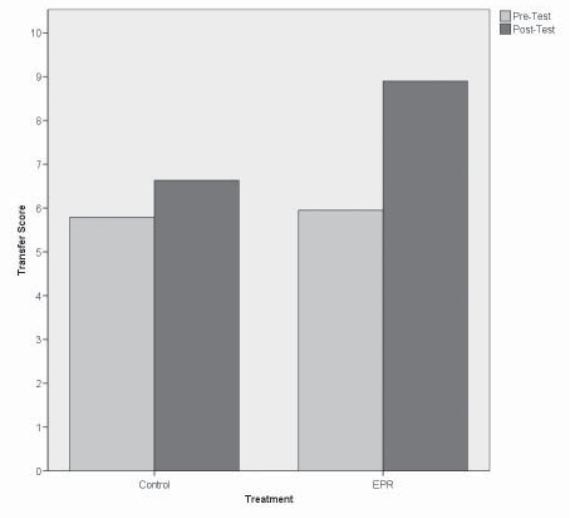
rationale for assigning scores to each response on the test and to educate students about a variety of issues related to critical thinking. After two three-hour training sessions, students were given the opportunity to score numerous CAT tests without any further training. The results of the pilot study showed that students could be taught to reliably score the CAT test using the EPR procedure.

We also compared students in the EPR pilot study to a control group. Both groups took a pre-test and post-test that included questions from the CAT instrument together with a new set of analogous transfer questions. Students who participated in the EPR training not only improved significantly more than the control group on the CAT questions they were trained on, they also improved significantly more than the control group on transfer questions (see Figure 3).

Figure 3. Performance on CAT Instrument & Transfer Test Questions



CAT Instrument



Transfer Questions

Student comments from a follow-up interview a year after the testing indicate lasting effects of the EPR training (e.g., "I'm a lot more critical about what I'm going to be listening to or believing in"; "I don't just look at one side and say this is definitely the way it is… I explore more sides and try to find alternative answers"; and "It helped me work through a personal serious life episode.").

We are currently conducting further studies of EPR using the CAT instrument and analog transfer questions. The pilot work done with training students to score the CAT instrument has also provided insights into better methods for training faculty to score the CAT instrument that could be incorporated into future workshops to train scoring workshop leaders at other institutions.

BROADER IMPLICATIONS

While there is broad agreement among faculty, administrators, educational experts, governmental officials, and employers that critical thinking skills are very important, efforts to improve students' critical thinking skills are currently hindered by the absence of an effective tool for assessing critical thinking skills that is valid, reliable,

and culturally fair. The CAT instrument is an innovative tool for assessing critical thinking that meets these criteria and, in addition, is designed to maximize faculty engagement in the assessment process and help motivate instructional improvement. These features of the instrument can help close the loop between assessment and the implementation of improvement initiatives. The CAT instrument is appropriate for both STEM and non-STEM disciplines. The interdisciplinary nature of the instrument is compatible with focused or broad-based institutional efforts to improve critical thinking across disciplines.

The CAT instrument may also be a useful assessment tool that can be used to strengthen other research projects designed to assess or improve student learning. We feel that combining the CAT instrument with initiatives that encourage the use of best practices to improve students' critical thinking and real-world problem solving to improve student learning can greatly improve undergraduate education in both STEM and non-STEM disciplines.

ACKNOWLEDGEMENTS

Partial support for this work was provided by the National Science Foundation's CCLI Program under grant 0404911.

REFERENCES

[1] Facione, P.A., Facione, N.C., Sanchez, C., and Gainen, J., "The disposition toward critical thinking," *Journal of General Education,* vol. 44, no.1, ERIC No. EJ499944, pp. 1-25, 1995.

[2] U.S. Congress, Goals 2000: *Educate America Act, H.R. 1804*, (U.S. Congress, Washington, D.C.), http://www.ed.gov/legislation/GOALS2000/TheAct/index.html, 1994.

[3] Bok, D., Our Underachieving Colleges: *A candid look at how much students learn and why they should be learning more,* (Princeton University Press, Princeton, N.J.), 2006.

[4] Bransford, J. D., Brown, A., L., and Cocking, R. R., (eds.), *How People Learn: Brain, Mind, Experience, and School*, (National Academy Press, Washington, D.C.), 2000.

[5] Ennis, R., "A Logical Basis for Measuring Critical Thinking Skills," *Educational Leadership*, vol. 43, no.2, pp. 44-48, 1985.

[6] Paul, R., and Nosich, G., *A Model for the National Assessment of Higher Order Thinking*, (National Center for Educational Statistics, Washington, D.C.), 1992.

[7] Pawlowski, D.R., and Danielson, M.A., "Critical Thinking in the Basic Course: Are We Meeting the Needs of the Core, the Mission, and the Students?" Paper Presented at the Annual Meeting of the National Communication Association, 1998.

[8] Resnick, L.B., *Education and Learning to Think,* Committee on Mathematics, Science, and Technology Education, Commission on Behavioral and Social Sciences and Education, National Research Council, (National Academy Press, Washington, D.C.), http://www.nap.edu, 1987.

[9] Siegel, H., *Education Reason: Rationality, Critical Thinking, and Education*, (Routledge, New York), 1988.

[10] Vygotsky, L. S., *Thought and language,* (MIT Press, Cambridge, Mass.), 1986.

[11] Wenglinsky, H., "Facts or critical thinking skills? What NAEP results say," *Educational Leandership*, vol. 62, no.1, pp. 32-35, 2004.

[12] Conley, D., A report from standards for success, Center for Educational Policy Research, http://www.s4s.org/cepr.s4s.php, 2003.

[13] See reference 3.

[14] Halpern, D. E., "Assessing the Effectiveness of Critical-thinking Instruction," *Journal of General Education*, vol. 42, no.4, pp. 238-254, 1993.

[15] Duchesne, R. E., "Critical Thinking, Developmental Learning, and Adaptive Flexibility in Organizational Leaders," paper presented at the Annual Meeting of the American Association for Adult and Continuing Education in Charlotte, N.C., 1996.

[16] Stein, B., Haynes, A., and Unterstein, J., "Assessing Critical Thinking," paper presented at the annual meeting of the Council on Colleges, Southern Association of Colleges and Schools, Atlanta, Ga. http://www.tntech.edu/cat/sacs2004stein.ppt, 2004.

[17] John J. Heldrich Center for Workforce Development, *Survey of New Jersey Employers to Assess the Ability of Higher Education Institutions to Prepare Students for Employment*, ERIC No. ED485290, 2005.

[18] Castellini, R., "Survey: More to see pay increase in 2006," http://www.cnn.com/2006/US/Careers/01/04/cb.aol.survey/index.html, 2006.

[19] Partnership for 21st Century Skills, *Learning for the 21st Century*, http://www.21stcenturyskills.org/images/stories/otherdocs/p21up_Report.pdf, 2002.

[20] U.S. Department of Education, National Center for Education Statistics, *The NPEC Sourcebook on Assessment, Volume 1: Definitions and Assessment Methods for Critical Thinking, Problem Solving, and Writing*, NCES 2000—172, Erwin, T.D. (ed.), for the Council of the National Postsecondary Education Cooperative Student Outcomes Pilot Working Group: Cognitive and Intellectual Development, (U.S. Government Printing Office, Washington, D.C.), 2000.

[21] Campione, J.C., and Brown, A.L., "Guided learning and transfer: Implications for approaches to assessment," in Frederiksen, N., and Glaser, R., et al. (eds.), *Diagnostic monitoring of skill and knowledge acquisition,* pp. 141-172, (Erlbaum, Hillsdale, N.J.), 1990.

[22] Feuerstein, R., *The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques,* (University Park Press, Baltimore), 1979.

[23] Lidz, C. S., *Dynamic assessment: An interactional approach to valuating learning potential,* (Guilford Press, New York), 1987.

[24] Samuels, M.T., "Assessment of post-secondary students with learning difficulties: Using dynamic assessment in a problem-solving process," In Lidz, C.S., and Elliott, J.G. (eds.), *Dynamic assessment: Prevailing models and applications,* pp. 521-542), (JAI/Elsevier Science, Amsterdam), 2000.

[25] Sternberg, R.J., and Grigorenko, E.L., *Dynamic testing: The nature and measurement of learning potential*, (University of Cambridge, Cambridge), 2002.

[26] Bloom, B.S., (ed.), *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain,* (Longmans, New York), 1956.

[27] Indiana University Center for Postsecondary Research, *National Survey of Student Engagement 2005 Annual Report: Exploring different dimensions of student engagement*, http://nsse.iub.edu/NSSE_2005_Annual_Report/index.cfm, 2005.

[28] Cecil, M.J., *The relationship between student responses on the National Survey of Student Engagement (NSSE) and performance on the Critical-thinking Assessment Test (CAT),* doctoral dissertation, Tennessee Technological University, 2006.

[29] Stein, B., Haynes, A., and Ennis, T., "Assessing Critical Thinking," paper presented at the Commission on Colleges of the Southern Association of Colleges and Schools Annual Meeting in Atlanta, Ga. in December 2005, http://www.tntech.edu/cat/sacs2005Stein.ppt, 2005.

[30] Chapman, O.L., and Fiore, M.A., "Calibrated Peer Review™," *Journal of Interactive Instruction Development*, vol. 12, no.3, pp. 1-15, 2000.

Developing an Outcomes Assessment Instrument for Identifying Engineering Student Misconceptions in Thermal and Transport Sciences

Ruth A. Streveler* - *Purdue University*
Ronald L. Miller - *Colorado School of Mines*
Mary A. Nelson - *University of Colorado - Boulder*
Monica R. Geist - *University of Northern Colorado*
Barbara M. Olds - *Colorado School of Mines*

OVERVIEW

This paper describes the development of the Thermal and Transport Concept Inventory (TTCI), which is designed to measure the presence of misconceptions held by undergraduate engineering students regarding important concepts in thermodynamics, fluid mechanics, and heat transfer. Concepts included in the inventory were identified using a Delphi study with experienced engineering faculty. Instrument items were developed using student responses to open-ended questions followed by beta testing of multiple-choice items using student language to create plausible, but wrong, answers. Psychometric analysis of beta test data collected at six engineering schools shows reasonable reliability among related items in the fluids, heat transfer, and thermodynamics portions of the instrument at acceptable levels of difficulty and discrimination.

INTRODUCTION

Evidence in the literature suggests that science and engineering students do not conceptually understand many fundamental molecular-level and atomic-level phenomena including heat, light, diffusion, chemical reactions, and electricity [1,2]. What is lacking is the students' ability to understand basic concepts in these fields. Although the word "concept" is a somewhat fuzzy term, we use it here to mean fundamental knowledge about how a process works — knowledge that goes beyond simply being able to perform the correct calculations [3]. Our research indicates that the problem is more than simply one of confusion or misunderstanding, but instead involves incorrect mental models, which we call misconceptions, by students about differences in the way that molecular-scale processes differ from observable, macroscopic causal behavior we experience in our daily lives [4]. Thus, students who can correctly apply macroscopic models of thermal or transport systems to solve problems in fluid dynamics, heat transfer, mass transfer, or thermodynamics often still believe that "heat flows from hot objects to cold objects" or that "molecular processes stop when they reach equilibrium." Any faculty member who has taught these subjects to engineering students will have heard similar comments.

For many traditional processes, the macroscopic models and metaphors ("heat flows") still work well and students must still be proficient in their use. However, they must now also understand when these models will break down and when the metaphors are no longer applicable to describe molecular-scale processes such as the ones listed above. Unfortunately, recent research suggests that students will persist in incorrectly applying macroscopic causal models to processes in which dynamic, molecular behavior dominates [4].

Why should student misconceptions be important to engineering educators? Simply put, prior knowledge contributes to the conceptual framework within which students will add new knowledge to form an updated mental model that describes each new learned concept (see [5], chapters 1, 3, and 10). If the framework contains

misconceptions, new topics will be difficult or impossible for the student to conceptualize correctly. The importance of assessing students' prior knowledge to improve learning has been recognized by assessment expert Thomas Angelo [6] and by the American Association for the Advancement of Science ([7], p. 198). In order to repair students' flawed mental models of thermal and transport processes, engineering educators must first determine what misconceptions exist and then discover as much about the nature of the misconceptions as possible.

Our goal in this project is to provide educators with a tool with which to identify these misconceptions. The instrument can be used as an outcomes assessment instrumental at the end of a program of study, or to determine learning gains by looking at differences in pre-test and post-test scores. We have based our assessment tool development on recent misconception work by Chi and colleagues that suggests the misapplication of macroscopic causal models as the primary source of student misconceptions when describing molecular-scale processes. We have patterned our instrument development after the highly successful Force Concept Inventory (FCI) developed for assessing Newtonian mechanics misconceptions in introductory college physics [8].

PROJECT OBJECTIVES AND EXPECTED SIGNIFICANCE TO ENGINEERING EDUCATION

Based on the overall goal of this project to develop an easy-to-use outcomes assessment instrument that will allow engineering faculty at the course and program levels to identify fundamental student misconceptions in thermal and transport sciences, we proposed to complete the project objectives listed below:

1) Develop a list of the most important student misconceptions in thermal and transport sciences by surveying experienced engineering faculty about important concepts that their students find difficult, and then validating this list through student interviews

2) Create a multiple-choice, pencil-and-paper instrument patterned after successful misconception instruments such as the Force Concept Inventory; misconceptions to be included in instrument questions would be based on the list created to meet the first objective

3) Field test the instrument to demonstrate validity and reliability of the misconception results obtained and usefulness of the instrument for both course-level and program-level assessment of student misconceptions in thermal and transport science topics

OBJECTIVE 1: IDENTIFYING IMPORTANT AND DIFFICULT CONCEPTS

As reported in more detail elsewhere [9], the project employed Delphi methodology to create a list of important but difficult concepts in thermal and transport science. The Delphi method is based on a structured process for collecting and distilling knowledge from a group of experts by means of a series of questionnaires interspersed with controlled opinion feedback [10]. In our project, 30 invited panelists (all experienced engineering professors from throughout North America) were asked to rate a list of 28 concepts, which the group itself had generated, on two scales: importance and difficulty. After iterating three times as prescribed by Delphi methodology [11], stable ratings were reached (see Table 1). These results were then used to create a list of the most important and difficult concepts in thermal and transport sciences.

Table 1. Results of Thermal and Transport Concepts Delphi Study[1]

Concept	"Understanding" Data Median (interquartile range)			"Importance" Data Median (interquartile range)		
	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
1.Adiabatic vs. Isothermal Processes	7.5 (6-8)	8 (6-8)	8(6.75-8.25)	9 (8-10)	9 (9-10)	9 (9-10)
2.Bernoulli Equation	7 (4-8)	6 (5-7)	6 (5-7)	9 (7-10)	9 (8-9)	9 (8-9)
3. Compressible vs. Incompressible Flow	5 (3-7)	6 (4-6.5)	6 (5-7)	7.5 (6-8)	7 (7-8)	7.5 (7-8)
4. Conservation of Linear Momentum	5 (3-6)	5 (4-6)	5.5 (5-6)	9 (8-10)	9 (8-10)	9(8-9.25)
5. Differential vs. Integral Analysis	4.5 (3-6)	4 (3-5.25)	4 (4-5)	7 (6-9)	8 (6-8)	8 (7-9)
6. Dimensional Analysis	6 (4-7)	5.5 (4.25-7)	6 (5-6.25)	7 (5-7)	6 (5-8)	7 (5-8)
7. Entropy & 2nd Law of Thermodynamics	4 (2-6)	4 (3-5)	5 (3-5.25)	8 (7-9)	9 (8-9)	9 (8-10)
8. Extensive and Intensive Properties	8 (6-9)	8 (7-8)	8 (7-9)	7 (6-9)	8 (7-9)	8 (7-9)
9. First Law of Thermodynamics	8 (7-9)	8 (7-9)	8 (8-9)	10 (10-10)	10 (10-10)	10 (10-10)
10. Fluid vs. Flow Properties	7 (5-8)	6 (5-7)	6 (5-6)	7 (5-9)	7 (5-8)	7 (5-8)
11. Heat Transfer Modes	8 (6-9)	8 (6.25-8)	8 (7-9)	9 (8-10)	9 (9-10)	9 (9-10)
12. Heat vs. Energy	6 (5-8)	6 (5-7)	6.5 (5-7)	9 (8-10)	9 (8-10)	9 (8-10)
13. Heat vs. Temperature	6 (4-8)	6.5 (5-8)	7 (6-8)	9 (8-10)	10 (9-10)	10 (9-10)
14. Ideal Gas Law	8 (7-9)	8 (8-9)	8 (8-9)	9 (8-10)	9 (9-10)	9 (9-10)
15. Internal Energy vs. Enthalpy	6 (3-7)	5 (4-6)	6 (5-6.25)	8 (7-9)	9 (8-9)	9 (8-9)
16. No-slip Boundary Conditions	8 (6-9)	8 (7-9)	8 (8-9)	8 (7-9)	9 (8-9)	9 (8-9)
17. Nozzles and Diffusers	6 (5-8)	6 (6-7.5)	7 (6-7)	7 (5-9)	7 (6-8)	7 (6-8)
18. Pressure	8 (6-9)	8 (7-8)	8 (7.75-9)	9 (8-10)	10 (9-10)	10(9.75-10)
19. Reversible vs. Irreversible Processes	5 (4-7)	5 (4-6)	5 (5-6)	8 (8-9)	9 (8-9)	9 (8-9)
20. Spatial Gradient of a Function	4 (3-7)	5 (4-6)	5 (4-5)	7 (3-9)	7 (6-8)	7 (6-8)
21. Specific Heat Capacity	7 (6-8)	7 (6-7)	7 (6-8)	8 (7-10)	9 (8-9)	9 (8-9)
22. Steady-state vs. Equilibrium Process	5 (3-8)	5 (3-6)	5 (4-5.25)	8 (5-10)	9 (7-9)	9 (8-9)
23. Steady-state vs. Unsteady-state Process	8 (7-8)	8 (7-8)	8 (7-8)	9 (8-10)	9.5 (9-10)	9.5 (9-10)
24. System vs. Control Volume	7 (4-8)	6 (5-7)	6 (6-7)	8 (6-10)	9 (8-10)	9 (8.5-10)
25. Temperature Scales	7 (5-9)	8 (8-9)	9 (8-9)	8 (6-10)	9 (8-10)	9 (9-10)
26. Thermal Radiation	6 (4-8)	5 (5-6)	5 (5-6)	7 (5-9)	8 (6.75-8)	8 (7-8.25)
27. Thermodynamic Cycles	7 (5-8)	7 (6-7)	7 (7-8)	8 (8-10)	9 (8-10)	9 (8-9.25)
28. Viscous Momentum Flux	5 (3-7)	4 (3.75-5)	4 (3-4)	7.5 (6-9)	8 (7-8)	7 (7-8)

"Understanding" Scale	"Importance" Scale
0 = no one understands the concept	0 = not at all important to understand the concept
10 = everyone understands the concept	10 = extremely important to understand the concept

[1]*Italicized* concepts are those the Delphi study identified as poorly understood but highly important

Based on these results we developed a list of 10 items that the study participants judged to be sufficiently important to engineering graduates that conceptual understanding is crucial but often not achieved. These 10 items include key topics such as the 2nd law of thermodynamics including reversible vs. irreversible processes, conservation of momentum, viscous momentum transfer, several energy-related topics (heat, temperature, enthalpy, internal energy), and steady-state vs. equilibrium processes.

OBJECTIVE 2: CREATING THE THERMAL AND TRANSPORT CONCEPT INVENTORY (TTCI)

After the list of most important and least understood concepts was identified, our next step was to create open-ended questions based on these concepts, which were then used in student interviews.

Using the ten concepts we identified through the Delphi study, we constructed sample questions and then conducted "think aloud" sessions with six Colorado School of Mines students (half chemical engineering and half mechanical engineering majors, two females and four males). We asked each of them to answer three or four questions, explaining their thought processes aloud as they did so. The students were volunteers who were selected on a first-come basis and compensated for their time. We followed a protocol during the sessions: first we gave the student a written copy of each question, and then read it aloud. We asked the student to tell us what concept s/he thought the question addressed and then asked him or her to solve the problem thinking aloud and using sketches, notes, etc., if desired. We asked questions intended to probe the students' thought processes as they solved the problems and tape-recorded the sessions for later transcription and analysis. We also collected all notes and equations the students produced. Based on these interviews, we hoped to accomplish two goals: 1) assure that the questions were clearly worded and illustrated so that students would not be confused by the framing of the questions, and 2) begin to develop distractors for our multiple-choice concept inventory.

As an example, Table 2 contains one of the questions that we asked students to solve using the "think aloud" method. Table 3 includes some of the dialogue from a student interview to give a flavor for the form that the interviews took. (We obtained informed consent from all student participants and have not used their real names in reporting on the interviews.)

Table 2. Sample "Think Aloud" Question

Question 6.A
A mass of air is contained in a rigid vessel at 100° C and 5 atm pressure. An equal mass of air is flowing at the same temperature and pressure through a circular pipe. Which of these samples of air can be used to produce more mechanical work?

Table 3. Excerpt from Student "Think Aloud" Interview

Interviewer: What are you looking at here if you were to describe the two things that you are being asked to look at? What are you looking at?
Diana: I'm not really sure, like, I'm just trying to think why both of them would have, uhm. I'm trying to think why one of them would produce more work than the other one. And in thinking about it I really don't think that both of them should produce different amounts of work.
Interviewer: Ok. So your initial thought on it would be that they would produce equal
Diana: Yeah.
Interviewer: Amounts of work. Ok. Why would you think that?
Diana: Well. Just because I don't think the work depends on the area in which the mass of air is contained.
Interviewer: Um-hum.
Diana: So, I mean if they're the same temperature and the same pressure and have the same volume so I just don't see why they would have a different amount of work.

After completing the "think aloud" interviews, we carefully reviewed the transcripts and then revised the questions to increase clarity and developed a set of distractors for each question using misconceptions discovered during the analysis of the sessions as well as, whenever possible, using the exact language of the student interviewees. If

students stated that they needed more information to answer a question, the question was rewritten to attempt to eliminate ambiguity. Table 4 contains the alpha test version of Question 6.A we developed using this process.

Table 4. Alpha Version of Question 6.A

Question 6.A
A mass of air is contained in a rigid vessel at 100 °C and 5 atm pressure. An equal mass of air is flowing at the same temperature and pressure through a circular pipe. Which of these samples of air can be used to produce more mechanical work? a. Both will produce the same amount of work because they are at the same temperature and pressure. b. Moving air will produce more work because it has kinetic energy and air in rigid vessel does not. c. Air in rigid vessel will produce more work because moving air will lose energy as friction dissipation. d. Moving air will produce more work because it contains flow work (pressure-volume work) and air in the rigid vessel does not. (Correct answer = d)

OBJECTIVE 3: FIELD-TESTING THE TTCI TO DEMONSTRATE VALIDITY AND RELIABILITY

In the fall of 2003, we administered an alpha version of the concept inventory consisting of 11 multiple choice questions to 93 students in two classes at the Colorado School of Mines—39 students in a senior-level chemical engineering course in transport phenomena and 54 students in a senior-level integrated laboratory course designed for students with a specialty in mechanical engineering. All of the students were seniors who had taken at least one course in thermodynamics, heat transfer, and fluids. (For more information about the results of the alpha test, and a list of questions used, see [12]).

After reviewing the results of alpha testing, at least two more questions had been written for each concept, and additional think-alouds conducted. Thirty-three concept questions were beta tested. Because of the complexity of many of the questions, the TTCI was administered in three sections: thermodynamics, heat transfer, and fluid mechanics. In that way, the questions were more targeted to specific classes and could be administered in a reasonable amount of time. Beta testing was then conducted at six universities, involving over 100 engineering students who were enrolled in or had completed appropriate coursework.

Two types of classical test statistics, item difficulty, and item discrimination, were used to determine if TTCI items were performing satisfactorily. Items with too low or too high difficulty, or with poor discrimination, were dropped from the TTCI. Remaining items were tested for reliability using Kuder-Richardson KR-20, a special case of Cronbach's alpha used when items are dichotomous. KR-20 values for the three sections of the TTCI (thermodynamics, fluid mechanics, and heat transfer) ranged between 0.56 and 0.66. A KR-20 value of 0.70 is generally thought to be acceptable, so the TTCI is approaching acceptable reliability. Continued refinement of the TTCI items is planned, and a future web-based version of the instrument will allow for easier beta testing by larger numbers of students. Revised items and a larger sample size should allow the TTCI to reach the acceptable levels of reliability needed for wide dissemination.

In addition to calculating reliability measures, we also tested to see if students were guessing on the TTCI. Some of the TTCI items are two-part questions, with one part asking what happens in a situation, and the second part asking why that happens. To determine if students were guessing, we calculated the correlation between these two-part questions, reasoning that for every "what happens" scenario there was a logical "why it happens"

response. We found that the correlation between the two-part questions ranged between 0.72 and 0.94, which we feel is evidence that students took the test seriously.

Our team has attempted to establish content and construct validity for the TTCI by a variety of means: expert review (by Delphi panelists and other content experts), literature review, and think-alouds by students working through the open-ended questions that eventually became TTCI items. The triangulation of these results allows us to be confident that the TTCI demonstrates content and construct validity. Thus, the TTCI is well on its way to becoming a valid, reliable instrument which can be widely disseminated to measure engineering student misconceptions in the thermal and transport sciences. For more detailed information about reliability and validity measures of the TTCI, see [13].

FUTURE DIRECTIONS

Based on the performance results of the current version of the TTCI and the need for additional questions and student response data, we have developed a new version of the instrument (version 3.0), which is now undergoing beta testing. A total of 24 new fluid mechanics, thermodynamics, and heat transfer items have been added focusing on key misconceptions that our results suggest are particularly robust among engineering students.

REFERENCES

[1] Reiner, M., Slotta, J.D., Chi, M.T.H., and Resnick, L.B., "Naive physics reasoning: A commitment to substance-based conceptions," *Cognition and Instruction*, vol. 18, pp. 1-34, 2000.

[2] Garnett, P.J., and Hackling, M., "Students' alternative conceptions of chemistry: A review of research and implications for teaching and learning," *Studies in Science Education*, vol. 25, p. 69, 1995.

[3] Case, J. M., and Fraser, D. M., "The challenges of promoting and assessing conceptual understanding in chemical engineering, *Chemical Engineering Education*, vol. 36, no. 1, pp. 42-47, 2002.

[4] Chi, M.T.H., "Commonsense conceptions of emergent processes: Why some misconceptions are robust," *Journal of the Learning Sciences,* vol. 14, no. 2, pp. 161-199, 2005.

[5], Bransford, J.D., Brown, A.L, and Cocking, R.R., (eds.), *How people learn*, (National Academy Press, Washington, D.C.), 2000.

[6] Angelo, T., "Doing assessment as if learning matters most: Guidelines from research and best practice," keynote address presented at the Best Assessment Processes IV Conference, Rose-Hulman Institute of Technology, Terre Haute, Ind., April 2001.

[7] *Science for All Americans: Project 2061*, (Oxford University Press, London), 1990.

[8] Hestenes, D., Wells, M., and Swackhamer, G., "Force concept inventory," *The Physics Teacher*, vol. 30, pp. 159-166, 1992.

[9] Streveler, R.A., Olds, B.M., Miller, R. L., and Nelson, M.A., "Using a delphi study to identify the most difficult concepts for students to master in thermal and transport science, *Proceedings of the American Society for Engineering Education Annual Conference and Exposition*, Nashville, Tenn., June 2003.

[10] Adler, M., and Ziglio, E., (eds.), *Gazing into the oracle: The delphi method and its application to social policy and public health*, (Jessica Kingsley Publishers, London), 1996.

[11] Linstone, H. A., and Turoff, M., (eds.), *The delphi method: Techniques and applications,* (Addison-Wesley, Reading, Mass.), 1975.

[12] Olds, B. M., Streveler, R. A., Miller, R. L., and Nelson, M. A., "Preliminary results from the development of a concept inventory in thermal and transport science," *Proceedings of the American Society for Engineering Education Annual Conference and Exposition*, Salt Lake City, Utah, June 2004.

[13] Streveler, R.A., Miller, R.L., Nelson, M. A., Geist, M. R., and Olds, B. M., "Development and psychometric analysis of the thermal and transport concept inventory," *Journal of Engineering Education,* under review.

The Signals and Systems Concept Inventory

Kathleen E. Wage* - George Mason University
John R. Buck - University of Massachusetts Dartmouth
Margret A. Hjalmarson - George Mason University

INTRODUCTION

The *Signals and Systems Concept Inventory* (SSCI) originated with our desire to quantitatively assess the changes in student learning when we adopted active and cooperative learning (ACL) in our undergraduate Signals and Systems (S&S) courses. These courses are a staple of electrical engineering curricula and are generally taught in the sophomore or junior year. Our interest in ACL for these upper division courses arose when UMass Dartmouth's participation in the NSF-funded Foundation Coalition (FC) [1] prompted a redesign of the integrated freshman year curriculum [2]. The discussions leading to this curriculum revision exposed us to the rich body of physics pedagogical research supporting the use of ACL [3,4]. Intrigued, we incorporated ACL methods into our S&S courses in the late 1990s. Our pedagogical instincts and student evaluations suggested that students were learning more when we used ACL, but we coveted the quantitative confirmation that Hestenes et al.'s *Force Concept Inventory* (FCI) provided for the physics reformers [5]. In the absence of a comparable instrument to the FCI for S&S, we set out to write one.

Our interest in writing the SSCI coincided with a new initiative in fall 2000 within the FC to promote concept inventories for upper division engineering courses. The FC funded our initial development of the SSCI during late 2000, and version 1.0 of the continuous-time (CT) SSCI was tested at UMass Dartmouth (UMD) and George Mason University (GMU) during the spring of 2001. The data obtained during these tests guided our revision of the CT instrument to produce version 2.0 and also the first version of the discrete-time (DT) SSCI. In addition to UMD and GMU, the revised instruments were tested at MIT, Old Dominion University, Rose-Hulman Institute of Technology, the U.S. Air Force Academy, and the U.S. Naval Academy during the 2001 to 2003 period. By the expiration of the FC seed funding in May 2003, the SSCI had been administered to more than 600 students by a dozen instructors. In September of 2005, the NSF Division of Undergraduate Education awarded us additional funding under the Assessment of Student Achievement (ASA) program to refine, validate, and then disseminate the SSCI. This new project includes a development team of twelve S&S faculty using the SSCI and providing feedback to the developers, described in more detail below. At the time of this writing, more than 30 faculty have administered the SSCI to over 1,400 students.

This paper collects and synthesizes information about the SSCI previously published in a number of different journals and conference proceedings [6,7,8]. The target audience of this paper is STEM faculty interested in developing new concept inventories for their own core classes, whereas the publications cited above are targeted at electrical engineering faculty teaching S&S courses. We hope to provide a useful perspective on the development, testing, and application of a concept inventory for an upper-division core course. This paper also presents new material on our expert peer review process to establish content validity. It also updates our data from employing the SSCI to quantify gains in students' conceptual understanding due to instruction, and to identify stubborn misconceptions held by students.

THE SIGNALS AND SYSTEMS CONCEPT INVENTORY INSTRUMENT

Signals and Systems (S&S) courses are nearly universal core requirements in the second or third year of electrical engineering curricula. S&S courses teach students to develop abstract mathematical models to represent linear and time-invariant systems, and then to use these models to understand how such systems behave. Many students struggle with this level of abstraction. They find it hard to master using the same techniques to model, for example, a circuit, a shock absorber, and a loudspeaker system. Moreover, mastering the techniques taught in S&S courses requires fluency with calculus and differential equations.

In developing the SSCI, we identified five key concepts that all S&S students need to master. These concepts are linearity and time-invariance, convolution, filtering, transform representations (e.g., Laplace and Fourier transforms), and background mathematics. Additionally, sampling is a key concept in any DT S&S course. There are separate CT and DT versions of the SSCI. Each consists of 25 multiple-choice questions covering the concepts listed above. The incorrect answers, or distractors, for each question capture common student misunderstandings, which we label misconceptions. It is worth noting that the term *misconception* can have different connotations in the science education literature. To physics educators such as Scherr, a misconception represents "a coherent framework of ideas that are stably present in students minds and present obstacles to instruction" [9]. Modell et al. use the term rather broadly to mean "conceptual and reasoning difficulties", though they also comment that the vocabulary used in the literature on misconceptions can be confusing since different labels are used to describe similar ideas [10]. Building on the work of diSessa [11], Nasr et al. think of a concept as being composed of a set of knowledge building blocks or "primitives", and they indicate that misconceptions occur when students combine these primitives in the wrong way or apply a primitive in the wrong context [12,13]. The papers by Nasr et al. are particularly relevant to the SSCI because they focus on conceptual understanding of signals and systems topics, albeit in the context of aeronautical (rather than electrical) engineering. In this article, we use the term misconception to mean incorrect or incomplete understanding, similar to both Modell and Nasr.

One of the challenges in developing a concept inventory for abstract material such as S&S is to develop good questions that probe students' conceptual understanding, and not simply their ability to carry out rote computations. Given the high level of mathematical sophistication required for S&S, some colleagues were skeptical whether it was possible to write conceptual questions in the same spirit as the FCI. Our best definition of a conceptual question is that when the student understands the concept being tested, they can choose the correct answer without computing anything. A good conceptual question contains few numbers, so that students who do not understand the concept have nothing to plug into memorized formulae. The journal article [8] describing the SSCI's development contains several sample questions from the CT and DT exams.

An important component of the expanded SSCI project under the NSF ASA program is the development team. This team consists of S&S instructors from ten schools in addition to the PIs' schools: Binghamton Univ., Duke Univ., Embry-Riddle Aeronautical Univ., Marquette Univ., Notre Dame, Rice Univ., Rose-Hulman Inst. of Tech., Univ. of California Berkeley, Univ. of Texas-El Paso, and Univ. of Wyoming. These instructors employ the SSCI in their classes in a pre-test/post-test protocol. They provide this data, along with linked demographic and academic data, to the SSCI project. The linked exam scores and demographic data will allow us to establish construct validity for the SSCI. The exam data from this pool of schools will provide an important baseline on student performance on the SSCI, and help us to characterize common and persistent student misconceptions. The next section presents highlights of the data analysis to date. The development team also meets annually to provide feedback to the SSCI

authors on the exam questions, as well as proposed new questions or distractors. These meetings provide expert review to establish content validity for the SSCI, as described below.

OVERVIEW OF SSCI RESULTS

Concept Inventories (CIs) can assess student achievement in several ways. First, CIs can measure students' gain in understanding in a class when comparing average scores for the same test administered at the beginning and end of a course (pre-test and post-test). Second, detailed analysis of answers to the multiple-choice questions provides valuable insights into student misconceptions. Third, correlating CI scores with other factors, such as grades in prerequisite courses, quantifies how students' prior preparation affects their ability to learn new material. Finally, CI's provide a fertile starting point for interviews probing student understanding and misconceptions. To encourage STEM faculty developing CI's to pursue these research directions, this section illustrates the first two of these applications and describes the interviews using examples from the SSCI study. For an example of correlation analysis, see the recent journal article [8].

As noted in the introduction, research by the physics community on the effectiveness of ACL methods motivated the development of the SSCI. Consequently, our analysis of student learning gains as a function of instructional mode derives from Hake's compelling comparison of the learning gains for traditional physics instruction versus ACL-based instruction. As a metric for student learning, Hake defines the normalized gain *(g)*:

$$\langle g \rangle = \frac{post\text{-}pre}{100\text{-}pre}, \qquad (1)$$

where the pre-test and post-test values are the averages for the course, computed using the set of students who took both tests. Based on an analysis of 62 Newtonian physics courses, Hake concludes that the 14 traditional lecture courses achieved normalized gain *(g)*=0.22 ± 0.04, while 48 ACL courses achieved a significantly higher *(g)*=0.48 ± 0.14. The SSCI study reveals a similar pattern in the gain for traditional and ACL courses. The average gain for 16 traditional S&S courses is *(g)*=0.22 ± 0.07 and the average gain for 16 ACL S&S courses is *(g)*=0.39 ± 0.06. Figure 1 displays the gain data for the 32 courses in the SSCI study using the format suggested by Hake. The plot shows the raw difference between the pre-tests and post-tests versus pre-test score. Note that the ACL courses cluster in the medium gain region, which Hake defined as gains between 0.3 and 0.7.

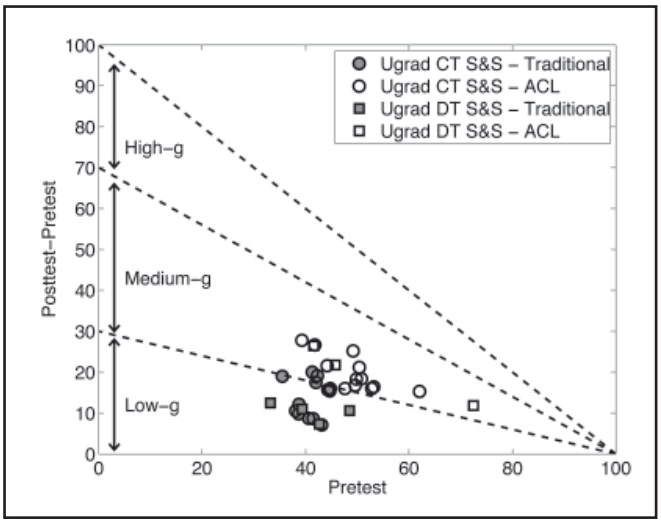


Figure 1. Comparison of raw gain versus pre-test score for 32 S&S courses. Each point represents a single course. The abscissa is the pre-test score and the ordinate is the raw gain, defined as the post-test average minus the pre-test average for the course. The low, medium, and high gain regions are those defined by Hake [4]. 16 of the 32 courses employed ACL methods.

While gain statistics are an indicator of students' overall performance, computing the difficulty index for a CI provides more detailed information about which concepts are the hardest for students to master. The difficulty index is defined as the percentage of students answering a question correctly. Figure 2 shows the difficulty index for the CT-SSCI computed using pre-test and post-test data from a pool of 445 students. The labels on the figure highlight the results for selected concepts. Students perform best on the four questions related to background mathematics. This probably reflects the numerous opportunities students have to practice their basic math skills during any S&S course. Figure 2 also indicates results for questions on core SSCI topics: time/frequency relationships (i.e., transform representations), convolution, and filtering. While there is substantial gain between pre and post for these three questions, it is disconcerting that, even on the post-test, less than 60 percent of students are responding correctly.

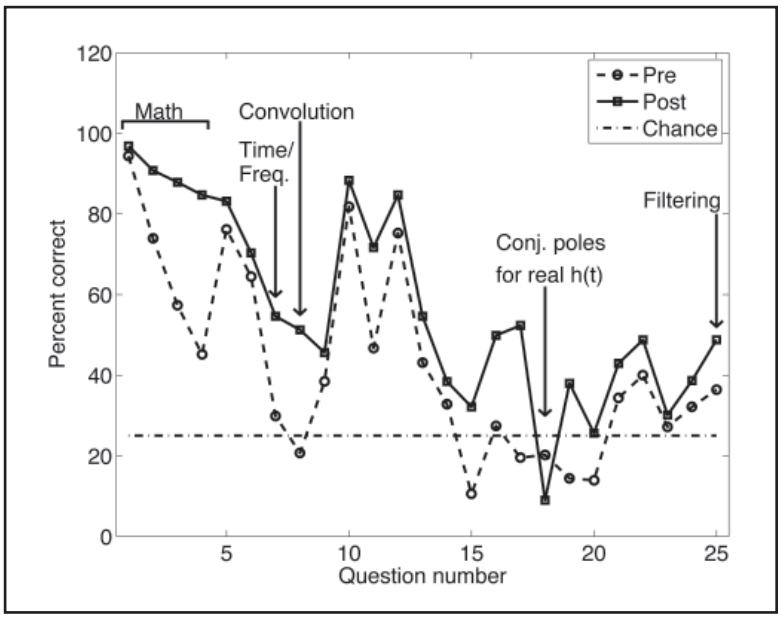


Figure 2. Pre-test and post-test difficulty index for the CT-SSCI based on data from 445 students. The dashed line at 25 percent indicates the expected result for random guessing with four possible choices.

Comparison of the pre and post difficulty indexes also identifies questions for which student performance is worse at the end of the semester than at the beginning, perhaps indicating that a little knowledge can be a dangerous thing. For instance, note that Figure 2 shows that fewer students answer Question 18 correctly on the post-test than the pre-test. They appear to be guessing on the pre-test since the percentage correct is just below the 25 percent predicted by chance. Analysis of the post-test answers indicates that most students are choosing a distractor that is partially correct. Question 18 displays a set of pole-zero plots and asks students which could correspond to real systems. Seventy-five percent of students choose the distractor indicating that they believe real systems must have all real poles and zeros, neglecting that systems with complex conjugate poles and zeros can also be real. The SSCI results indicate that this misconception is very persistent, meaning that it is resistant to instruction. We define the persistence of a distractor by the fraction of the students who choose that same distractor on both the pre-test and post-test [8]. The main distractor for Question 18 is chosen by 42 percent of students on both the pre- and post-tests, in contrast to the 6.25 percent predicted by chance if students were only guessing on both the pre-test and post-test. This suggests that S&S courses do little to alter students' misconceptions about where the poles and zeros of a real system must lie. Several other distractors on the CT-SSCI are significantly persistent. Analysis of persistence is useful because it indicates where new explanations and exercises are most needed to help students overcome resistant misconceptions.

Quantitative analysis of multiple-choice test data from a CI does not provide a complete picture of students' thought processes. Conducting student interviews based on CI questions gives additional insight into how students approach the subject, and may reveal unanticipated misconceptions or mental models used by students. For these reasons, the SSCI project includes funding to interview students about their thought processes in

solving SSCI questions. The first round of interviews was conducted in spring 2006 [7], and additional interviews are ongoing. To date the interview questions have focused on concepts related to frequency-selective filtering. To understand filtering students must also have a good understanding of sinusoidal signals, linear time-invariant systems, and the relationship between time and frequency. The CT-SSCI includes four questions on these basic concepts. The interviews revealed that students find the Fourier transform to be difficult, and that they have misconceptions about the roles of the magnitude and phase of a system's frequency response. One unexpected outcome of the interviews was the discovery that the connotations of the everyday use of the term "filter" (e.g., coffee filter, spam filter) may limit students' concept of how filters work in S&S. There is also some indication that the heavy emphasis placed on ideal lowpass filters as examples in S&S courses may encourage students to over-generalize how filters work, leading to problems when they try to apply the same ideas in other contexts.

EXPERT REVIEW FOR CONTENT VALIDITY

As part of the current phase of SSCI development, we are holding day-long meetings with the development team consisting of faculty users of the SSCI. The first goal is to obtain feedback about the questions on the SSCI since there is variation among institutions and faculty in terms of course structure, course sequencing, and course content in S&S. The second goal of the meetings is to ensure the content validity of the concept inventory. We want to know if the CT and DT exams are covering the content experts in the field feel is important and if the content is covered accurately. At the first team meeting, eleven professors from different institutions attended (in addition to the authors). All of them are experienced S&S instructors. Most had already used the SSCI with their students, and the remainder had committed to using the SSCI in their classes during the 2006–2007 academic year. The development team members come from a variety of institutions, ranging from small, private universities, to large, public universities. All have an interest in the teaching aspects of their work, and a few are in primarily teaching positions.

In the first part of the meeting, we asked the development team to identify the three hardest concepts and the three most important concepts in S&S. After they compiled lists individually, we held a group discussion of the topics. The two most frequently identified as difficult concepts were convolution and the relationship between the poles of the system function and time domain behavior. More generally, these difficult concepts are related to students' understanding of functions and transformation of functions in the context of S&S. Under important concepts, faculty also identified convolution but included frequency and time relationships, as well as the examination of non-linear functions. As an example, properties and transformations of periodic functions play a significant role in S&S. Complex number representations, Fourier transforms, Fourier series, non-linearity, and linear time invariance were also highlighted. Overall, the main challenge for students is the application and integration of mathematical topics in the context of signals and systems. In the second part of the day-long meeting, the development team completed either the CT or DT SSCI in order to help them understand students' experience of the exam, and to identify any potential improvements in wording or formatting of the questions.

The expert review of the exam, as well as the identification of both difficult and important concepts, confirmed that the SSCI was covering topics important to the discipline. One aspect of the discussion focused on how important it is to include symbolic representations of functions in the questions or whether questions could be stated without symbolic representations. As identified by at least one student interviewee, the symbols can be a point of difficulty for students and their use may vary. The symbols used may add another layer of abstraction to an abstract topic. Another issue was whether discrete time signals were covered in parallel with continuous time signals or whether the material was covered in two separate courses. While there are relationships between the concepts, departments organize the sequencing differently.

CONCLUSION

To date, the SSCI project has obtained several important results. First, the analysis of gains on the SSCI reveals a striking similarity to Hake's results for the FCI. For a data set consisting of 32 S&S courses, we found that on average, students in ACL courses gain roughly twice as much conceptual understanding as those in traditional lecture courses ($< g > = 0.39$ vs. $< g > = 0.22$). Second, an analysis of students' answers highlighted several persistent misconceptions that students have about key S&S topics. Third, our pilot interview study found that students bring unexpected preconceptions about filtering into signals and systems courses, probably based on commonplace uses of the word "filter." Finally, the faculty development team has provided significant feedback that will be used to revise the SSCI exams in the upcoming year.

The SSCI project is an ongoing effort funded by NSF's ASA program through 2008. In addition to the types of analysis and peer review described in this paper, we are also investigating the reliability of the SSCI and checking for gender and racial bias using student data provided by the development team. The goal of the SSCI project is to provide faculty with a reliable, validated instrument that can be used for formative feedback, as well as for accreditation assessment. It is exciting to note that several instructors are currently using the SSCI to assess the effects of different instructional methods, such as project-based courses [14,15] and graphical vs. text-based programming methods [16].

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the National Science Foundation through grants DUE-0512686 and DUE-0512430 under the Assessment of Student Achievement program, as well as NSF's prior support for the initial development of the SSCI through grant EEC-9802942 to the Foundation Coalition.

REFERENCES

[1] Foundation coalition web site, http://www.foundationcoalition.org/index.html.

[2] Pendergrass, N.A., Laoulache, R. N., Dowd, J. P., and Kowalczyk, R.E., "Efficient development and implementation of an integrated first-year engineering curriculum," *Proc. of Frontiers in Education*, Tempe, Ariz., November 1998.

[3] Mazur, E., *Peer Instruction: A User's Manual,* (Prentice Hall, Englewood Cliffs, N.J.), 1997.

[4] Hake, R. R., "Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *American Journal of Physics*, vol. 66, pp. 64–74, 1998.

[5] Hestenes, D., Wells, M., and Swackhamer, G., "Force Concept Inventory," *The Physics Teacher,* vol. 30, March, pp. 141–158, 1992.

[6] Buck, J. R., and Wage, K. E., "Active and Cooperative Learning in Signal Processing Courses," *IEEE Signal Processing Magazine,* vol. 22, no. 2, pp. 76–81, 2005.

[7] Wage, K.E., Buck, J. R., and Hjalmarson, M. A., "Analyzing misconceptions using the signals and systems concept inventory and student interviews," *Proceedings of the Fourth IEEE Signal Processing Education Workshop*, Grand Teton National Park, Wyo., pp. 123–128, September 2006.

[8] Wage, K. E., Buck, J. R., Wright, C.H.G., and Welch, T. B., "The Signals and Systems Concept Inventory," *IEEE Trans. on Educ.*, vol. 48, pp. 448–461, August 2005.

[9] Scherr, R. E., "Modeling student thinking: An example from special relativity," *American Journal of Physics,* vol. 75, pp. 272–280, March 2007.

[10] Modell, H., Michael, J., and Wenderoth, M.P., "Helping the learner to learn: The role of uncovering misconceptions," *The American Biology Teacher,* vol. 67, pp. 20–26, January 2005.

[11] DiSessa, A. A., "Toward an epistemology of physics," *Cognition and Instruction*, vol. 10, pp. 105–225, 1993.

[12] Nasr, R., Hall, S. R., and Garik, P., "Student misconceptions in signals and systems and their origins," *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference,* Boulder, Co., pp. T2E23–T2E28, November 2003.

[13] Nasr, R., Hall, S. R., and Garik, P., "Student misconceptions in signals and systems and their origins – part ii," in *Proceedings of the 35th ASEE/IEEE Frontiers in Education Conference*, Indianapolis, Ind., pp. T4E26–T4E31, October 2005.

[14] Nelson, J.K., "Using project-based assignments in a graduate-level digital signal processing course," *Proceedings of the Fourth IEEE Signal Processing Workshop*, Grand Teton National Park, Wyo., pp. 135–140, September 2006.

[15] Padgett, W.T., "Teaching fixed-point algorithm development in a systems context," *Proceedings of the Fourth IEEE Signal Processing Workshop,* Grand Teton National Park, Wyo., pp. 297–301, September 2006.

[16] Yoder, M.A., and Black, B.A., "Teaching DSP first with LabVIEW," *Proceedings of the Fourth IEEE Signal Processing Workshop,* Grand Teton National Park, Wyo., pp. 278–280, September 2006.

Assessing Students' Value for Science and Math Literacy

Mark Wood*, Vickie Luttrell, Bruce Callen, Charles Allen, and Donald G. Deeds - *Drury University*

INTRODUCTION

The current national climate in higher education is marked by increasing demands for accountability by both public and private institutions [1]. If taken seriously, these demands can provide the impetus for developing meaningful quantitative measures of learning, and using those measures to enhance all areas of undergraduate education. In this paper, we report on one avenue of this research: an attempt to create simple and powerful assessment tools aimed at strengthening undergraduate science and math education, particularly for general education students. We have developed two instruments—the Science Value Inventory (SVI) and Math Value Inventory (MVI)—that measure the value students place on their science and math education. We plan to use these tools to build a national database with which universities can measure their progress toward shaping a scientifically literate public [2]. Our inventories will also provide significant feedback for universities initiating curricular reform, by measuring the success that particular pedagogies have on establishing students as life-long learners in science and math.

Our project grew out of the need to measure the success (or failure) of our integrated interdisciplinary curriculum, developed as part of a comprehensive reform of science and math education at Drury University. The success of curricular reform can only be established with valid measures of student outcomes, both to determine its effectiveness as well as to provide meaningful results for others to draw upon. Therefore, we initially attempted to measure student attitudes about science, math, and technology. These efforts revealed that following the completion of Drury's new curriculum, only 29 percent of students thought that mathematics and statistics were required for good citizenship. This was troubling, especially because one of our primary goals was to help students "develop an appreciation of the importance and relevance of mathematics and science to our history, our current status, and our global future" [2]. In response, we initiated a much more focused effort to understand the affective variables that enhance (or inhibit) student learning. Drawing on 50 years of research on motivation and learning, we came to appreciate the fundamental importance of the affective domain.

By adapting and extending well-developed models, we created the SVI and MVI to assess the "value" that general education students place on their undergraduate science and math courses. According to Rokeach [3], "[v]alues are core conceptions...that serve as standards or criteria to guide not only action but also judgment, choice, attitude, evaluation, argument, exhortation, rationalization, and, one might add, attribution of causality." More specifically, values mediate student decision making regarding pursuit of scholastic events and activities, such as course selection [4].

The SVI and the MVI were derived from models that draw on the foundational work of Rotter [5] and Atkinson [6], who proposed that students' expectancies for success and the inherent value they place on that success serve as mediators of achievement-related behavior. Grounded in the early work of Atkinson, Eccles and colleagues developed a comprehensive theory, proposing that students' academic performance, persistence, and scholastic choices are directly affected by their expectancy-related and task-value beliefs [7,8]. Eccles' model has been used extensively for understanding K-12 students' achievement in math and science, as well as for exploring gender-related differences in academic accomplishment and career choices (e.g., [9,10,11,12]). Eccles' work provided the theoretical foundation for our research, and served as the roadmap for the development of the SVI and MVI.

INSTRUMENT DEVELOPMENT

Over the past four years we have developed the SVI and MVI and explored their psychometric properties. Establishing that the results obtained from these (or any) inventories are valid and reliable is a multi-step process (see [13,14,15]). The first step is to provide evidence for content validity: that is, that the area of interest is identified completely and that candidate items adequately reflect the multiple aspects of the area of interest. A second step is to provide evidence for *factorial validity*: that in large-scale tryouts, students respond to the items as predicted by our theoretical framework. When this process produces a final version of the inventory, reliability studies must be conducted to ensure that student responses are stable over short periods of time. We have completed this work for both the SVI and MVI, as we describe in more detail below. The final stage of inventory development is construct validation, in which logical predictions following from the area of interest are tested (such as students who major in science will have higher SVI scores than those majoring in other subjects).

Item Generation and Content Validation

We conceptualized the value students place on becoming well educated in science and math as consisting of four interrelated domains, which we tentatively labeled *interest value, utility value, attainment value,* and *personal cost* [7,8]. After creating definitions for each domain, we constructed a pool of items to reflect each of the domains for both science and math. For both sets of items, we chose a five-point Likert-type format, with response options ranging from 1 (strongly disagree) to 5 (strongly agree).

For the SVI, we asked five experts in science education to evaluate the domain descriptions, whether these domains encompassed all facets of value, and the logical validity of each of the items. We selected five experts in math education to assist us with a similar evaluation of the MVI domains and associated items. For both inventories, we used expert recommendations to modify or eliminate certain items from the pool. As a final step before large-scale testing, we asked graduate students in measurement classes to review the SVI and MVI items for technical adequacy, clarity of meaning, and content.

Large-scale Item Tryouts

Following the content validation procedures, we conducted our first round of large-scale item tryouts of the 91-item SVI (N=1587 non-science majors, 60.4 percent female, 91.3 percent Caucasian) and the 70-item MVI (N = 944 non-math majors, 71.9 percent female, 90.1 percent Caucasian). We used these data to conduct exploratory item level, factor structure, and internal consistency analyses. Each item was evaluated for skewness, kurtosis, and inter-item correlations. Those items with non-normal distributions were eliminated, and highly inter-correlated items (Pearson's $r \geq .70$) were examined for redundancy of content and possible elimination.

To examine factor structure, we subjected the data to a principal components analysis with maximum likelihood extraction. In addition to examining the results of the Scree test and the Kaiser-Guttman Criterion (i.e., eigenvalue greater than 1), we also applied a parallel analysis [16] to decide on the number of factors to retain. As predicted, the MVI data produced a 4-factor solution, with those factors matching our four domains of *interest value, utility value, attainment value,* and *personal cost*. However, on the SVI, *utility value* separated into two factors (*present* and *future*) and *personal cost* separated into three factors (*time pressures, threats to self-efficacy,* and *conflicts with religiosity*) creating a 7-factor solution.

As we expected, the factors, although conceptually independent, were correlated on both inventories. Consequently, the retained factors were subjected to oblique (promax) rotation to allow for these correlations when determining the principal factor to which an item belonged, and during data reduction. We removed items from the SVI and MVI for the following reasons:

- *Initial communality below .20*
- *Pattern/structure coefficient below .45 on the principal factor*
- *Complex pattern/structure coefficients on uncorrelated factors*

Complex pattern/structure coefficients were defined as loadings ≥ .45 on the principal factor *and* ≥ .30 on one or more other factors. This final criterion was selected to minimize factor overlap and enhance the specificity of each subscale. As a result, 46 items were eliminated from the SVI and 51 were eliminated from the MVI.

We then examined the internal consistency of the items in the subscales (factors) using the .70 criterion suggested by Nunnally and Bernstein [17]. On the SVI, Cronbach alphas ranged from .79 to .92, suggesting item interrelatedness at the subscale and total scale level. On the MVI, interrelatedness was stronger, with Cronbach alphas ranging from .87 to .95.

We generated a few new items to ensure equal-item subscales on the final inventories, and then conducted a second round of large-scale item tryouts with a 69-item SVI (N = 872 non-science majors, 58.2 percent female, 90.4 percent Caucasian) and a 32-item MVI (N = 1,096 non-math majors, 59.0 percent female, 92.6 percent Caucasian) using the same procedures described earlier. For each of the seven factors in the SVI, we retained the six items with the highest structure coefficients and also showed the largest reduction in Cronbach's alpha when deleted. Thus, the SVI is a 42-item inventory with seven, equal-item subscales renamed as:

- *Interest*
- *Present Utility*
- *Future Utility*
- *Need for High Achievement*
- *Threats to Self-Efficacy*
- *Time Pressures*
- *Conflict with Religiosity*

Domain descriptions for the seven subscales are provided in Table 1.

Item scores range from 1 to 5, with subscale scores ranging from 6 to 30. The maximum SVI total score is 210, with higher scores reflecting greater perceived value of science. Table 2 provides the pattern/structure coefficients from the final principal components analysis (primary factor only), as well as SVI subscale means, standard deviations, and Cronbach alphas (item numbers followed by a lowercase "r" are reverse scored). The seven factors are shown in Figure 1, along with their interrelationships (Pearson correlations).

On the MVI, a similar procedure resulted in a 28-item inventory with four seven-item subscales, renamed as:

- *Interest*
- *General Utility*
- *Need for High Achievement*
- *Costs*

Domain descriptions for the four subscales are found in Table 3.

Subscale scores range from 7 to 35. The maximum MVI total score is 140, with higher scores reflecting greater perceived value of math. Table 4 provides the pattern/structure coefficients from the final principal components analysis (primary factor only) as well as subscale means, standard deviations, and Cronbach alphas. Figure 2 shows the four factors and their interrelationships (Pearson correlations).

Test-retest Reliability Estimates

To examine the test-retest reliability of scores on the SVI, we administered it to a sample of undergraduate students enrolled in a general education math course for non-science majors. Students were asked to complete the SVI again two weeks later, and 76 students (69.3 percent females, 91.9 percent Caucasian) participated in both sessions. Test-retest correlations for each SVI subscale and the SVI total score were strong, with Pearson's correlations ranging from $r = .74$ to $r = .92$.

For the MVI, 55 introductory psychology students (70.9 percent females, 90.9 percent Caucasian) participated in both the test and the retest sessions. Test-retest correlations for each MVI subscale and the MVI total score were very strong, with Pearson's correlations ranging from $r = .88$ to $r = .96$.

Gender-related Differences on the SVI and MVI

Using the undergraduate sample from the second SVI item tryout ($N = 872$), scores for men and women (non-science majors) on each of the seven subscales and on the inventory as a whole were compared using independent-sample t-tests. Results suggested that females placed *greater* value on high achievement in science, $t(870) = 2.57$, $p = .01$ [effect size = .18], but also reported *less* science self-efficacy, $t(865) = -4.48$, $p < .001$ [effect size = -.31], less interest in science $t(863) = -6.12$, $p < .001$ [effect size = -.42], and *lower* present utility of science, $t(866) = -2.84$, $p < .01$ [effect size = -.20] than males. SVI total-scale score differences were also statistically significant, with females reporting a *lower* valuing of science in general, $t(825) = -2.04$, $p < .05$ [effect size = -.14]. In contrast, there were no statistically significant differences between males and females on perceived future utility of science, time pressures, or conflicts with religiosity.

In contrast, identical analysis of the undergraduate sample from the second MVI item tryout ($N = 1,096$) showed *no* statistically significant gender differences on any MVI subscale or on the MVI total score.

Relations between Previous College Coursework and SVI/MVI Scores

We also used the second SVI/MVI item tryouts to explore the relationship between previous college coursework in science and math and SVI/MVI scores. Using one-way ANOVA with Tukey HSD post-hoc testing, we found that sudents who had completed 3 or more science courses valued science more than students who had completed 0, 1, or 2 courses. Specifically, those completing 3 or more science courses found science to be *more* interesting, *more* useful (both in the present and in their futures), and *less* threatening to their sense of self-efficacy than those who had completed fewer than 3 courses. Need for high achievement, time pressures, and conflicts with religiosity did not differ across these groups.

Additional analyses revealed that SVI scores also varied according to previous coursework in math. In particular, students completing 3 or more math courses found science to be *more i*nteresting and *more* useful to their present and future lives than students who had completed fewer than 3 courses. Need for high achievement and perceived costs did not vary across these groups.

We conducted the same analyses to examine relations between previous coursework and MVI scores. Consistent with the SVI findings, students who had completed 3 or more courses in math found math to be *more* interesting and useful than students who had completed fewer than 3 courses. These students also placed *more* value on high achievement in math and experienced *fewer* perceived costs. Thus, scores on each MVI subscale also varied as a function of prior coursework in the discipline.

Likewise, compared to other students, students who had completed 3 or more courses in *science* found math to be *more* interesting and useful and placed *greater* value on high achievement in the discipline. However, the perceived costs of learning math were not related to previous coursework in science.

Relations between Intended Academic Major and SVI Scores

As part of our construct validation, we hypothesized that students who did *not* intend to major in science would value science *less* than those who did intend to major in science, and would perceive the costs of learning science to be greater. To test this, we administered the SVI to 351 incoming first-year students (58.7 percent female, 91.0 percent Caucasian), of whom approximately 70 percent did not intend to major in science.

As hypothesized, students who did not intend to major in science had *lower* scores on four of the SVI subscales: Interest in Science, $t(326) = -7.77$, $p < .001$ [effect size = -1.0]; Present Utility, $t(325) = -9.42$, $p < .001$ [effect size = -1.18]; Future Utility, $t(323) = -14.91$, $p < .001$ [effect size = -2.02]; and Need for High Achievement, $t(325) = -9.05$, $p < .001$ [effect size = -1.18]. Total SVI scores also varied by intended academic major, $t(312) = -10.21$, $p<.001$ [effect size = -1.33]. However, scores on the three cost subscales did not vary by intended major. Thus, contrary to our expectations, the costs associated with learning science are the same for those majoring in science and those studying in other disciplines. We are currently examining relations between MVI scores and intended major.

DISCUSSION

Over the past three years, we conducted a series of studies in order to develop the SVI and MVI, inventories that measure how much college students value science and math literacy. The development phases of our project have involved expert evaluators as well as over 6,600 undergraduate participants. To date, the 42-item SVI has

seven domains that we believe may be relevant to achieving science literacy: Interest, Present Utility, Future Utility, Need for High Achievement, Time Pressures, Threats to Self-Efficacy, and Conflicts with Religiosity. The 28-item MVI is comprised of four domains: Interest, General Utility, Need for High Achievement, and Costs. Both inventories have identifiable factor structures that are theoretically consistent with published empirical literature, and our analyses provide evidence that both inventories are psychometrically sound. These factor structures are slightly different in math and science, which has important implications for instruction.

As we begin to examine the predictive validity of SVI and MVI scores, we have already gained insights into how men and women value science and math, how values are affected by instruction, and what student values suggest about the structure of effective curricula at all levels. With additional data from a variety of partner institutions, we hope to soon have data on how students at historically black colleges and universities (HBCUs), private liberal arts colleges, and large public universities may differently value these subjects.

We believe this is precisely the kind of data that must be available to evaluate, refine, and support reform in all areas of higher education. For those who share our conviction that the affective domain has been neglected, particularly with those students who choose majors and careers outside of science and math, these tools will provide measures of the community's effectiveness in convincing students that these subjects matter. Indeed, the framework for our research strongly implies that the learning we all desire takes place only when students find such value in their education.

Table 1. Science Value Inventory: Domain Definitions

INTEREST IN SCIENCE This domain refers to the importance a student places on science because of genuine interest in the subject. Items that assess interest in science ask students to reflect on the intrinsic satisfaction they receive from learning about science either inside or outside the classroom.
UTILITY OF UNDERSTANDING SCIENCE Utility is the importance a student places on understanding science because it will help him or her to accomplish a variety of short- or long-term goals. Items that assess utility ask students to reflect on what they have to gain personally as a result of understanding scientific concepts or to reflect on the personal benefit of taking science courses. These items assess why it is useful to understand science, but do not assess what is lost or sacrificed. Utility has two facets: ***Present Utility*** *This facet refers to the usefulness of science in helping students accomplish short-term goals. These benefits may include improving skills or abilities required for success in other college classes, or other things that produce similarly immediate benefits.* ***Future Utility*** *Future Utility refers to the usefulness of science in helping students meet their long-term goals after completing their study of science. These include the value of science in improving career prospects as well as performing well in future college courses.*

NEED FOR HIGH ACHIEVEMENT IN SCIENCE

Need for high achievement is the importance a student places on doing well in science. Items that assess need for high achievement ask students to reflect on how important it is to develop a good understanding of science or to achieve at high levels in their science courses. These items do not assess why it is important to have a good understanding of science.

PERSONAL COST OF UNDERSTANDING SCIENCE

Personal Cost refers to the sacrifices a student believes are required to develop an understanding of science or to do well in science courses. Items that assess personal cost ask students to reflect on what may be lost, given up, or compromised in order to master scientific concepts. Personal Cost has three facets:

Threats to Self-efficacy

This refers to the perceived impact of learning science on a student's judgment of his or her own abilities. Items that assess this facet ask students to reflect on how struggling with science may pose threats to their self-esteem or lead them to question their abilities or intelligence.

Time Pressures

This facet refers to the conflict that may arise between the time and effort required to understand science, and the time needed to accomplish other goals in a student's life. Items assessing time pressure ask students whether learning science takes too much time, particularly in the context of competing daily activities.

Conflict with Religiosity

This refers to the conflict that students may find between scientific ideas and their religious beliefs. Items that assess this facet ask students to consider how articles of their religious beliefs may be compromised or challenged in order to succeed in science classes.

Table 2. Science Value Inventory: Pattern and Structure Coefficients from Principal Components Analysis (Primary Factor Only) with Subscale Means, Standard Deviations, and Cronbach Alphas

			Pattern Coefficient	Structure Coefficient
SVI Full Scale Statistics: M = 136.73, SD = 22.31, α = .91				
I. Interest (M = 18.29, SD = 5.85, α = .90)				
	10	I enjoy reading magazine articles about science.	.89	.87
	15	Learning about scientific discoveries is interesting to me.	.75	.83
	31	Learning scientific material is interesting.	.67	.82
	38	I like reading about science on the internet.	.84	.81
	27	I enjoy watching science programs on TV.	.82	.80
	40	Science fascinates me.	.70	.79
II. Present Utility (M = 19.99, SD = 5.01, α = .89)				
	34	Studying science improves my problem-solving skills.	.85	.89
	1	Solving science problems improves my critical thinking skills.	.82	.85
	35	Science gives me insight into real-world problems.	.72	.81

	21	Learning about science helps develop my ability to think creatively.	.66	.76
	19	My lifestyle choices will be healthier if I have a good knowledge of science.	.54	.71
	24	My life will be better if I understand science.	.37	.67
III. Future Utility (M = 17.61, SD = 5.41, α = .90)				
	37	Understanding science improves my chances of getting a good job.	.85	.89
	32	A good knowledge of science will give me more career options.	.78	.85
	13	My future income will be higher if I have a good understanding of science.	.87	.84
	5	Studying science is necessary to prepare me for my career.	.81	.80
	16	I'll be better off in the future if I have a good understanding of science.	.53	.74
	12	Understanding science allows me to better understand my other classes.	.58	.73
IV. Need for High Achievement (M = 19.57, SD = 5.26, α = .84)				
	7	Only a course grade of "A" in science is acceptable to me.	.84	.85
	30	If I do not receive an "A" on a science exam, I am disappointed.	.80	.80
	11	It is important to me to get one of the top grades in my science classes.	.78	.80
	25	I must do well in my science classes.	.70	.74
	6r	I don't care if I'm a poor science student.	.58	.64
	20	It bothers me when other students perform better than me on a science exam.	.68	.64
V. Threats to Self-Efficacy (M = 19.52, SD = 5.79, α = .90)				
	41r	Taking science exams makes me less confident in my academic ability.	.83	.87
	36r	Learning science makes me question my intelligence.	.83	.85
	26r	Learning science makes me question my general academic ability.	.82	.83
	4r	My self-esteem suffers when I take science exams.	.79	.81
	23r	It makes me feel stupid when others are successful at learning science and I'm not.	.79	.79
	14r	I have to study much harder for science than for other classes.	.57	.69
VI. Time Pressures (M = 17.09, SD = 4.99, α = .84)				
	2r	It's hard to focus on my science classes when I have so much else to do.	.82	.82
	42r	I find it difficult to study science when I have so many other things to do.	.76	.77
	28r	I don't have enough time to study for science.	.69	.74
	22r	I would do better in science if other obligations took less of my time.	.72	.72
	39r	Work-related activities interfere with my studying science.	.75	.71
	8r	Studying science takes an unreasonable amount of my time.	.53	.68
VII. Conflict with Religiosity (M = 24.58, SD = 5.15, α = .84)				
	17r	Learning science conflicts with the sacred text(s) of my religious faith.	.82	.82
	18r	Learning science undermines my religious beliefs.	.80	.81
	9r	Studying science is difficult for me, because of the religious intolerance of science teachers.	.79	.80
	33r	Understanding science requires me to accept theories that conflict with my religious faith.	.74	.74
	29r	I must hide my religious faith when taking courses in science.	.70	.70
	3	I find no conflict between my religious faith and my understanding of science.	.65	.64

Inventory item numbers are listed to the left of each item; reverse scored items are indicated with an "r".

Table 3. Math Value Inventory: Domain Definitions

Interest in Math refers to the importance a student places on math because of a genuine interest in the subject. Items that assess interest in math ask students to reflect on the intrinsic satisfaction they receive from learning about math either inside or outside the classroom.
General Utility of Math is the importance a student places on understanding math because it will help him or her to accomplish a variety of short- or long-term goals. Items that assess utility ask students to reflect on what they have to gain personally as a result of understanding mathematical concepts or to reflect on the personal benefit of taking math courses. These items assess why it is useful to understand math, but do not assess what is lost or sacrificed.
Need for High Achievement is the importance a student places on doing well in math. Items that assess need for high achievement ask students to reflect on how important it is to develop a good understanding of math or to achieve at high levels in their math courses. These items do not assess why it is important to have a good understanding of math.
Costs refers to the sacrifices a student believes are required to develop an understanding of math or to do well in math courses. Items that assess personal cost ask students to reflect on what may be lost, given up, or compromised in order to master mathematical concepts.

Table 4. Math Value Inventory: Pattern and Structure Coefficients from Principal Components Analysis (Primary Factor Only) with Subscale Means, Standard Deviations, and Cronbach Alphas

			Pattern Coefficient	Structure Coefficient
Total MVI Scale Statistics: M = 90.11, SD = 21.80, α = .95				
I. Interest (M = 18.22, SD = 7.41, α = .95)				
	24	Solving math problems is interesting for me.	.95	.93
	12	I find many topics in mathematics to be interesting.	.95	.92
	20	I am interested in doing math problems.	.87	.90
	27	Mathematics fascinates me.	.94	.89
	16	It's fun to do math.	.86	.89
	2	Learning new topics in mathematics is interesting.	.80	.84
	9	I find math intellectually stimulating.	.67	.80
II. General Utility (M = 26.16, SD = 6.27, α = .92)				
	3r	There are almost no benefits from knowing mathematics.	.90	.86
	17r	I see no point in being able to do math.	.84	.86
	13r	Having a solid background in mathematics is worthless.	.82	.83
	6r	I have little to gain by learning how to do math.	.79	.82
	21	Understanding math has many benefits for me.	.71	.79
	13r	After I graduate, an understanding of math will be useless to me.	.78	.78
	23r	I don't need math in my everyday life.	.74	.74

III. Need for High Achievement (M = 24.49, SD = 6.83, α = .92)				
	19	Earning high grades in math is important to me.	.88	.89
	11	It is important to me to get top grades in my math classes.	.88	.85
	25	Only a course grade of "A" is acceptable to me.	.84	.82
	28	I must do well in my math classes.	.82	.82
	14	I would be upset to be just an average student in math.	.77	.81
	4	If I do not receive an "A" on a math exam, I am disappointed.	.85	.81
	14	Doing well in math courses is important to me.	.72	.80
IV. Costs (M = 21.06, SD = 7.38, α = .91)				
	5r	Taking math classes scares me.	.83	.87
	22r	Trying to do math causes me a lot of anxiety.	.85	.86
	26r	Math exams scare me.	.89	.85
	1r	I have to study much harder for math than for other courses.	.82	.82
	7r	I worry about getting low grades in my math courses.	.83	.79
	18r	Solving math problems is too difficult for me.	.67	.76
	15r	Mathematical symbols confuse me.	.67	.71

Inventory item numbers are listed to the left of each item; reverse scored items are indicated with an r.

Figure 1. Seven domains of value of the Science Value Inventory (SVI). Subscale correlations are shown by curved arrows. Subscale correlations below 0.3 have been omitted.

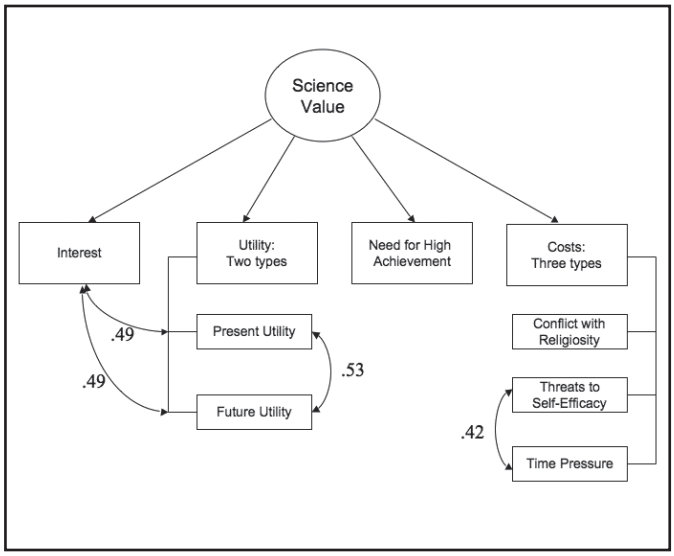
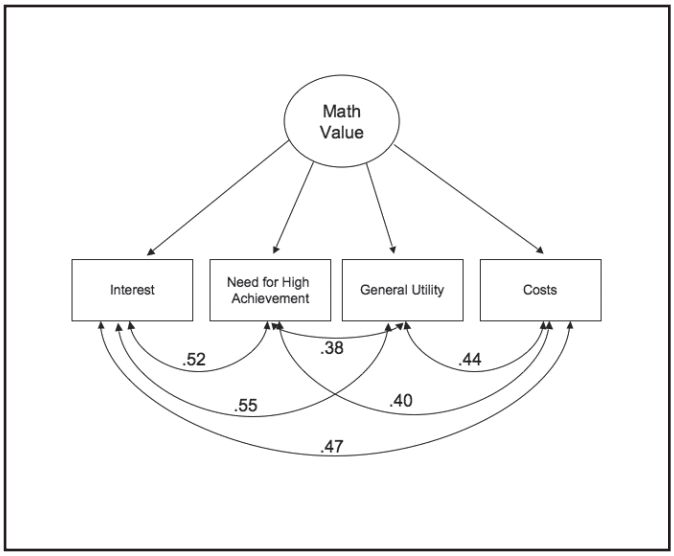


Figure 2. Four domains of value on the Math Value Inventory (MVI). Subscale correlations are shown by curved arrows. As the figure illustrates, all domains on the MVI are moderately interrelated.



ACKNOWLEDGMENTS

We thank our experts for their assistance in establishing the validity of the domains and corresponding items. For the SVI, we are grateful to Dr. Rodger Bybee, Biological Sciences Curriculum Study; Dr. Patty Elmore, Southern Illinois University–Carbondale; Dr. David Hanson, Stony Brook University; Dr. Jay Labov, National Research Council; and Dr. Priscilla Laws, Dickinson College. For the MVI, we thank Dr. John Ed Allen, North Texas State University; Dr. John Chapman, Southwestern University; Col. Gary Krahn, United States Military Academy; Dr. Carol Schumacher, Kenyon College; and Dr. Lynn Steen, St. Olaf College.

REFERENCES

[1] U.S. Department of Education, *A test of leadership: Charting the future of U.S. higher education*, (U.S. Department of Education, Washington, D.C.), 2006. Available at http://www.ed.gov/about/bdscomm/list/hiedfuture/reports/final-report.pdf

[2] Deeds, D. G., Wood, M. D., Callen, B. W., and Allen, C.S., "Contemporary student attitudes about mathematics, science and technology," *Journal of College Science Teaching*, vol. 29, pp. 86–92, 1999.

[3] Rokeach, M., "From individual to institutional values with special reference to the values of science," in Rokeach, M., (ed.), *Understanding human values*, (The Free Press, New York), pp. 47–70, 1979.

[4] Feather, N. T., "Expectancy-value approaches: Present status and future directions", in Feather, N.T., (ed.), "*Expectations and actions: Expectancy-value models in psychology*," (Lawrence Erlbaum Associates, Hillsdale, N. J.), pp. 395-420, 1982.

[5] Rotter, J. B., *Social learning and clinical psychology,* (Prentice-Hall, New York), 1954.

[6] Atkinson, J. W., "Motivational determinants of risk-taking behavior," *Psychological Review*, vol. 64, no. 6, pp. 359–372, 1957.

[7] Eccles (Parsons), J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., and Midgley, C., "Expectancies, values, and academic behaviors", in Spence, J. T., (ed.), "*Achievement and achievement motivation*", (W.H. Freeman, San Francisco), pp. 75–146, 1983.

[8] Eccles, J., Adler, T., and Meece, J. L. "Sex difference in achievement: A test of alternate theories," *Journal of Personality and Social Psychology,* vol. 46, no. 1, pp. 26–43, 1984.

[9] Eccles, J. S., "Why doesn't Jane run? Sex differences in educational and occupational patterns," in Horowitz, F.D., and O'Brien, M., (eds.), *The gifted and talented: Developmental perspectives*, (American Psychological Association, Washington, D. C.," pp. 251-295, 1985.

[10] Eccles, J., "Gender roles and women's achievement-related decisions," *Psychology of Women Quarterly*, vol. 11, pp. 135-172, 1987.

[11] Eccles, J. T., Barber, B. L., and Jozefowicz, D.M., "Linking gender to educational, occupational, and recreational choices: Applying the Eccles et al. model of achievement-related choices," in Swann, W. B., Langlois, J. H., and Gilbert, L.C., (eds.), *Sexism and stereotypes in modern society: The gender science of Janet Taylor Spence*, (American Psychological Association, Washington, D. C.), pp. 153–192, 1999.

[12] Jacobs, J. E., and Eccles, J.S., "Parents, task values, and real-life achievement-related choices," in Sansone, C., and Harackiewicz, J.M., (eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance*, (Academic Press, San Diego), pp. 405–439, 2000.

[13] Haynes, S. N., Richard, D. C. S., and Kubany, E.S., "Content validity in psychological assessment: A functional approach to concepts and methods," *Psychological Assessment*, vol. 7, pp. 238-247, 1995.

[14] Smith, G. T., and McCarthy, D.M., "Methodological considerations in the refinement of clinical assessment instruments," *Psychological Assessment*, vol. 7, no. 3, pp. 300-308, 1995.

[15] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *The standards for educational and psychological testing*, (AERA, Washington, D. C.), 1999.

[16] Velicer, W. F., Eaton, C. A., and Fava, J. L., "Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components," in Goffin, R.D., and Helmes, E., (eds.), *Problems and solutions in human assessment Honoring Douglas Jackson at seventy,* (Kluwer, Boston), pp. 41–71, 2000.

[17] Nunally, J. C., Bernstein, I.H., *Psychometric theory (*3rd ed.), (McGraw-Hill, New York) 1994.

Collaborative Project: Using Students' Design Tasks to Develop Scientific Abilities

Xueli Zou* - *California State University, Chico*

OVERVIEW

This article describes experimental design tasks for students that have been developed, tested, and implemented in introductory physics labs. These design tasks are aimed at developing physics students' scientific abilities [1, 2]. Detailed examples of the design tasks will be illustrated below. This work is a collaborative effort with the Astronomy and Physics Education Research Group at Rutgers led by Alan Van Heuvelen and Eugenia Etkina. For more information on scientific abilities and related rubrics, visit http://paer.rutgers.edu/scientificabilities; for more examples of different experimental design tasks, visit http://www.csuchico.edu/~xzou.

DEVELOPMENT AND IMPLEMENTATION OF STUDENTS' DESIGN TASKS

Four different types of hands-on design tasks—observational experiments, testing experiments, application experiments, and investigation experiments—have been developed for calculus-based introductory physics labs. Students working in small groups are engaged in designing their own experiments either to explore some physical phenomena, verify a physics principle, build a real-life device, or conduct an investigation. These design activities, including both qualitative and quantitative tasks, help students not only better understand concepts, but also think creatively and develop important workplace skills in design, complex problem solving, scientific investigation, teamwork, communication, and learning how to learn. Identified scientific abilities in conducting the design tasks, related sub-abilities, and scoring rubrics are reported in detail in references 1 and 2.

An *observational experiment* asks students to "play around" using some given equipment so as to explore and observe certain physical phenomena. Figure 1 describes an example of a qualitative observational experiment used in a calculus-based electricity and magnetism course. These observational experiments are given to students before related concepts are discussed in class. In a traditional instruction setting, these observational experiments are usually conducted by the instructor as lecture demonstrations while associated concepts are introduced. For many students, the experiments are more like magic shows; they have no ideas about how they work. As a result, of course, it is difficult for students to construct the experiments as part of their own knowledge.

While using these experiments as design tasks, students themselves need to come up with some ideas and conduct the experiments. Students not only gain a better understanding of the phenomena, but no longer view them as magic. The experiments also arouse their curiosity to further study the phenomena, which is evident from the many questions students ask when the phenomena are discussed in subsequent classes.

Many observational experiments have been developed covering the important concepts in calculus-based introductory mechanics and electricity and magnetism courses. For some electrostatic phenomena, which are difficult for students

Figure 1. Example of a qualitative observational experiment

Can you generate currents without a battery?

Using the given materials, including a magnet, a Galvanometer, some PVC wire coils, and wires, conduct some experiments so as to generate a current through the Galvanometer. In your lab notebook, please 1) sketch your experimental setup and describe what you have performed, 2) record in which ways the magnitude of the generated current could be changed, and 3) record in which ways the direction of the generated current could be changed.

themselves to perform successfully due to physical constrictions (e.g., room humidity) in a student lab room, digital videos have been made for students to observe and construct the concepts. Those videos are available at http://www.csuchico.edu/~xzou.

A *testing experiment* asks students to devise an experiment to test a conceptual or mathematical model. Figure 2 describes a qualitative testing design task given in a calculus-based mechanics laboratory. In contrast to a traditional verification experiment, students are not provided with step-by-step instructions and ready-to-go setups. Instead, they are provided with certain materials, equipment, a procedure outline (see below) [3], and detailed rubrics (see examples in [1]) as real-time feedback.

Basic Procedure of Testing Experiment
Model: *State the model to be tested*
Plan: *Describe some phenomenon that can be analyzed using the model*
Prediction: *Predict the outcome of the situation*
Experiment: *Set up, conduct the experiment, and record, analyze data*
Conclusion: *Decide if experimental evidence supports the model*

Figure 2. Example of a qualitative testing experiment

Using two carts and the other given equipment on your table, design and perform some situations for which you predict that the force exerted by cart A on cart B is greater than the force exerted by cart B on cart A (or vice versa).

Students work in small groups of 3 or 4 to conduct their experiments, and are encouraged to follow explicitly the basic process that guides practical scientists in conducting their own experiments. It is observed that many introductory physics students have difficulties conducting such testing experiments as a scientist would since they have no similar previous experiences. It is important at the very beginning of lab to discuss with students what an acceptable scientific design is and what a poor, naïve design is using real examples. In particular, hypothetical-deductive reasoning [4, 5] and the power of predication in science are addressed and emphasized explicitly with students [6, 7].

Any traditional verification experiments can be readily revised as testing experiments. In addition, students' common-sense ideas or identified alternative conceptions can be used as "hypotheses" to be tested (see the example in Figure 2). Ample evidence in physics education research has shown that it is very difficult to address students' "misconceptions" effectively. An experimental design task to test students' own "misconceptions" serves an effective way to help students reconstruct their knowledge.

An *application experiment* requires students to build a real-life device or devise an experimental approach to measure some physical quantities. Figure 3 describes a quantitative design task in measuring the maximum coefficient of static friction between a wooden block and a wooden board. To conduct such application designs, students are provided with some materials and equipment and encouraged to follow a basic process of effective problem solving (see below) [3]. A significant component of an application experiment is *evaluation*—students are required to devise an additional approach or experiment to evaluate their own design or result.

Basic Procedure of Application Experiment
Plan: *Define the problem and plan an experiment*
Assumption: *State any assumptions or estimations*
Prediction: *Theoretically predict outcomes of the experiment*
Experiment: *Set up, conduct the experiment, and record, analyze data*
Evaluation: *Perform additional experiment to evaluate the result*

Figure 3: Example of quantitative application experiment

Devise and perform an experiment to determine the maximum coefficient of static friction between a wooden block and a wooden board.
Design and perform an additional experiment to evaluate your result.

An *investigation experiment* is a more challenging case study combining the other three design tasks described above. For example, toward the end of a calculus-based introductory electricity and magnetism lab, students are asked to investigate the magnetic field inside a current-carrying slinky before the concepts are addressed in class. In this lab, students conduct some selected observational experiments, come up with some conceptual ideas to account for the observations, identify possibly related physical quantities, design and conduct their own experiments to explore a quantitative model for the magnetic field inside the slinky, and finally design and conduct an additional experiment to verify or apply their own mathematical model.

In summary, as part of the outcomes of this NSF Assessment of Student Achievement (ASA) project, two lab manuals have been developed for calculus-based introductory mechanics and electricity and magnetism courses that contain design components in most labs. Over thirty individual design experiments can be supplemented in any lab. In addition, over twenty video problems can be used as observational experiments and homework problems. These curricular materials will be made available on the web as they are completed at http://www.csuchico.edu/~xzou.

ASSESSMENT OF STUDENTS' ACHIEVEMENT

Student experimental design tasks have been implemented and tested in calculus-based introductory physics courses with a class enrollment of about 48 students each semester at California State University, Chico since 2003. Each week the students have three one-hour lectures and are divided into two lab sections to conduct one three-hour lab. Typically, about 60 percent of the students are majors in engineering, 30 percent in computer science, and 10 percent in natural sciences.

To assess the impact of these innovative design tasks on students' development of conceptual understanding and scientific abilities, three different types of assessment have been administered. To assess student conceptual understanding, for example, a research-based multiple-choice test, Conceptual Survey of Electricity and Magnetism (CSEM) [8], has been given to calculus-based physics students every seminar since fall 2003. Students' normalized gains [9] are consistently higher than 0.5 [10], which is typical for interactive engagement courses and almost double the national norm from similar traditional classes [8].

How do students perceive the lab with design components? Do their perceptions mirror the goals of the design tasks? To answer these questions, a Q-sort instrument, Laboratory Program Variables Inventory [11], was given to students since fall 2003. For example, the students from spring 2004 listed the following as the "MOST DESCRIPTIVE" three features for the design lab: *1) "Students are asked to design their own experiments." 2) "Lab reports require the interpretation of data." 3) "Laboratory experiments develop skills in the techniques or procedures of physics."* This result [12] shows that the goals of the design tasks had been successfully conveyed to the students.

To assess students' scientific abilities, a performance-based task was given to students in both the design and non-design labs. A short, theoretical, but *fake* "scientific paper" [13] was given to the students a week before the lab. The paper theoretically argues that for falling cotton balls the distance as a function of time should be s = 1/2kt2, where k is a constant but less than g = 9.8 m/s2. The students were asked to design, conduct, and report an experiment to test the theory in a three-hour lab.

As shown in Table 1, most students from both the design and non-design labs explicitly followed the procedures emphasized in each lab to report their results. Then students' reports were analyzed anonymously

and independently by two researchers using the identified scientific abilities (see Table 2) [1, 2], and response taxonomies (see Table 3) developed particularly for this lab based on related rubrics [1, 2]. Discrepancies between scores given to a particular ability in a particular report were carefully discussed between the two researchers and agreements were reached at the end. Percentages of students' scores on the six identified scientific abilities are shown in Table 4 for the design lab and in Table 5 for the non-design lab. One major difference shown in Tables 4 and 5 is that about 70 percent of the non-design lab students did not make any predictions in their experiments, while about 90 percent of the design lab students made reasonable predictions. To compare the students' overall performance from the two labs, the data were run using a Mann-Whitney U Test (two-tailed). The result is that the total score of the design lab is statistically significantly higher than the total score of the non-design lab ($p < 0.002$). This preliminary result indicates that the scientific abilities demonstrated by the design lab students are more explicit than those by the non-design lab students.

Table 1. Format of students' lab reports

Design Lab (Students = 42, Lab groups = 17)	Non-Design Lab (Students = 25, Lab groups = 10)
81 percent *explicitly* wrote their reports following the format emphasized in the lab for a testing experiment: **Model | Plan | Prediction | Experiment | Conclusion**	96 percent *explicitly* wrote their reports following the format emphasized in the lab: **Introduction | Procedure (or Experiment) | Conclusion**

Table 2. Scientific abilities used to analyze students' reports from the deign and non-design labs

1. Able to identify the model to be tested
2. Able to make a reasonable prediction based on the model to be tested
3. Able to design a reliable experiment to test the prediction
4. Able to analyze data appropriately
5. Able to decide whether or not to confirm the prediction based on the experiment results
6. Able to make a reasonable judgment about the model

Table 3. Example of response taxonomy for analyzing students' lab reports

Score/Ability	0	1	2	3
2. Is able to make a reasonable prediction based on the model to be tested	No *explicit* prediction	"My prediction is that the distance formula $s=1/2kt^2$ is not valid because my experience is that $s=1/2gt^2$."	"If our model is correct the cotton balls will take longer to fall than the heavier object. Therefore $k<g$."	"If the model is true, then the slope of s vs. 1/2t2 should be constant $k<g$."

Table 4. Percentages of students' scores on six identified scientific abilities from the design lab

Score/Ability	0	1	2	3
1. Identifying model	2%	17%	64%	17%
2. Making prediction	2%	7%	76%	15%
3. Designing experiment	0%	9%	17%	74%
4. Analyzing data	0%	26%	43%	31%
5. Confirming prediction	12%	7%	17%	64%
6. Making Judgment	9%	64%	17%	10%

Table 5. Percentages of students' scores on six identified scientific abilities from the non-design lab

Score/Ability	0	1	2	3
1. Identifying model	12%	8%	24%	56%
2. Making prediction	72%	4%	24%	0%
3. Designing experiment	0%	28%	24%	48%
4. Analyzing data	0%	68%	0%	32%
5. Confirming prediction	64%	16%	4%	16%
6. Making Judgment	8%	92%	0%	0%

DISSEMINATION AND FUTURE RESEARCH

In summary, this paper has reported on four different types of hands-on design tasks that have been developed, tested, and implemented as formative assessment activities to advance physics students' scientific abilities.

To disseminate the project outcomes, seven peer-reviewed papers have been published or accepted to be published in *The Physics Teacher* [e.g.,14, 15, 16], *American Journal of Physics* [e.g., 7], and in *Physics Education Research Conference Proceedings* [e.g., 17, 18]. More than fifteen talks or posters have been presented at international, national, and local physics conferences such as the AAPT meetings and the PER conferences. For more design tasks, identified scientific abilities, and related rubrics, visit http://www.csuchico.edu/~xzou and http://paer.rutgers.edu/scientificabilities.

Based on the scientific abilities and scoring rubrics developed in this project, a simulated physics experiment is being investigated and potentially used as a diagnostic tool to assess students' scientific abilities learned in the introductory physics laboratory.

ACKNOWLEDGMENTS

Special thanks go to Eugenia Etkina, Alan Van Heuvelen, and other members of the Astronomy and Physics Education Research Group at Rutgers for their collaboration and support. The NSF funding for this project (DUE #0242845) is also greatly appreciated.

REFERENCES

[1] Etkina, E., Brookes, D.T., Murthy, S., Karelina, A., Villasenhor, M.R., and Van Heuvelen, A., "Developing and assessing student scientific abilities," in this volume, 2007.

[2] E. Etkina, Van Heuvelen, A., White-Brahmia, S., Brookes, D. T., Gentile, M., Murthy, S., Rosengrant, D., and Warren, A., "Scientific abilities and their assessment," *Phys. Rev. ST Phys. Educ. Res*., vol. 2, no. 2, 020103, 2006.

[3] The procedure outline is used as a guide not only for students to design their experiments but also to help them explicitly acquire some basic procedural knowledge in conducting scientific experiments. It is observed that the procedure outline is important at the beginning of the lab to lead students in conducting their experiments successfully. As their abilities and interests in design tasks grow, the outline is not needed anymore.

[4] Lawson, A. E., "The generality of hypothetico-deductive reasoning: Making scientific thinking explicit," *Am. Bio. Teach*., vol. 62, no. 7, pp. 482 – 495, 2000.

[5] Lawson, A. E., "The nature and development of hypothetico-predictive argumentation with implications for science teaching," *Int. J. Sci. Ed.*, vol. 25, no. 11, pp. 1387 – 1408, 2003.

[6] Zou, X., and Davies, O., "Development and assessment of students' skills in designing and conducting introductory physics experiments: Part I," National Conference of American Association of Physics Teachers, Sacramento, Calif., August 2004.

[7] Etkina, E., Murthy, S., and Zou, X., "Using introductory labs to engage students in experimental design," *Am. J. Phys.*, vol. 74, no. 11, pp. 979 – 986, 2006.

[8] Maloney, D.P., O'Kuma, T.L., Hieggelke, C.J., and Van Heuvelen, A., "Surveying students' conceptual knowledge of electricity and magnetism," *Am. J. Phys*., vol. 69, S12, 2001.

[9] Hake, R.R., "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *Am. J. Phys*., vol. 66, no. 1, pp. 64-74, 1998.

[10] Zou, X., and Murthy, S., " Development and assessment of students' skills in designing and conducting introductory physics experiments: Part II," National Conference of American Association of Physics Teachers, Sacramento, Calif., August 2004.

[11] Abraham, M., "A descriptive instrument for use in investigating science laboratories," *Journal of Research in Science Teaching*, vol. 19, no. 2, pp. 155-165, 1982.

[12] Zou, X., Lin, Y., Demaree, D., and Aubrecht, G., "From students' perspectives: A Q-type assessment instrument," National Conference of Physics Education Research 2004: Transfer of Learning, Sacramento, Calif., August 2004.

[13] Erickson, T., "Using fake scientific papers as a prompt," National Conference of American Association of Physics Teachers, Sacramento, Calif., August 2004.

[14] Zou, X., "The sandbag collision," *The Physics Teacher*, vol. 44, p. 448, 2006.

[15] Zou, X., "Conductor and Insulator: A QuickTime Movie," *The Physics Teacher*, vol. 43, no. 7, pp. 460-462, 2005.

[16] Zou, X., "Making 'internal thermal energy' visible," *The Physics Teacher,* vol. 42, no. 6, pp. 343-345, 2004.

[17] Aubrecht, G. J., Lin, Y., Demaree, D., Brookes, D., and Zou, X., "Student Perceptions of Physics by Inquiry at Ohio State," *Proceedings of the National Conference of Physics Education Research,* pp. 113-116, 2005.

[18] Y. Lin, D. Demaree, X. Zou, and G. J. Aubrecht, "Student perceptions of laboratory in introductory physics courses: A Q-sort approach," *Proceedings of National Conference of Physics Education Research*, pp. 101-104, 2005.

AWARD #0125651

CHEMQUERY: AN ASSESSMENT SYSTEM FOR MAPPING STUDENT PROGRESS IN LEARNING GENERAL CHEMISTRY

Investigator(s): Angelica Stacy (astacy@berkeley.edu)
Co-PI(s): Mark Wilson
Sponsor: University of California-Berkeley

Students in general education science courses typically do not appreciate that science advances through a series of complex paradigm shifts involving considerable debate and discussion. Their perception that scientific understanding of issues proceeds in a cut-and-dry process impedes their grasp of scientific concepts, and their ability to apply science reasoning to issues in societal contexts. Unfortunately, there is no way to quantify these misconceptions and thereby address them with a concerted curricular response. This project is producing an assessment instrument, the Science Perceptions Survey, designed to measure these dispositions and changes in them. The Science Perceptions Survey is being devised in the context of general education chemistry courses and validated through correlation with results of written essays and interviews. An independent evaluator who is not associated with the development of the assessment tool is managing these validation protocols. Once validated, the instruments will be freely available to other institutions considering changes in curricula focused on enhancing the appreciation of science methods and their place in general education courses. The Principal Investigators anticipate that the Science Perceptions Survey will be a valuable tool in delineating deficiencies in students' ability to understand the way in which scientific arguments are constructed and to guide curriculum development in correcting these deficiencies. In addition to the curriculum assessment format, the Science Perceptions Survey will be disseminated in a web-based format for student self-evaluation. This implementation will provide students real-time analysis of their tolerance for ambiguity in scientific findings and their ability to use the scientific method. This ongoing project can be used by instructors as an assignment in the general education courses to help students understand how science advances knowledge.

AWARD #0127650

REAL-TIME MULTI-DIMENSIONAL ASSESSMENT OF STUDENT LEARNING

Investigator(s): David Hanson (david.hanson@sunysb.edu)
Co-PI(s): David Ferguson, Troy Wolfskill, Janice Grackin
Sponsor: Stony Brook University, Stony Brook, NY

A web-based assessment instrument that provides real-time, multi-dimensional, formative assessment of student learning is being developed. This instrument provides measures of learning by topic and by level of mastery. The levels of mastery are defined as the information level, the algorithmic level, the conceptual level, and the problem-solving level. The information level is characterized by memorization and the ability to recall, repeat pieces of information, and identify information that is relevant. The algorithmic level is characterized by the ability to mimic, implement instructions, and use memorized information in familiar contexts. The conceptual level is characterized by the ability to visualize, rephrase, change representations, make connections, and provide explanations. The problem-solving level is characterized by the ability to use material in new contexts; to analyze problems; to identify the information, algorithms, and understanding needed to solve them; to synthesize these

components into a solution; and to evaluate the quality of this solution. Four products are resulting from the project: 1) A web-based assessment system founded on a coded database of questions; 2) A coded database of questions for use in measuring student learning in introductory college chemistry, calibrated to reproduce measurements generated through student interviews; 3) Protocols for developing and calibrating databases for use in other SMET courses; and 4) Quantitative and qualitative analyses of the use of this system by both students and teachers, including the impact of the instrument on the users' learning and teaching strategies and the impact of the instrument on student performance.

Students are using the instrument to self-assess their learning after each assignment; to clarify expectations for their performance, particularly with regard to conceptual understanding and problem solving; and to identify strategies for improving their achievement. Faculty are using it to prepare formative and summative assessments, to target particular practices and topics for improved instruction, to identify groups of students who need immediate assistance, and to assess the effectiveness of new teaching strategies and materials.

AWARD #0127694

SUPPORTING ASSESSMENT IN UNDERGRADUATE MATHEMATICS
Investigator(s): J. Michael Pearson (pearson@maa.org), Thomas Rishel (Former Principal Investigator)
Co-PI(s): Bernard Madison
Sponsor: Mathematical Association of America, Washington, DC

The objective of this project is to support faculty members and departments in efforts to assess student learning in at least one of the following: 1) Coherent blocks of courses of undergraduate mathematical sciences, including entire degree programs; and 2) Individual courses, especially reform courses, using various assessment tools across varieties of institutions. The targeted blocks of courses are: a) the major in mathematics; b) courses for future teachers; c) school mathematics as a preparation for college mathematics, usually called college placement programs; and d) general education courses, including those aimed at quantitative literacy. This latter block includes study of assessments of the mathematical and quantitative literacy achieved in entire degree programs, recognizing that much mathematics is learned outside mathematics courses. Assessment cycles that use assessment for program improvements are of special interest, including those that use research on learning.

Support from this work includes: 1) Nationwide distribution and discussion of a 1999 volume of case studies on assessment practices in undergraduate mathematics; 2) Compilation and nationwide distribution of a second volume of case studies and syntheses of coherent sets of case studies; 3) Construction and maintenance of a website to contain an annotated bibliography, literature synopses, interviews, and assessment designs; and 4) Development and operation of a series of workshops for faculty working on campus-based assessment programs, including both face-to-face and electronic venues.

Assessment cycles for program improvement have not been integrated into core operations of departments. This project greatly facilitates assessment program development through wide dissemination of knowledge gained through experiences with assessment.

AWARD #0127751

ASSESSING COGNITIVE COMPONENTS OF DESIGN FOR COURSE-BASED IMPROVEMENT

Investigator(s): Judith Sims-Knight (jsimsknight@umassd.edu)
Co-PI(s): Richard Upchurch, Nixon Pendergrass, Paul Fortier
Sponsor: University of Massachusetts, Dartmouth, Dartmouth, MA

This project is developing, implementing, and evaluating assessment tools and methods that support continuous improvement in student preparation in electrical and computer engineering design and related areas. Educational research has developed findings which suggest that engineering education would be improved by deliberately and carefully assessing important educational outcomes and building continuous improvement loops around courses and curricula. One of the most important goals of engineering education is preparing students to succeed in developing solutions to large-scale, ill-structured design problems that are typical of senior capstone design project courses and the professional world of engineering practice.

The project is using the knowledge of expert functioning provided by cognitive science and instructional research to identify crucial aspects of the design process. It is developing multiple measures and gathering data to verify that these aspects are significant components of design skill. In broad terms, the components of expertise being addressed are declarative knowledge, procedural knowledge, and metacognitive processing knowledge. These are being illuminated by the development of quantitative ways to measure quality of design. Declarative knowledge (including "facts") tends to build and become refined with experience, and its structure tends to become organized in more abstract schemas with experience, allowing it to be applied confidently to a greater range of problems. Procedural knowledge is understanding how to address problems efficiently. Metacognitive processing knowledge is particularly significant for developing lifelong learning skills. It is developed as students learn to think broadly about how they are solving particular problems and seek better or more efficient approaches for future use. This is knowledge that requires planning, self-monitoring, and reflection. Its growth responds to deliberate efforts by instructors to develop it.

The approach is being class tested in sophomore design courses and senior project design courses in two engineering areas: electronic and computer software design. Participating instructors are evaluating the assessments, writing improvement plans, and implementing these over a period of three years. The project is producing three distinctive products. First, it is producing assessments of various components of design skill, with reliability and validity findings. Second, it is evaluating the effectiveness of these assessment measures for use in course-based continuous improvement of student learning. Finally, the assessments of sophomore knowledge in the first year of the project are being used to predict students' performance in their senior design courses.

AWARD #0127806
DEVELOPING AN OUTCOMES ASSESSMENT INSTRUMENT FOR IDENTIFYING ENGINEERING STUDENT MISCONCEPTIONS IN THERMAL AND TRANSPORT SCIENCES
Investigator(s): Ronald Miller (rlmiller@mines.edu)
Co-PI(s): Barbara Olds, Ruth Streveler
Sponsor: Colorado School of Mines, Golden, CO

This project is creating an outcomes assessment instrument to identify engineering student misconceptions in thermal and transport science courses, such as thermodynamics, fluid mechanics, heat transfer, and mass transfer. The instrument focuses on misconceptions concerning fundamental molecular-level and atomic-level phenomena, including heat, light, diffusion, chemical reactions, and electricity, which differ in significant ways from observable, macroscopic causal behavior.

Important student misconceptions identified by surveying experienced engineering faculty are validated through student interviews for inclusion in a multiple-choice pencil-and-paper instrument patterned after successful misconception instruments such as the Force Concept Inventory.

The instrument is being field-tested to demonstrate its validity and reliability, and its usefulness for both course-level and program-level assessment of student misconceptions in thermal and transport science topics. The instrument is designed to allow for pre-testing (at the beginning of a course or curriculum) and post-testing (at the end of a course or curriculum) to measure changes in student mental model development.

AWARD #0350395
CONCEPTUAL UNDERSTANDING OF THREE DIMENSIONS OF EARTH PROCESSES IN GENERAL EDUCATION AND INTRODUCTORY COURSES: TEST DEVELOPMENT AND VALIDATION
Investigator(s): Julie Libarkin (libarkin@msu.edu)
Sponsor: Ohio University, Athens, OH

This project falls into the New Development track of the program guidelines. An assessment tool is being developed that can be used as both a diagnostic tool and a measurement of instructional effects, with primary focus on student conceptual understanding in the geosciences. Conceptual understanding and change are being targeted for two reasons. First, conceptual understanding implies both a familiarity with content and the ability to apply it to complex questions. Second, a number of studies have suggested that prior knowledge can be as important to understanding as pedagogy. As such, students' personal understanding of Earth systems may impact the way in which they understand and retain the formal geoscience they are exposed to. The primary goal of this study is the dissemination of a reliable and valid assessment tool to geoscience faculty around the nation, for use in general education and introductory geoscience courses, as a means of both diagnosing student preconceptions and assessing one aspect of course effectiveness. This test can thus be used as a cross-course, cross-university assessment instrument and as a means for comparing a variety of instructional styles and other variables, including disparate student outcomes related to characteristics such as age and gender. This goal will be achieved by:

1) Identification of alternative conceptions of geological processes through a comprehensive literature search and interviews with students. Geoscience covers a range of interdisciplinary studies, and we have narrowed

our focus to three dimensions: Earth's crust (including topographic expression and geographic expression), Earth's interior, and Earth through time. Roughly 70 student interviews will be conducted at four different institutions with dissimilar student populations.

2) Development of a multiple-choice conceptual assessment tool using commonly held misconceptions as distractors. These misconceptions will be catalogued from the student interviews described above.

3) Qualitative and statistical validation of the assessment tool, to ensure robustness as a comparative instrument. We will ensure both validity (the ability of a test to measure a specific characteristic) and reliability (the internal consistency of the test items and test reproducibility) of this test. Reliability and validity will be ensured through initial piloting with novice students, educators and expert geoscientists, and Item Response Theory statistics.

The availability of an assessment tool involving fundamental conceptions in geology will be invaluable for university faculty interested in assessing introductory and non-major courses in the geosciences. Additionally, the use of misconceptions as distractors allows interested faculty to use the pre-test as a diagnostic tool, to determine the kinds of alternative ideas held by their students. Teachers can then modify course structures to specifically target these preconceptions. Finally, the availability of a standard test will allow geoscience faculty to compare courses at different universities, and in so doing allow a basis by which different instructors, teaching methodologies, curricula, and technologies can be compared. This type of evaluation is critical if we are ever truly going to answer the question, "What works in the geoscience classroom?"

AWARD #0127828
DEVELOPMENT OF FACULTY COLLABORATIVES TO ASSESS ACHIEVEMENT OF STUDENT LEARNING OUTCOMES IN CRITICAL THINKING IN BIOLOGY CORE COURSES
Investigator(s): Judith Kandel (jkandel@fullerton.edu)
Co-PI(s): Joyce Ono, Merri Casem, William Hoese
Sponsor: California State University-Fullerton Foundation, Fullerton, CA

The Department of Biological Science at California State University Fullerton (CSUF) is implementing a major curricular revision of the biology major's program in fall 2002. Planning for this revision involved the majority of the 24 faculty in the department working as collaborative teams and was based on explicit identification of student learning outcomes for the entire curriculum and for each of the four new core courses that replace the eight core courses previously offered. Faculty worked in groups called Teaching Collaboratives, according to their expertise, to develop the four new core courses. The new core courses incorporate active learning and inquiry-based activities in both the lecture and laboratory sections. The Faculty Collaboratives are now developing and adapting assessment instruments and techniques for critical thinking and problem-solving skills, linked to the major concepts and themes within the courses. Faculty teaching the same core courses (members of the Teaching Collaboratives) are sharing their assessment items and developing new ones with the guidance of assessment software and in collaboration with an assessment consultant and process facilitator. In addition, sample student work resulting from these assessment items is being examined by the Teaching Collaboratives to identify strengths and weaknesses of the assessment items and of the learning environment. The end result is development of: an informed faculty; an electronic system for gathering the data to monitor the impact of curricular changes on

student achievement, attitudes, and retention; and an accessible but secure database that allows faculty to share assessment items and to monitor student profiles throughout the biology program.

AWARD #0127725

ASSESSING STUDENT LEARNING AND EVALUATING FACULTY CAPACITY DEVELOPMENT IN THE NSF-FUNDED REGIONAL WORKSHOPS PROJECT

Investigator(s): Raffaella Borasi (raffaella.borasi@rochester.edu), Richard Iuli (Former Principal Investigator)
Co-PI(s): Susan Millar, Mark Connolly, Susan Lottridge
Sponsor: University of Rochester

An NSF-funded national dissemination project, "Disseminating Successful Strategies for Fieldwork in Undergraduate Science Curricula" (DUE 0088217), also known as the "Regional Workshops Project" (RWP), began a 5-year effort in 2001 to conduct 15 regional workshops for 300 undergraduate faculty at the rate of 3 per year. The RWP is engaged in establishing professional learning communities of faculty who 1) create and deliver undergraduate SMET courses that demonstrate that environmental problem solving is an integrative, challenging, effective way to engage undergraduate majors and non-majors, 2) use concepts and field/laboratory techniques suitable for teaching undergraduates how science is done in a real-world, problem-solving context, and 3) use research-based knowledge of how to assess student learning and support faculty capacity for development as educators on an environmental problem-solving-based curriculum for undergraduate SMET courses.

This ASA project is simultaneously examining the effects of the NSF-funded Regional Workshops Project (RWP) on improved faculty capacity to foster increased student learning. The approach is to study both the faculty capacity development initiated by the RWP workshops and sustained by interactions and resources (both planned and unplanned) available to faculty following their workshop experience, and the effects of these faculty capacity building processes on student learning. More specifically, this project is studying: 1) the development of RWP participants' capacity to use research on SMET learning and assessment practices in ways that result in changes in the participants' approaches and attitudes toward teaching; 2) the extent to which faculty are supported while undertaking significant change in their curricula and pedagogy (where pedagogy includes assessment activities) in ways that result in improvements in the participants' abilities to develop, sustain, and institutionalize their new environmental problem-solving-based courses; and 3) the extent to which students in SMET courses : a) learn SMET concepts in a meaningful way; b) construct a view of SMET disciplines that is consistent with views held by experts in those disciplines; c) construct integrative conceptual frameworks to facilitate their understanding of SMET disciplines; and d) develop positive attitudes and perceptions about SMET disciplines.

The project is providing the RWP PIs and regional workshop leaders with formative evaluation of the effectiveness of the regional workshops, and tested instruments and an easy-to-use analysis and report process that workshop leaders can use to undertake formative evaluation of other faculty development workshops with similar goals. These formative evaluations will be designed to help workshop leaders improve the format, content, delivery, and "climate" of the workshop. A tangible outcome will be a tested longitudinal assessment/evaluation process that SMET faculty can adapt to gather credible, dependable, transferable, and confirmable feedback that 1) guides course changes in support of improved student learning and 2) fosters their own professional growth and development.

AWARD #0206977

THE STATISTICAL CONCEPTS INVENTORY (SCI): A COGNITIVE ACHIEVEMENT TOOL IN ENGINEERING STATISTICS
Investigator(s): Teri Reed-Rhoads (trhoads@purdue.edu)
Co-PI(s): Teri Murphy
Sponsor: University of Oklahoma Norman Campus, Norman, OK

This project focuses on the development of a new assessment instrument, applicable to multiple undergraduate engineering programs, to measure students' understanding of statistics and its applications. The statistical understanding measure developed under this research, called the Statistical Concepts Inventory (SCI), provides score profiles that specifically describe students' abilities to design and conduct experiments as well as to analyze and interpret data.

This project is timely because an increasing number of post-secondary engineering programs are endorsing "outcome requirements" that depend on statistical thinking and problem-solving skills. Within engineering, these requirements are precipitating major changes in engineering education, specifically in general and engineering statistics education. Engineering curricular objectives in many respects are being driven by the ABET EC 2000 criteria. Of special relevance to the proposed project are:

1) Criterion 3, Program Outcomes and Assessment, which states that "Engineering programs must demonstrate that their graduates have: a) an ability to apply knowledge of mathematics, science, and engineering, and b) an ability to design and conduct experiments, as well as to analyze and interpret data."

2) Criterion 8, Program Criteria, in which 16 of the 24 listed programs directly indicate the need to demonstrate that students have acquired facility with statistics.

Industrial Engineering has historically taught statistics as a service course to other engineering programs and continues to use statistics as a foundation for much of its own curriculum. Recently approved are the Criteria for Accrediting Computing Programs (Computing Accreditation Commission, December 30, 2000), which covers Computer Science programs and states in its curriculum standards that "Course work in mathematics must include probability and statistics."

This project is also exploring the links between cognitive and attitudinal aspects of introductory statistics courses. In its second phase, this project is gathering SCI profile scores from students in combination with an existing affective instrument, the Survey of Attitudes Toward Statistics (SATS).

AWARD #0206571
THE WEB-BASED ARTIST PROJECT
Investigator(s): Joan Garfield (jbg@tc.umn.edu)
Co-PI(s): Robert delMas, Beth Chance
Sponsor: University of Minnesota-Twin Cities, Minneapolis, MN

This project develops Web-based resources for first courses in statistics, called the Assessment Resource Tools for Improving Statistical Thinking (ARTIST). The Web ARTIST project produces the following products:

1) A collection of high-quality assessment items and tasks, coded according to content (e.g., normal distribution, measures of center, bivariate data) and type of cognitive outcome (e.g., statistical literacy, reasoning, or thinking).

2) A Website that contains the assessment items and tasks, provides online testing, offers guidelines for using the assessment items/tasks in various ways, and allows for the collection and compilation of data for research and evaluation purposes.

3) Faculty development workshops and mini-courses to encourage and assist statistics instructors in how to use the assessment resources to improve student learning, improve their courses, and evaluate course outcomes.

4) A comprehensive test that measures desired outcomes of a first course in statistics.

The ARTIST website includes a variety of item formats and types of performance assessments. Instructors have a centralized resource to help them better evaluate student attainment of particular outcomes, rather than global measures of achievement. Specifically, outcomes to evaluate include statistical literacy (e.g., understanding words and symbols, being able to read and interpret graphs and terms), statistical reasoning (e.g., reasoning with statistical information, using statistics to make predictions or judgment), and statistical thinking (e.g., the type of thinking that statisticians use when solving problems that involve data, such as choosing appropriate procedures and checking assumptions).

AWARD #0206924
C-TOOLS: CONCEPT-CONNECTOR TOOLS FOR ONLINE LEARNING IN SCIENCE
Investigator(s): Douglas Luckie (luckie@msu.edu)
Co-PI(s): Diane Ebert-May, Janet Batzli
Sponsor: Michigan State University, East Lansing, MI

In this project, an interdisciplinary team of faculty from the College of Natural Science and the Lyman Briggs School at Michigan State University are working to develop and validate a new assessment tool for courses in biology, chemistry, physics, and geology. The idea is for students to construct concept maps of their own. The Concept Connector consists of a web-based, concept mapping Java applet with automatic scoring and feedback functionality. The tool enables students in large introductory science classes to visualize their thinking online and receive immediate formative feedback. The assessment tool and the methods of its application in the classroom are

being designed to motivate students to reflect, revise, and share their thinking with peers as an extension of the learning process.

We predict that faculty will become better able to identify their students' areas of incomplete, vague, or incorrect understanding of science; and that students will confront their misconceptions and become better able to reflect on, organize, and integrate their learning. Three primary goals of this project are to: 1) develop and validate an online concept mapping tool that can be used to provide immediate feedback (computer automated) to both students and instructors about student understanding of conceptual relationships, 2) detect and document students' misconceptions regarding relationships between concepts (e.g., interdisciplinary; relation between ecology and quantum physics, or discipline specific; structure and function of DNA), and 3) implement and sustain faculty development workshops to help faculty design curricula and instruction that can better assess student learning and enable students to use visual models to represent their understanding.

AWARD #0206943
ASSESSING STUDENT TRANSFER AND RETENTION OF LEARNING IN MATHEMATICS, PHYSICS, AND ENGINEERING COURSES
Investigator(s): Andrew Bennett (bennett@math.ksu.edu)
Co-PI(s): N. Sanjay Rebello
Sponsor: Kansas State University, Manhattan, KS

This project is based on the theme that assessment is most useful when it is closest to instruction, in both form and time. Summative assessments at the end of a program provide necessary feedback, but are often of limited value in guiding improved achievement. Such assessment takes place several years after instruction in basic concepts, and changes in instructors and instruction during the intervening time may make the feedback appear irrelevant. In addition, faculty members are often distrustful of reports where the faculty members lack first-hand knowledge of the assessment procedure. To be most effective in improving design and conduct of actual courses, the instructors in the course should use the assessment tools themselves, preferably during the conduct of the course. Because the main reason for STEM majors taking core engineering science courses is to prepare them for future courses, the most important aspect of assessment is the ability of students to transfer their learning to new contexts in later courses. Assessment should not only serve the purpose of the instructor or the institution, it should also aid the students in recognizing their own achievements and in guiding the students to improve their understanding.

The goal of this project is to design online assessment tools that can be easily integrated into core engineering science courses and that are capable of answering the following questions:

1) What specific material have the students learned in core engineering science courses in mathematics and physics?

2) What understanding do the students have of the material they have learned? Is it just disconnected facts and procedures, a broad conceptual picture informed by careful understanding of the details, or something in between? If it is something in between, can we describe more exactly what understanding they have gained?

3) How much (and what type of) knowledge do the students retain after specific classes have ended?

4) Can the students use the material they have learned in new situations in their professional courses? How consistently do they use the understanding developed in core engineering science courses when encountering these ideas in new contexts? In the best case, can we predict in advance whether students have gained the necessary understanding to successfully apply their knowledge in new contexts?

Early versions of the sought-after assessment tools have been developed under an earlier grant: "Technology- & Model- Based Conceptual Assessment: Research on Students Applications of Models in Physics and Mathematics" funded by the NSF ROLE program (REC-0087788). Model Analysis is being used to develop and validate the tools. These tools are providing great insight into student conceptual understanding and learning styles, and this project is expanding their use to more core classes, and extending the focus from basic research on student learning to assessment of learning and conceptual understanding. The assessment tools we are developing are providing real-time feedback to both instructors and students, enabling both to adjust the teaching and learning process to improve student achievement in terms of conceptual understanding and the ability to transfer learning to new contexts.

AWARD #0206952
ASSESSING, UNDERSTANDING, AND IMPROVING THE TRANSFER OF LEARNING IN UNDERGRADUATE MATH, SCIENCE, AND ENGINEERING
Investigator(s): Bradford Lister (listeb@rpi.edu)
Co-PI(s): William Siegmann, Kenneth Connor, Karen Cummings, Frank Lee
Sponsor: Rensselaer Polytechnic Institute, Troy, NY

This project is examining students' ability to transfer knowledge and skills within the engineering curriculum. It is focused on how much knowledge is transferred from Calculus, Differential Equations, and Physics to subsequent engineering courses. We are developing a series of diagnostic exams that measure the degree of transfer of basic skills and concepts from these areas into Engineering Dynamics, and Fields and Waves. These two courses were chosen because of their wide application in engineering and because of their extensive use of concepts from these earlier courses. We are also exploring the strength of relationship between high grades and a student's ability to transfer knowledge from Calculus, Differential Equations, and Physics. We are evaluating classroom practices that strengthen transfer of skills. Among the products of this project is the "Transfer Environment and Student Readiness Instrument" being developed at RPI's Anderson Center for Innovation in Undergraduate Education. This instrument assesses a range of environmental and affective variables known to correlate with improved transfer capability. It is meant to help students evaluate their own perception about their courses as being integrated or disconnected and fragmented.

AWARD #0206947
ASSESSING IMPLEMENTATION OF COLLABORATIVE LEARNING STRATEGIES
Investigator(s): Eric Mazur (mazur@physics.harvard.edu)
Sponsor: Harvard University, Cambridge, MA

In recent years, problems in American science education, from elementary schools to universities, have been widely publicized and have aroused great concern. It has been repeatedly established that students often fail to master key concepts under the lecture approach that is commonly used in high school and undergraduate science courses. Some progress has been made in developing alternative pedagogies based on more active forms of learning, which are more effective in producing student learning. However, such pedagogies are still not widespread. A number of strategies have been developed to improve the teaching of introductory science courses, several of which can be used in a variety of disciplines, and extensive assessment has demonstrated the effectiveness of these strategies. Over the past ten years, Peer Instruction has been established as one effective way to improve student learning in undergraduate science courses by including collaborative exercises within the context of traditional lectures. Peer Instruction is presently used by hundreds of instructors around the world, and the majority of those instructors testify to its effectiveness and ease of implementation.

The goal of this project is to assess the effectiveness of Peer Instruction at a variety of institutions and to determine the implementation factors that contribute to its success. The results will improve understanding of what makes Peer Instruction work and, thus, allow for more effective implementation of Peer Instruction. In addition, the work serves to increase the body of research on collaborative learning and to stimulate additional faculty to begin using collaborative teaching methods such as Peer Instruction. Assessment tools are being developed and enhanced to efficiently assess student achievement on standard conceptual instruments. The project involves a number of collaborating faculty from across the country whose courses are participating in the focused study of Peer Instruction. The results will enable science faculty around the nation to teach more effectively with Peer Instruction. The results thus have the potential to impact not only the 50,000 students who are currently taught using Peer Instruction, but also the more than 150,000 students of faculty who have expressed interest in using Peer Instruction in their classes.

AWARD #0243265
A PILOT INVESTIGATION OF FUNCTIONAL ROLES ON ENGINEERING STUDENT TEAMS
Investigator(s): Linda Schmidt (lschmidt@eng.umd.edu)
Co-PI(s): Janet Schmidt, Paige Smith
Sponsor: University of Maryland College Park, College Park, MD

The project team is fast becoming a primary learning environment for engineering students. From the perspective of education, positive team project experiences can motivate students to perform at higher levels. The definition of what factors specifically contribute to a successful team experience in engineering education requires further clarification and empirical study. One of the most consistent aspects of the literature on teams is the importance of team roles to successful team functioning. The notion behind the importance of team roles is that certain predictable processes and behaviors must occur if a team is to thrive. Increasing team members' awareness of and training in these roles improves the chances that the appropriate behaviors will occur and the team will

meet its goals. This project is an exploratory study seeking to better understand and empirically demonstrate the significance of functional roles to engineering project teams.

This study uses a focus group-based investigation of a variety of engineering team classes from freshman to seniors. In addition, focus groups of major engineering disciplines, women, minorities, professional engineers, and students who do not persist in an engineering program, are being conducted to identify individual differences in the team experience and role taking by these diverse constituencies. Additionally, the instructors of the capstone courses are being individually interviewed by the Principle Investigators. If the presence and importance to team functioning and student learning of functional roles is empirically supported, then the pilot study will be followed by a larger scale proposal to create a psychometrically appropriate assessment measure for evaluating student performance in functional roles.

AWARD #0243258
FORMATIVE ASSESSMENT MATERIALS FOR LARGE-ENROLLMENT PHYSICS LECTURE CLASSES
Investigator(s): Thomas Greenbowe (tgreenbo@iastate.edu) , David Meltzer (Former Principal Investigator)
Sponsor: Iowa State University, Ames, IA

This project is developing new formative assessment materials for large introductory lecture-based general physics courses. Among the project goals are to 1) analyze the reliability and validity of these materials, 2) evaluate their effectiveness in the process of instruction, and 3) acquire baseline data regarding student performance that will be of value to other instructors who make use of the materials.

The assessment materials themselves consist of carefully sequenced sets of multiple-choice questions, each focused on a specific topic. The individual items are primarily conceptual questions that downplay algebraic manipulations, and instead make heavy use of diagrammatic, graphical, and pictorial elements. The materials are intended for use in large lecture classes, and they are specifically designed to allow for rapid and reliable assessment of student learning during the course of a single class. The structure and sequencing of the questions are formulated to maximize intense student-instructor interaction on a minute-by-minute basis even in large-enrollment classes. The materials will be used in classes organized along active-learning lines in which a classroom communication system is available. This type of system, either electronic or one based on flash cards, allows students to rapidly signal multiple-choice responses to questions posed by the instructor. The instantaneous feedback they provide will allow instructors to make immediate alterations, as needed, in their presentations and in planned instructional activities. The materials and baseline data gathered from these materials will be made available via websites, CD-ROMs, and through other dissemination methods to reach the education community and other interested audiences.

AWARD #0243254
DESIGNING A PEER EVALUATION INSTRUMENT THAT IS SIMPLE, RELIABLE, AND VALID
Investigator(s): Matthew Ohland (ohland@purdue.edu)
Co-PI(s): Cynthia Finelli, Richard Layton, Lisa Bullard, Misty Loughry
Sponsor: Clemson University, Clemson, SC

A simple, reliable, cooperative learning group peer evaluation method is the focus of this project. The project addresses an issue that is significant for the effective use and wider acceptance of cooperative learning in engineering education. The instrument being designed and tested will not provide the same level of feedback as more complex and difficult-to-use instruments, but its reliability and validity should be comparable or higher than those of the best instruments currently available, and its simplicity will encourage widespread adoption by faculty who are not ready to make the commitment to more ambitious approaches. This project builds from what is known about cooperative learning. A multi-item instrument with good administration has good potential for becoming a successful instrument. The project is working on crucial issues in peer evaluation (inflation of scores, identical scores, and bias) and in reliability (number of raters, number of administrations). Because there is no true measure of the quality of teamwork, the project is using multiple assessment methodologies and seeking concurrence among them. Instrument validity is being established through verbal protocol analysis, behavioral observation, and concurrence with results obtained with other validated instruments. Both test-retest reliability and inter-rater reliability are being assessed. The project is being conducted in a distributive fashion in five participating institutions. There are nine faculty across these institutions, experimenting with different methods of assessment in eleven different courses. The breadth of courses and instructors promises to make the developed method robust.

This peer evaluation instrument is expected to significantly improve both team-based engineering education and assessment activities. Instructors using it are able to improve the teaching of team skills by giving students formative peer feedback. Engineering programs using it can document that their students have the ability to work in teams. The successful dissemination of the prototype instrument through publications and presentations has been promising. The simplicity, validity, and reliability of this instrument can be expected to speed its adoption at other institutions. The involvement of schools from multiple Engineering Education Coalitions is also likely to accelerate its dissemination. The assessment instrument and its associated rubric and instructions are also of potential interest to a large number of academic programs. The instrument under development is useful anywhere cooperative groups are employed. Extensive testing for validation and reliability evaluation for this tool should also minimize any unintended gender or racial bias.

AWARD #0243227
ADAPTATION OF CONCEPT INVENTORIES FOR RAPID FEEDBACK AND PEER-ASSISTED LEARNING IN CORE ENGINEERING COURSES
Investigator(s): John Chen (jchen@rowan.edu)
Co-PI(s): Dexter Whittinghill, Jennifer Kadlowec
Sponsor: Rowan University, Glassboro, NJ

Our study is being conducted under the Adaptation track of the Assessment of Student Achievement in Undergraduate Education Program. The goal of our project has been to adapt the Concept Inventory for frequent classroom use, and to implement it in a system to provide rapid feedback to students of their understanding

of key concepts during classroom lectures. This rapid feedback system acts as the focal point and catalyst to encourage students, working in pairs, to assist each other in correcting misconceptions or deepening each other's understanding of the topic at hand. Furthermore, the system allows the professor to assess the students' level of comprehension (or misconception) in a just-in-time fashion, and thus guides his or her pacing and coverage of the material. The system and methodology has been implemented in at least three core engineering courses at two institutions.

AWARD #0243209
WRITING FOR ASSESSMENT AND LEARNING IN THE NATURAL AND MATHEMATICAL SCIENCES
Investigator(s): Nancy Simpson (n-simpson@tamu.edu)
Co-PI(s): Michael Stecher, A. Lewis Ford, Arlene Russell, Comer Patterson
Sponsor: Texas Engineering Experiment Station, Texas A&M University, College Station, TX

The Writing for Assessment and Learning in the Natural and Mathematical Sciences (WALS) project adapts the Calibrated Peer Review (CPR), developed and widely disseminated in chemistry. The project creates and implements writing assignments for students in biology, mathematics, and physics. The assignments assess student understanding, enhance student learning, and promote faculty inquiry into how students learn.

The WALS project involves carefully crafted writing assignments that promote and assess student conceptual learning and faculty-guided peer review that promote both critical thinking and student self-assessment. Through the workshops and writing activities, faculty participants increase their own understanding of how and what students understand. Synthesizing multiple classroom research projects provides significant insight into the nature of student learning in the sciences and mathematics. The faculty participants engage in activities lead by a team that represents expertise in the prescribed content areas, faculty development, the CPR protocol, writing assignments development, and technology-mediated instruction.

The broader impacts of the WALS project are realized through the collaboration between faculty and students from the Texas Collaborative for Excellence in Teacher Preparation partner institutions. Of these, three are Hispanic serving, one a historically black university, and another 90 percent female. Diversity issues relating to the implementation of CPR are being investigated. The project impacts approximately 5,000 students taking the affected courses over the project term, several of which are taken by future science and mathematics teachers. The reach of the project therefore extends to the students of these pre-service teachers.

AWARD #0243207
VALUING LITERACY: THE SCIENCE AND MATHEMATICS VALUE INVENTORY (SaM-VI)
Investigator(s): Donald Deeds (ddeeds@drury.edu)
Co-PI(s): Charles Allen, Mark Wood, Bruce Callen, Vickie Luttrell
Sponsor: Drury University, Springfield, MO

This project introduces the Science and Math Values Inventory (SaM-VI), a comprehensive inventory that determines the value undergraduate students place on science and math literacy. Although this instrument is useful in examining educational outcomes for all students, the main aim is for those who do not major in STEM

fields. The SaM-VI provides a mechanism to review, modify, and enhance interdisciplinary math/science curricula. More importantly, this inventory supports national curricular reform and improved literacy in mathematics and science by providing universities and colleges a means to determine if their general education math and science programs have truly made a difference in the lives of students. It also reinforces the constructivist tenet that the effective application of knowledge depends not only on how much knowledge students have and how they organize it, but also on the feelings they associate with that knowledge and especially the value they place on it.

The major component of this project is the development of the Science and Math Values Inventory (SaM-VI). This inventory investigates how (or if) the value students place on math and science literacy changes following formal instruction. This project has five significant outcomes. The development of the SaM-VI allows: 1) the assessment of the relationship between the perceived value of science and math literacy and achievement, 2) the assessment of longitudinal changes in students' perception of the value of science and math literacy, 3) the examination of gender-related differences in perceived value of science and math literacy, 4) the exploration of the impact of ethnicity on perceived value of science and math literacy and, 5) the dissemination of an assessment tool that 35 partner institutions plan to use to evaluate curricula.

AWARD #0243184
CURRICULUM IMPROVEMENT IN PRACTICE-ORIENTED BIOLOGY AND COMPUTER SCIENCE PROGRAMS USING STUDENT PORTFOLIOS
Investigator(s): Kostia Bergman (k.bergman@neu.edu)
Co-PI(s): Viera Proulx, Melvin Simms, Veronica Porter
Sponsor: Northeastern University, Boston, MA

This project moves the assessment of practice-oriented science education from indirect to direct measures of student learning. Northeastern University is a major research university with a long tradition of practice-oriented education. Through its well-established cooperative education program, most students alternate periods of full-time course work with periods of full-time paid employment relevant to their major. Even in on-campus settings science education has moved beyond the traditional classroom, where instructors maintain direct oversight over students, to include new forms of lab instruction, collaborative research projects, and independent study. Both the cooperative education model and the other new forms of instruction raise questions about what is learned in each setting, how the acquired skills can be assessed, and how this learning can be integrated into the overall curriculum.

The College of Computer Science and the Biology Department of the College of Arts & Sciences at Northeastern University have surveyed students and their co-op employers to evaluate student skills and determine where and how they were learned. The two units have become campus leaders in assessing learning to improve curricula and teaching. Building on this base, this project develops portfolios for students to showcase and archive their work. Student learning is then evaluated according to well-defined rubrics that set standards of measurable progress toward proficiency in desired learning goals. The portfolios are stored in digital form on computer servers (so-called electronic portfolios) to facilitate communication between the various learning environments. Development of rubrics to evaluate student learning in different settings is the first major task of the project. This process requires in-depth, facilitated discussions among academic faculty, co-op faculty, employer supervisors, and experienced students. The second task is the designing of the portfolio system prototype

following the recommendations of the rubric development teams and working with an outside database and web-front-end designer. The portfolio is piloted during the third year of the project, using 200 students in computer science co-op preparation classes, and 200 Biology students from all five years of their program. During the last two months of each year the project team reviews progress by consulting with curriculum committees, employer groups, student groups, and an Advisory Committee formed for this project. After the final evaluation the results are disseminated to the university community and to the wider audience of scientists and educators. The process of rubric development and portfolio construction can be used as a model to understand the relative contribution of different settings to desired student learning outcomes, and as a vehicle for regular review and improvement of curriculum. The model can be used in a variety of settings with a broad range of students; for instance to help improve skills and clarify goals for students who are the first in their family to attend college and students for whom English is their second language.

AWARD #0243126
DIAGNOSTIC QUESTION CLUSTERS: DEVELOPMENT AND TESTING IN INTRODUCTORY GEOLOGY AND BIOLOGY
Investigator(s): Joyce Parker (parker@lite.msu.edu)
Co-PI(s): Duncan Sibley, John Merrill, Merle Heidemann, Gerd Kortemeyer
Sponsor: Michigan State University, East Lansing, MI

Diagnostic assessment is the foundation for instructional improvement. It alerts instructors to student difficulties in learning particular concepts. In addition, truly diagnostic assessment yields information as to why students have difficulties, identifying common conceptual barriers students encounter. This information can be used to guide instructors in making changes to their teaching. This project is working to achieve two major goals. One is the development of a generalized system for creating and using diagnostic assessment in an undergraduate discipline. The other is to implement this system in geology and biology.

Towards the first goal we are piloting a procedure for developing diagnostic questions that can be used in any field, developing protocols for online peer review and publication of diagnostic questions. Further, we are programming the LON CAPA network to support the online peer review, and publication of diagnostic questions and the export of data on students' responses to questions. (LON CAPA is the Learning Online Network with Computer Assisted Personalized Assistance. It is a system developed at Michigan State University for delivery of online instructional materials that is freely available to other institutions.) Finally, we are engaging faculty in other institutions in development work that will make them aware of the diagnostic question clusters and help them learn how to integrate diagnostic questions into their teaching.

This project is responding to the dearth of diagnostic questions in geology and biology by producing clusters of diagnostic questions addressing topics in geology and biology. These clusters cover three topics in geology: systems, the global carbon cycle, and the water cycle. In biology they cover two large topics: nitrogen cycling and carbon cycling. Each of these two are being covered at three different levels: global, macro, and micro. We are working with biology and geology faculty, teaching them to use the diagnostic questions as part of their instruction in order to learn about the conceptual barriers encountered by their students and modify their instruction in light of this information. Our ultimate goal is to have faculty in all fields using this procedure for developing diagnostic questions and adding to the pool. Those who use LON-CAPA will be aided by the new data export features. The LON-CAPA library of diagnostic questions is a permanent place for storing peer-reviewed questions available to STEM teaching faculty.

AWARD #0241078

COLLABORATIVE RESEARCH: USING FORMATIVE ASSESSMENT TO DEVELOP INTRODUCTORY PHYSICS SKILLS

Investigator(s): Eugenia Etkina (etkina@rci.rutgers.edu), Xueli Zou (xzou@csuchico.edu)

Co-PI(s): Alan Van Heuvelen

Sponsor: Rutgers University New Brunswick, New Brunswick, NJ , California State University, Chico Research Foundation, Chico, CA

This collaborative effort is developing activities that integrate knowledge building and formative assessment in the evaluation of learning at the introductory physics level. Included in this effort is the development of research-based formative assessment tools. These tools involve active engagement methods with appropriate feedback in the form of test kits of activities that are common in the practice of science and engineering. Such activities have been found to be the most effective intervention to help students achieve desired learning outcomes. Each kit includes six different types of activities, along with templates and rubrics that instructors can use for evaluation, and that students can use for self and small-group assessment of their own work. The scoring rubrics assess students' development of science and engineering abilities, and allow instructors to evaluate outcomes, improve the quality of feedback (the most important part of formative assessment), and establish the validity of the activities—do they assess the desired skills. The components of the kits are being developed and tested in large-enrollment introductory physics courses in two universities (one large research and one medium comprehensive university) with 1,600 students each year, in a high school physics course (100 students/year), and in methods courses for pre-service teachers (15 teachers/year). Dissemination of the kits occurs via the web and by a publisher (hard copy and a CD), peer-reviewed papers, workshops, and talks at national meetings.

AWARD #0404924

COLLABORATIVE RESEARCH: TRANSFERABLE ASSESSMENTS FOR CAPSTONE ENGINEERING DESIGN

Investigator(s): Denny Davis (davis@wsu.edu)

Co-PI(s): Michael Trevisan

Sponsor: Washington State University, Pullman, WA

This project is creating a versatile system for reliable assessments of student learning outcomes for capstone design courses across engineering disciplines and institutions. It builds on a decade of engineering design education and assessment achievements of the Transferable Integrated Design Engineering Education (TIDEE) consortium in the Pacific Northwest. Prior to this project, the investigators engaged regional and national collaborators to: 1) develop a consensus profile of a top-quality engineer; 2) define engineering design learning outcomes at mid-point and end-of-program; 3) develop and test reliable assessment instruments for engineering design; and 4) facilitate dozens of workshops to help faculty define and assess engineering design learning outcomes. In this project, a multi-institution faculty team is developing assessment tools with input from a diverse set of expert consultants and stakeholders. Assessment instruments are being built around frequently used classroom assignments to measure achievement of high-level integrated performances in capstone design courses. Performance expectations are being derived from a profile of a top-quality engineer for applicability to a full range of capstone engineering design courses. The developed assessment instruments are being tested for quality in diverse institutions and student populations to demonstrate their transferability.

AWARD #0405007
BUILDING A BASIC BIOLOGY CONCEPT INVENTORY
Investigator(s): Michael Klymkowsky (klym@spot.colorado.edu)
Co-PI(s): Ronda Garvin-Doxas
Sponsor: University of Colorado at Boulder, Boulder, CO

This project is developing a basic Biology Concept Inventory (BCI) that will enable the field to reliably quantify student learning at the introductory college level. The objective is to arm the field with reliable data on student learning as biology departments around the country attempt to improve student achievement. The project is inspired by, and will follow the development strategy of, similar efforts in physics and astronomy, which have been highly successful in significantly improving teaching in these fields. The development of valid, reliable instruments relies critically on the identification of the dominant misconceptions students carry into the classroom with them. Misconceptions are incorrect mental models of physical phenomena and processes that students hold before instruction. They present a significant barrier to learning, and they are best addressed explicitly with specifically designed learning activities. Once these misconceptions have been identified, an experienced team of content experts and evaluators can, through repeated review and test cycles, develop valid instruments that can reliably diagnose students' misconceptions. These instruments can then form the basis of course and curriculum transformation efforts that aim to improve student achievement. Development and assessment of the instrument involves collecting data on misconceptions (essays, questionnaires, and interviews) from over 2,000 students in four schools, including a significant population of pre-service teachers.

The proposed project is using well-tested methodology to develop highly needed evaluation capacity for biology. The existence of a Biology Concept Inventory has the potential to impact the teaching of biology to thousands of undergraduates throughout the country, in the same way that the Force Concept Inventory and the Astronomy Diagnostic Test have impacted the teaching of Physics and Astronomy.

AWARD #0404988
CLASSACTION: A MODEL RAPID-FEEDBACK AND DYNAMIC FORMATIVE ASSESSMENT SYSTEM
Investigator(s): Kevin Lee (klee6@unl.edu)
Co-PI(s): Edward Schmidt, Michael Guidry, Timothy Slater, Todd Young
Sponsor: University of Nebraska-Lincoln, Lincoln, NE

The ClassAction project is developing a model for interactive classroom materials for formative assessment and rapid student feedback in introductory college science courses. Taking the form of carefully crafted multiple-choice questions in electronic databases, the formative assessment items are specifically designed for student voting that promotes collaborative discussion. Based on student misconceptions research, they are designed to allow rapid and reliable formative assessment of student learning in large classrooms equipped with electronic personal response systems. The questions are illustrated visually with animations and simulations and are dynamic in that instructors can easily transform them into alternative representations. Considerable background instructional resources are included in each module to allow instructors to provide feedback.

Initially set in the context of introductory astronomy for non-science majors, the questions target common student misconceptions that have proven difficult to overturn. The materials are designed using a

learning cycle approach to encourage students to elicit, confront, and resolve their misconceptions. These efforts effectively balance research informed pedagogy and creativity. The ClassAction project has considerable impact on the way in which formative assessment is used to improve astronomy instruction and the underlying design is easily transferable to other disciplines. The database infrastructure and created materials are being widely disseminated throughout the science education community.

AWARD #0404986
ADAPTING DIAGNOSTIC DIGITAL PORTFOLIO TECHNOLOGY TO TRACK ASSESSMENTS OF ADVANCED STUDENT LEARNING OUTCOMES IN ANALYSIS AND PROBLEM-SOLVING ABILITIES
Investigator(s): Lauralee Guilbault (lauralee.guilbault@alverno.edu)
Co-PI(s): Leona Truchan, Susan Pustejovsky
Sponsor: Alverno College, Milwaukee, WI

This project adapts Alverno College's Digital Diagnostic Portfolio (DDP) technology for advanced level SMT courses to address several goals. These include monitoring and building coherence within the major programs in SMT; identifying and describing how the implementation of DDP in SMT curricula affects assessment processes; identifying, describing, and comparing how DDP in non-SMT curricula affects assessment in SMT; and using DDP to study the differences and commonalties in analytic problem solving. This will provide important understanding of the ways in which adapting technological innovations impacts not only the natural science fields but, through comparative analyses, other disciplines as well. The wider benefit of this project outside of Alverno College will be in analyzing and describing the process of adapting the DDP assessment technology and in attempting to describe the effects of such assessment technology upon student learning and achievement.

AWARD #0404975
CHEMX: ASSESSING COGNITIVE EXPECTATIONS FOR LEARNING CHEMISTRY
Investigator(s): Stacey Lowery Bretz (bretzsl@muohio.edu)
Sponsor: Youngstown State University, Youngstown, OH

The heart of teaching and learning chemistry is the ability of the teacher to provide experiences that share a conceptually abstract, mathematically rich subject with novice learners. This includes not only chemistry concepts, but also knowledge about how to learn chemistry. Students' expectations for learning chemistry in the university classroom impact their success in doing so.

Physics education research has explored the idea of student expectations with regard to learning physics, resulting in the development of MPEX (the Maryland Physics EXpectation survey). We are adapting MPEX to develop a chemistry survey regarding student expectations for learning chemistry: CHEMX. In particular, CHEMX explores the role of laboratory in learning chemistry as shaped by Johnstone's work with the macroscopic, particulate, and symbolic representations of matter.

Data collection from university chemistry faculty, undergraduates, and graduate students in chemistry programs approved by the ACS Committee on Professional Training allows examination of differences in expectations across the disciplines of chemistry. Data collection from high school chemistry teachers examines the environment in which entering undergraduates develop their expectations. Collaborations with the POGIL Project

(Process Oriented Guided Inquiry Learning, NSF Award 0231120) and the MORE Project (Model-Observe-Reflect-Explain, NSF Award 0208029) focus upon explicit efforts to shift student expectations and explore correlations with student achievement.

AWARD #0404927
COLLABORATIVE RESEARCH: WRITING FOR LEARNING AND ASSESSMENT IN ENGINEERING DESIGN COURSES
Investigator(s): D. Millard (millard@rpi.edu)
Sponsor: Rensselaer Polytechnic Institute, Troy, NY

Our project focuses on communication assignments and formative assessment to improve the teaching of engineering design. Specifically, we use Calibrated Peer-Review (CPR)—an end-to-end computer-mediated learning environment developed at UCLA—that seamlessly integrates writing as a vehicle for critical thinking into a technical or content course. We are testing the premise that a series of well-designed communication assignments can serve as enablers for students to enact the "habits of mind" fundamental to professional engineering practice. These assignments fully exploit CPR's four web-delivered, guided-inductive workspaces that 1) teach students how to recognize levels of accomplishment for specific activities, 2) guide peer-review sessions that produce both qualitative and quantitative formative assessment data, and 3) encourage deep-structured student self-reflection both on the task product and on the task process.

AWARD #0404911
PROJECT CAT: ASSESSING CRITICAL THINKING
Investigator(s): Barry Stein (bstein@tntech.edu)
Co-PI(s): Joseph Redding, Ada Haynes
Sponsor: Tennessee Technological University, Cookeville, TN

The primary goal for this project is to build on past efforts and refine a promising assessment instrument to evaluate critical thinking skills, CAT (Critical thinking Assessment Test). The refinement of this assessment tool focuses on five critical areas: 1) Refining the CAT assessment tool so that it has high face validity when evaluated by a broad spectrum of faculty across the country in STEM and non-STEM disciplines, 2) Refining the CAT assessment using expert evaluation in the area of learning theory and learning sciences to establish construct validity, 3) Refining the CAT assessment so that it continues to have high criterion validity when compared to other instruments that measure critical thinking and intellectual performance, 4) Refining the CAT assessment so that it is culturally fair, and 5) Refining the CAT assessment so that it has high reliability. In the process of achieving these objectives, Tennessee Technological University is partnering with six other institutions across the country (University of Texas, University of Colorado, University of Washington, University of Hawaii, University of Southern Maine, and Howard University) to use and score a critical-thinking assessment test that provides insights into their own students' critical-thinking skills. Schools were carefully selected to provide geographic, ethnic, racial, public versus private, socio-economic, and size diversity. Once the assessment tool is fully developed the project can become self-sustaining by providing the test at cost to other institutions. There is little question that as a result of an increasingly technological and information driven society the ability to think critically has become a cornerstone to both workplace development and effective educational programs. Critical thinking is central to both the National Science Standards and the National Educational Technology Standards. Despite the central importance

of critical thinking in the workplace and education, existing assessment tools are plagued by problems related to validity, reliability, and cultural fairness. According to Bransford et al. [1], "a challenge for the learning sciences is to provide a theoretical framework that links assessment practices to learning theory." The intellectual merit of this project is to refine such an assessment device for critical thinking based upon current theories of learning and cognition. Another merit of this project is that it seeks to refine an assessment device for critical thinking that has high face validity for a broad spectrum of faculty across the country. The perceived validity of the assessment tool is important because it will have a direct bearing on the motivation of faculty to improve their students' critical thinking skills. The broader impact of this project is that by providing an instrument for assessing critical thinking and problem-solving skills to other universities across the nation, one can encourage improvements in the quality of students' critical-thinking and problem-solving skills. Institutional goals for improving education are often assessment driven, hence, the development of a valid, reliable, and culturally fair critical-thinking assessment tool would encourage more institutions to focus on the teaching of critical-thinking skills in universities across the country. These higher order thinking skills are essential in an economy that has shifted from manufacturing to information technology and services.

REFERENCE

[1] Bransford, J. D., Brown, A. L., and Cocking, R. R., (eds.), How people learn: Brain, mind, experience, and school, (*National Academy Press*, Washington, D.C.), 2000.

AWARD #0404818

DEVELOPMENT AND VALIDATION OF A CALCULUS CONCEPT INVENTORY TEST

Investigator(s): Jerome Epstein (jepstein@duke.poly.edu)
Co-PI(s): Deane Yang
Sponsor: Polytechnic University of New York, Brooklyn, NY

This project designs, develops, pilots, and validates a test, the Calculus Concept Inventory (CCI), that measures conceptual understanding of the few most basic principles of differential calculus. Both the framework for the CCI and its development and validation are rooted in the experience of the Force Concept Inventory (FCI) in physics. Both tests are based on the most common student conceptual misunderstandings that block any meaningful ability to understand and apply the subject. It is essential that schools and faculty have a validated test to measure whether students leaving calculus have conquered these misconceptions.

The great intellectual merit is in the use of modern scientific methods to both develop and validate such an instrument, and the investigators are well versed in these methods. The FCI has had a dramatic effect in improving physics education, and is abundantly shown to be able, in a reproducible way, to document results dramatically better from Interactive-Engagement methodologies than from standard Lecture-Demonstration approaches. It is very important to know whether this critically important finding remains valid in calculus (and other math courses). The CCI is able to do this. Such an instrument is needed in mathematics. The ability to document convincingly whether a given teaching methodology really does what it claims to do (in mathematics as well as physics) has a broad national impact. Without such documentation, decisions on how to teach are based largely on the personal faith of faculty, and subjective decisions. A well-validated CCI, which this project provides, gives an independent and reproducible measure of whether teaching methodology is the dominant factor for

conceptual understanding in calculus that the FCI has shown it to be in physics. The CCI lends itself to greatly improved teaching of calculus to thousands of college and high school students and the spawning of similar conceptual tests, also needed, in other parts of the math curriculum.

AWARD #0404802
MEASURING STUDENT AND FACULTY ENGAGEMENT IN ENGINEERING EDUCATION
Investigator(s): Norman Fortenberry (nfortenb@nae.edu)
Sponsor: National Academy of Sciences, Washington, DC

The Center for the Advancement of Scholarship on Engineering Education (CASEE) of the National Academy of Engineering (NAE) is building upon the National Survey of Student Engagement, the Faculty Survey of Student Engagement, and the ABET EC2000 Study in order to develop integrated assessment tools that will measure faculty and student engagement in instructional and learning practices that may correlate with desired student learning outcomes in undergraduate engineering programs. To provide the context and basis for these assessment tools, we are developing: 1) a set of synthesized learning outcomes for undergraduate engineering education that reflect the desires of the academic and employer communities; and 2) a pilot database summarizing the education research literature that links desired student outcomes to specific educational "best practices". The outcome of this work will be an integrated set of pilot instruments, vetted with a representative cross-section of the engineering community, to assess the current state of instructional practice and student learning outcomes.

AWARD #0511940
COLLABORATIVE RESEARCH: ASSESSING CONCEPT KNOWLEDGE AND ATTITUDES IN INTRODUCTORY COMPUTER SCIENCE COURSES
Investigator(s): Stephen Cooper (scooper@sju.edu)
Sponsor: St Joseph's University, Philadelphia, PA

This project is developing two assessment instruments to measure student-learning outcomes and student attitudes in introductory computing courses. One instrument measures student-learning outcomes for introductory programming courses and one measures students' attitudes toward computers and computer science. Each is designed to measure fundamental concepts that are not language specific. The validity and reliability of the resultant instruments is being demonstrated through extensive testing.

The project is based on the need to devise new assessment tools and to update two-decades-old assessment tools for computer science education. The need for new and updated assessment tools is particularly crucial in a dynamic field where changes in, and availability of, computing technology has reverberating effects on pedagogy and student experience in the classroom. The learning outcomes instrument is based on the content domain defined by the IEEE/ACM Computing Curriculum 2001 for a first course in object-oriented programming.

A primary goal of many computer science education projects is to determine the extent to which a given instructional intervention has had an impact on student outcomes. However, valid and reliable assessment instruments that measure the desired goals and outcomes across different platforms are not currently available. This project is filling that gap. Careful attention is dedicated during the validation process to the impact that gender and ethnicity have on the validity of the resultant instruments.

AWARD #0512526

COLLABORATIVE RESEARCH: ADAPTING IMMEX TO PROVIDE PROBLEM-SOLVING ASSESSMENT MATERIALS FROM THE ACS EXAMS INSTITUTE

Investigator(s): Thomas Holme (tholme@csd.uwm.edu)

Sponsor: University of Wisconsin-Milwaukee, Milwaukee, WI

This project is a collaboration between the American Chemical Society (ACS) Examinations Institute, Clemson University, and the IMMEX project, to develop an alternate assessment instrument for the first two years of college chemistry by adapting IMMEX technology. The goals and objectives consist of the following. 1) IMMEX cases with content emphasis on chemical structure and function are being collated and adapted for summative problem-solving assessment, and additional cases are being devised as needed to complement current cases with structure and function content bases. 2) Analysis of dynamic problem-solving behavior for individual cases within IMMEX is being adapted to allow the measurement of the structure and function cognitive domain across several cases, and cognitive measures. 3) The internal reliability and validity of the comparison of structure and function problem solving is being established across multiple IMMEX cases using a combination of measures including Item Response Theory analysis as well as performance and progress measures derived by neural network and hidden Markov modeling. These include measurement rubrics for transitions between dynamic problem solving states for students to determine gains in the use of successful problem-solving strategies for students in the first two years of college chemistry. 4) A remote IMMEX server is being established at the offices of the ACS Examinations Institute to enhance the ability to provide IMMEX cases to a growing national pool of students. 5) A reporting system is being devised that will be published from the Exams Institute to provide instructors with reports of the gains in successful problem-solving strategies for their students participating in this project.

This project is grounded in the prior research on assessment of problem solving using IMMEX technologies. A range of problems, based on the relationship between structure and function, is being developed and the resulting assessment data are being used to develop predictive models for individual student problem solvers. The resultant database of student problem-solving behavior also provides a resource for chemical educators.

This project involves the extension of new assessment technologies to a national population. The resulting reliable, normed assessments from a trusted source (the ACS Examinations Institute) that focus on deeper aspects of problem solving will be available to a broader audience. The availability of these instruments is expected to facilitate the spread of reformed curricula that emphasize deep learning and problem solving.

AWARD #0512686

COLLABORATIVE PROPOSAL: THE SIGNALS AND SYSTEMS CONCEPT INVENTORY

Investigator(s): Kathleen Wage (kwage@gmu.edu)

Co-PI(s): Margret Hjalmarson

Sponsor: George Mason University, Fairfax, VA

This project is refining, validating, and disseminating an assessment instrument, the Signals and Systems Concept Inventory (SSCI), for a core course in the electrical and computer engineering curricula. The SSCI, modeled after the Force Concept Inventory for Newtonian mechanics, consists of 25 multiple-choice questions

emphasizing conceptual understanding, not rote problem solving. The incorrect answers, or distractors, capture common student misconceptions. A group of faculty members from 12 diverse institutions, referred to as the Development Team, is providing pre-test and post-test SSCI scores linked to grades, gender, and race data. These data are being used to examine the construct validity of the instrument and to check for evidence of bias in the results. The Development Team is assessing the content validity of the instrument through expert peer review of the SSCI questions. A test/re-test protocol is being used to calibrate the reliability of the instrument. Student interviews are providing additional data for the validity analysis. The group is preparing an SSCI Instructor Manual that will describe which student misconceptions are captured by each distractor. This manual will also guide the interpretation of results by providing baseline data on student performance and pre-test/post-test gains for the SSCI. Annual meetings of the Development Team are being used to monitor progress toward the project's objectives, each of which has been translated into a set of measurable outcomes and evaluation questions. Journal articles, conference presentations, the signals-and-systems.org website, and workshops at signal processing and engineering education conferences are being used to disseminate the SSCI and the study results.

AWARD #0512527

AN SRL PERFORMANCE-BASED ASSESSMENT SYSTEM FOR ASSOCIATE DEGREE ELECTROMECHANICAL ENGINEERING TECHNOLOGY STUDENTS

Investigator(s): Seymour Blank (sblank@citytech.cuny.edu)
Co-PI(s): John Hudesman, Barry Zimmerman
Sponsor: CUNY New York City College of Technology (NYCCT), Brooklyn, NY

This project is adapting a social cognitive self-regulated learning model to assist students in developing and assessing their SRL skills more effectively. For six years, the SRL program at NYCCT has concentrated on developing students' metacognitive skills in ways that enable them to become motivated, goal oriented, resilient, and successful. Using the SRL model as a framework, a variety of instructional interventions have been implemented with successful outcomes for first-year students enrolled in remedial math and writing courses. The SRL program has also incorporated the use of technology by developing an SRL software application for personal digital assistants and a student-focused SRL website. This project is building on earlier initiatives by developing and implementing an individualized SRL assessment system for a sequence of first-year electromechanical courses. Many incoming students at institutions like NYCCT are academically underprepared, resulting in high rates of poor academic achievement and attrition. (Within the school's EM program, only 15 percent of the students earn an associate degree within 5 years.) Students who become skilled, self-regulated learners develop a greater understanding of how to monitor and manage their learning processes. Attainment of SRL skills has been shown to have a higher correlation with academic success than previous grades or SAT scores.

We expect students to be able to more accurately assess changes in their SRL skills and link these changes to performance-based outcome measures including weekly quizzes and major examinations. Because self-regulated learning entails the cyclical use of feedback to plan, implement, and evaluate learning activities, the SRL Performance Assessment System focuses on providing detailed, content-specific feedback to the student within multiple learning opportunities. The project is creating an assessment form designed to help students develop and track the effectiveness of their SRL skills and improve self-evaluative beliefs. Ultimately, this form will be integrated with an EM/SRL website that will provide each student with a personal profile. A comprehensive evaluation plan is being implemented throughout the project, including formative and summative elements. The SRL Performance

Assessment System's predicted outcomes include: 1) improved academic achievement, 2) higher self-efficacy ratings on quizzes, 3) shifts in attributions for incorrectly solved EM problems from uncontrollable to controllable causes, 4) improved selection and articulation of appropriate academic strategies, 5) more accurate estimates of academic performances, and 6) greater self-satisfaction with academic performance.

Throughout the project, dissemination efforts at the local, regional, and national levels have been made. On a local level, the SRL program is working with the dean and faculty of the School of Technology to incorporate the assessment system into other academic departments. Based on past success with the math and English departments and the anticipated success in electromechanical engineering, we plan to work with colleagues to achieve transfer into other academic departments. On a regional level, the assessment work is being shared at educational research and technology meetings, e.g., the Annual CUNY Technology and the Classroom Conference. On a national level, the findings are being shared at forums such as the annual meetings of the American Education Research Association and the Institute of Electrical and Electronics Engineers. We are also writing scholarly articles for publication in refereed journals.

AWARD #0512596
ASSESSING PEDAGOGICAL CONTENT KNOWLEDGE OF INQUIRY SCIENCE TEACHING
Investigator(s): David Schuster (david.schuster@wmich.edu)
Co-PI(s): William Cobern, Renee Schwartz, Ralph Vellom, Edward Applegate
Sponsor: Western Michigan University, Kalamazoo, MI

Inquiry science teaching refers to teaching methods (pedagogy) that reflect the investigative attitudes and techniques that scientists use to discover and construct new knowledge. Undergraduate students are routinely assessed for their science content knowledge. A complementary component of our national effort to improve science education is the ability to effectively assess teachers' pedagogical knowledge of inquiry science teaching. Currently, such an assessment tool is not available. This project is developing a field-validated, objective assessment tool for testing undergraduate pre-service elementary teachers' pedagogical knowledge of inquiry science teaching (POSIT, the Pedagogy of Science Inquiry Teaching test). The assessment items can be used both in summative evaluation and as a formative tool in undergraduate instruction.

The U.S. Department of Education has recently called for educational revisions to be informed by research and scientifically-based evidence. This project helps to meet that standard. It involves a careful plan for test development followed by two rounds of piloting and revision, concluding with a blinded field-test classroom observation component for studying the validity of POSIT. The intellectual merit of the project lies in its coherent combination of innovative and research-based features in its design—a design that includes features of authenticity, problem-based learning, epistemology, exemplification, methodology, and summative and formative functions. The project is focused with a clear goal—one that serves a national need as articulated by NSF, the National Research Council, and the AAAS. This project is also interdisciplinary. Participating are ten universities, several geographically and demographically diverse public school districts, and a national panel of experts. The final instrument will conform to The Student Evaluation Standards and the Standards for Educational and Psychological Testing.

The "gold standard" for pedagogy assessment is what teachers actually do in the classroom. POSIT is being blind-tested against such a standard. Once it has been developed and validated against the criteria by observation of classroom practice, educators at any undergraduate institution could use it to evaluate and improve science teacher education programs. Improving the preparation and quality of teacher graduates will help ameliorate inequalities in schools that serve underrepresented populations, by increasing the number of teachers well qualified to teach science. POSIT will also be of value to researchers in science education seeking to improve the education and practice of science teachers. Professors teaching undergraduate and graduate SMET courses may find POSIT useful. Previous research indicates that science professors can be motivated to improve their own teaching practice as a consequence of considering how science is best taught to young people.

Project Goals:

1) Develop criteria for items reflecting the pedagogy of inquiry to guide item development and evaluation.
2) Compose, pilot, and revise a set of problem-based objective items for K-8 science curricula, based on realistic teaching vignettes of elementary science teaching, and meeting inquiry item criteria.
3) Pilot the test items with undergraduate, pre-service teachers reflecting racial and gender diversity.
4) Establish standardized scoring and test administration directions.
5) Establish initial estimates of reliability.
6) Validate the instrument by studying its predictive power with respect to actual teacher practice of inquiry teaching in classrooms.
7) Disseminate assessment items and project reports via the Internet, conference presentations, and journal articles.

AWARD #0512725
ASSESSMENT OF MODEL-BASED REASONING IN BIOLOGY
Investigator(s): Randall Phillis (rphillis@bio.umass.edu)
Co-PI(s): Neil Stillings
Sponsor: University of Massachusetts Amherst, Amherst, MA

Approximately 700,000 college and university students enroll in introductory biology courses each year. The majority of these courses deliver the basic facts and terms of biology to large numbers of passive students. Few of these courses have learning scientific reasoning skills among their stated objectives. This project is developing a set of teaching tools and tests in biology centered around the primary learning objective of model-based reasoning (MBR), the central intellectual activity of professional biologists. The intellectual merit of the project lies in the development, evaluation, and dissemination of a set of methods and tools for teaching and testing model-based reasoning in college-level introductory biology courses. An independent panel of experts, drawn from among professional biologists nationwide, is rating model-based reasoning questions. These expert ratings are compared with student performance in a classroom in which MBR problems are used for teaching and assessment. This study uses open-ended essays to investigate changes in students' descriptions of their reasoning process at several points during the course. Improvements in reasoning skills are being compared between students in MBR-based courses and traditional lecture-based instruction. A series of valid, multiple-choice summative examinations designed specifically to assess model-based reasoning skills are also being developed. Model-based reasoning instruction has the potential for broad impact in introductory biology courses nationwide. It exploits the strengths of two technologies that are being widely adopted. The first is web-based course support, and the second is the use of Classroom Communications Systems (personal response devices) in the lecture hall. In partnership with a major publisher, a teacher's guide and student study materials are being published.

AWARD #0512776

ASSESSING STUDENT TEAM EFFECTIVENESS

Investigator(s): P.K. Imbrie (imbrie@purdue.edu)
Co-PI(s): Susan Maller, Heidi Diefes-Dux
Sponsor: Purdue University, West Lafayette, IN

The project is developing both an instrument and a procedure that are specifically designed to allow students to assess their individual and team effectiveness using a web-based peer evaluation process. The process is being designed to correct for rater-bias (e.g., halo, leniency, severity) automatically. With this instrument, engineering faculty can determine: 1) how successful a teaming experience actually is for participating students, 2) the impact of team training methods on the teaming experience, 3) what team formation strategies best promote course learning objectives, and 4) the types of intervention strategies that will improve overall team functionality. Indicators of effective teams are being based on peer evaluations for each team member, as provided by other team members. Vignettes are being used to investigate the accuracy of students' ratings of team members, and to correct for rater bias, as compared to expert judgments, in actual peer ratings. Students' perceptions of functionality are being operationalized in terms of a self-report instrument requiring students to indicate the degree to which their team is working together across a range of domains, including interdependency, learning, potency, and goal setting. In evaluating their project, the investigators are comparing the attitudes of students who have used the instrument to those of earlier students who did not. They also are cross-validating effectiveness prediction with an instructor's observations of team effectiveness and with the students' course grades. In addition, an Assessment Review Panel meets semi-annually to monitor progress and identify problems. The approach is being tested at three other institutions. The investigators plan to present and publish their work in both engineering education and general higher education research venues.

Index

AAAS (American Association for the Advancement of Science) science education benchmarks, 6
AAC&U (Association of American Colleges and Universities), 22
accountability for student learning, 4, 16–26
Accreditation Board for Engineering and Technology (ABET), 82–84, 190–93, 240, 274
accreditation issue, 4, 23–25, 274
Achenbach, Joel, 13
achievement (see learning, student)
active and cooperative learning (ACL) method
  collaborative learning, 124, 125, 342, 348
  and SSCI, 307–12
  strategies, 209–10
  (see also interactive engagement (IE) method; peer instruction)
Adequate Yearly Progress (AYP), 5
Advanced Placement (AP) program, 7, 20
Alice Curricular Materials Team, 194–95
Allen, Charles, 314–23
Aman, Susan, 27–36
American Chemical Society, 126, 354
American Competitiveness Initiative, 17
analysis skills and e-portfolio tool, 100–109, 350
  (see also problem solving)
Anderson, Charles, 219–26
Anderson, Steven W., 148–56
Angelo, Thomas, 301
ANN (Artificial Neural Network), 120, 121, 122, 124
AP (Advanced Placement) program, 7, 20
Applegate, Brooks, 247–65
application experiment type, 327
Armstrong, Robert, 37–44
Arons, Arnold, 61, 66, 181
Artificial Neural Network (ANN), 120, 121, 122, 124
ARTIST (Assessment Tools for Improving Statistical Thinking) project, 93–99, 339
ASA (Assessment of Student Achievement in Undergraduate Education), 7, 8, 110
assessment
  college leadership view of, 17, 23–26
  federal concerns, 4–9, 17–26
  future directions in supporting, 333
  NRC's guidelines for, 65
  (see also testing)
assessment-for-learning approach, 39
Assessment of Student Achievement in Undergraduate Education (ASA), 7, 110
Assessment Practices in Undergraduate Mathematics (Marion and Gold), 167
Assessment Tools for Improving Statistical Thinking (ARTIST) project, 93–99, 339
Assessment Wizard, 212
Association of American Colleges and Universities (AAC&U), 22
astronomy, 141–46
Atkinson, J. W., 314
at-risk students, failure of interventions for, 37–38
attainment, higher education, failure to graduate problem, 18, 41
attitudes, student
  assessment instruments for measuring, 194–200, 353
  and CHEMX survey, 163
  expectations about STEM learning, 159–64, 350–51
  and POSITT, 256
  value decisions on math and science, 314–23, 345–46
attrition rates, 37, 41
Ausubel, D., 160
AYP (Adequate Yearly Progress), 5
BCI (Biology Concept Inventory), 130–39, 349
behaviorally anchored rating scale (BARS), 203–4, 205
Behrens, Nathan, 194–200
Bergman, Kostia, 27–36
Berry, Frederick C., 190–93
BESTEAMS (Building Engineering Student Teams and Management Systems), 240
best practices, importance of sharing, 25–26
Beyerlein, S. W., 53–57
biology/biochemistry
  critical thinking in, 336
  diagnostic question clusters, 347
  e-portfolio initiatives, 27–36, 346–47
  faculty collaborative groups method, 209–17
  identifying and correcting problematic thinking, 220–21, 222–26
  model-based reasoning, 357
  self-assessment tools, 101–2
Biology Concept Inventory (BCI), 130–39, 349
Bjorklund, Stefani, 81–91, 82
Blackboard, 213
Blank, Seymour, 37–44
Blinder, Alan, 20–21
Bloom's taxonomy, 210, 252, 267–68, 292
Boehner, John, 17
Bransford, J. D., 46, 352
Bretz, Stacey Lowery, 159–64
bridging of ideas, 175
Brookes, David T., 68–77
Brown, Robert, 202
Bruner, J., 181
Buck, John R., 307–12
Building Engineering Student Teams and Management Systems (BESTEAMS), 240

Calculus Concept Inventory (CCI), 60–66, 352–53
Calibrated Peer Review (CPR), 190–93, 266–73, 295, 345, 351
Callen, Bruce, 314–23
Campbell, D. T., 252
CAOS (Comprehensive Assessment of Outcomes in a First Statistics) course, 93, 94, 95–97
capstone engineering design, transferable assessments for, 53–57, 348
Carlson, Patricia A., 190–93
CASEE (Center for the Advancement of Scholarship on Engineering Education), 82, 83–84, 353
Casem, Merri Lynn, 209–17
CAT (Critical Thinking Assessment Test), 290–97, 351–52
CATME (Comprehensive Assessment of Team-Member Effectiveness), 202–7
CCI (Calculus Concept Inventory), 60–66, 352–53
Center for the Advancement of Scholarship on Engineering Education (CASEE), 82, 83–84, 353
Chance, Beth, 93–99
chemistry, 103, 118–27, 159–64, 350–51, 354
ChemQuery, 332
CHEMX, 161–64, 350–51
Chen, John C., 46–52
CLA (Collegiate Learning Assessment), 24
Claesgens, Jennifer, 227–37
ClassAction, 141–46, 349–50
CMAP (IHMC Concept Map Software), 278–81
Cobern, William, 247–65
Cognition-Observation-Interpretation assessment triangle, 150, 222

cognitive expectations, 160–61, 350–51
Cognitive Laboratories, 62
coherence in education, 20
collaborative learning, 124, 125, 342, 348
  (see also interactive engagement (IE) method)
Collegiate Learning Assessment (CLA), 24
Commission on the Future of Higher Education, 17–26
Committee on the Undergraduate Program in Mathematics (CUPM), 166–67
competitiveness, labor, 7, 17, 20
Comprehensive Assessment of Outcomes in a First Statistics (CAOS) course, 93, 94, 95–97
Comprehensive Assessment of Team-Member Effectiveness (CATME), 202–7
computer science, 194–200, 346–47, 353
Computing Sciences Accreditation Board (CSAB), 274
Concept-Connector tools (C-Tools), 339–40
concept inventories
  BCI, 130–39, 349
  CCI, 60–66, 352–53
  vs. diagnostic question clusters, 219
  FCI, 60, 61, 130–31, 141, 148
  GCI, 148–56
  overview, 60–61
  for peer instruction, 344–45
  SCI, 338
  SSCI, 307–12, 354–55
  vs. testing, 131–32
  TTCI, 300–305
concept maps
  C-Tools, 339–40
  engineering design process, 274, 278–81, 284
  RWP, 111–12, 113–14, 116
conceptual understanding
  ChemQuery's support for, 228–37
  ClassAction tool, 145–46, 349–50
  computer science, 194–200, 353
  earth sciences processes, 148–56, 335–36
  experiment design tasks, 328
  identifying and correcting problematic thinking, 219–26
  large-enrollment classes, 173–83
  RWP, 110–16
  and student expectations about STEM learning, 160
  (see also concept inventories; misconceptions, student)
constructed response testing instrument, 252
content analysis technique, 133–34
continuous improvement tools for engineering design, 274–85, 334
Cooper, Melanie, 118–27
Cooper, Stephen, 194–200
Cooperative Education Program (Co-op), 29–36, 346–47
costs of higher education, 19
coursework records, 18
CPR (Calibrated Peer Review), 190–93, 266–73, 295, 345, 351
criterion-referenced measurement, ChemQuery as, 227, 228
critical thinking
biology courses, 336
CPR, 351
faculty collaborative groups to assess, 209–17
importance of, 12–13, 290–91, 351–52
Critical Thinking Assessment Test (CAT), 290–97, 351–52
Critical Thinking Longitudinal Embedded Assessment (CTLEA), 210–13
critical thinking/problem-solving (CT/PS) curriculum, 209–17
crossover experimental design, 48, 50
CSAB (Computing Sciences Accreditation Board), 274
CTLEA (Critical Thinking Longitudinal Embedded Assessment), 210–13
C-Tools (Concept-Connector tools), 339–40
CT/PS (critical thinking/problem-solving) curriculum, 209–17
culture, academic department, and assessment development, 171–72
culture and society
  and engineering instruction, 81
  mistake of focusing blame on, 18
  value of liberal education for, 21
CUPM (Committee on the Undergraduate Program in Mathematics), 166–67
curricula
  Alice, 194–95
  capstone engineering design, 53–57, 348
  and conceptual understanding focus, 181
  CT/PS, 209–17
  e-portfolio initiatives, 27–36, 346–47
  EPS, 110, 111
  faculty collaborative assessment, 209–10
  faculty's lack of attention to, 142
  high school vs. college, 20
  innovation in science, 23
  inquiry science teaching, 248
  and mathematics assessment, 167
  media computation method, 195–96
A Curriculum Guide for Engineering Faculty, 240

Dann, Wanda, 194–200
Davis, D. C., 53–57
DDP (diagnostic digital portfolio), 100–109, 350
declarative vs. procedural knowledge, 118–19, 284, 334
Deeds, Donald G., 314–23
deep vs. surface approaches to learning, 111–12, 113
delMas, Robert, 93–99
Delphi method, 210, 301–2
democracy, educating citizens for responsibilities in, 21–22
design
  assessment, 250, 252–55
  engineering, 53–57, 274–85, 334, 348, 351
  physics, 326–30
Design Process Knowledge Test (DPKT), 274, 277–78, 284
diagnostic digital portfolio (DDP), 100–109, 350
diagnostic question clusters, 219, 347
difficulty indexes, 309–10
disadvantages students in two-year technical colleges, 37
discovery method, 175, 181, 210, 247, 248
diversity issue and CPR, 345
Dollhopf, Sherry, 100–109
domains of subject area, 160, 318, 319–20, 322, 323
Doxas, Isidoros, 130–39
DPKT (Design Process Knowledge Test), 274, 277–78, 284
Drewery, Malcolm, 81–91
Driskill, L., 192
dynamic assessment, CAT as, 291

earth sciences, 148–56, 335–36, 347
Eccles, J. S., 314
Ed's Tools, 130–39
education majors, 14, 247–65
efficiency in higher education institutions, 19, 20, 21
E-FSSE (Engineering Faculty Survey of Student Engagement), 81–91, 353
electromechanical engineering technology, 37–44, 355–56
electronic portfolios, 27–36, 100–109, 346–47, 350
elementary teachers, inquiry science approach, 247–65

elicit-confront-resolve method, 175–76, 177, 178–79
employers and assessment development, 27, 29–30, 33–34, 35, 346
engagement, student, 24, 81–91, 141, 142–46, 353
(see also interactive engagement (IE) method)
engineering
  ACL method, 307–12
  CATME tool, 202–7
  CPR, 190–93, 351
  design process, 53–57, 274–85, 334, 348, 351
  electromechanical engineering technology, 37–44, 355–56
  project teams and learning, 240–45, 342–43
  rapid assessment, 46–52, 344–45
  SCI, 338
  student engagement studies, 81–91, 353
  thermal and transport sciences, 335
  TTCI, 300–305
Engineering Faculty Survey of Student Engagement (E-FSSE), 81–91, 353
Engineering National Survey of Student Engagement (E-NSSE), 81–91, 353
The Engineer of 2020: Visions of Engineering in the New Century, 81
Enhanced Peer Review (EPR), 295–96
E-NSSE (Engineering National Survey of Student Engagement), 81–91, 353
environmental problem-solving (EPS) curriculum, 110, 111
Epstein, Jerome, 60–66
ethnicity variable, 244, 345
Etkina, Eugenia, 68–77
Ewell, Peter, 168
expectations, student (see attitudes, student)
experiments, designing and conducting, 70–77, 326–30

faculty
  ARTIST tool usage, 94, 99
  CAT development, 291, 292–93
  CATME, 204–5
  CPR, 266, 267, 268, 270–71, 345
  dissemination of assessment methods to, 167–72
  engineering student engagement study, 85, 86, 87
  e-portfolio tools, 100, 102, 103–4
  focus on student learning, 14
  IE vs. TL methods, 61, 64
  lack of attention to curriculum, 142
  lack of coordination with psychometrics, 63
  project team dynamics, 244, 245
  rapid assessment tools, 46, 50
  real-time multi-dimensional assessment, 333
  RWP, 110–17
  salaries and costs to colleges, 19
  scientific abilities assessment role, 76–77
  on self-assessment by students, 101
  Spellings Commission's lack of consideration for, 19
  SSCI development, 311
  students' motivations, 159–60, 162
  (see also teachers)
faculty collaborative groups, 209–17, 336
FCI (Force Concept Inventory), 60, 61, 130–31, 141, 148
federal government
  assessment policies, 4–9, 17–26
  support for private vs. public higher education, 17
feedback systems
  ClassAction, 141–46, 349–50
  e-portfolios, 27–36, 100–109, 346–47, 350
  flashcard method, 48, 50
  rapid assessment, 46, 344–45
  scientific abilities assessment, 69
Felder, R. M., 202
Fiske, D. W., 252
flashcard feedback method, 48, 50
Force Concept Inventory (FCI), 60, 61, 130–31, 141, 148
formative assessment methods
ClassAction, 141–46, 349–50
in-class, 173–83, 343
scientific abilities, 69–77, 348
FormSite, 86
Fortenberry, Norman, 81–91
Fortier, Paul, 274–85
foundation engineering courses, 46–52
Fraknoi, A., 141
Freeman, M., 46
Friedman, Francis, 180–81
Fuller, H., 202
fully interactive lecture method, 178–79
functional role specialization, 240–45
fundraising, higher education, 19

Garfield, Joan, 93–99
Garvon-Doxas, Kathy, 130–39
GCI (Geoscience Concept Inventory), 148–56
Geist, Monica R., 300–305
gender factor, 162, 244, 317
Geoscience Concept Inventory (GCI), 148–56
geosciences, 148–56, 335–36, 347
globalization and education policy, 7, 17, 20, 81
Gold, Bonnie, 167
grading, traditional, ineffectiveness as motivational tool, 38–39
grounded theory, 150
Grove, Nathaniel, 159–64
guided inquiry, 175, 181, 210, 247–65, 336, 356–57
Guidry, Michael, 141–46
Guilbault, Lauralee, 100–109
Gunersel, Adalet Baris, 266–73
Guzdial, Mark, 194–200

Hake, R. R., 60, 61, 130, 141, 176, 309
Halloun, I., 61
Harrison, Olakunle, 53–57
Haynes, Ada, 290–97
Heidemann, Merle, 219–26
Hestenes, D., 61
Hewitt, Nancy, 159
Hidden Markov Modeling (HMM), 120, 121, 124
higher education
  accountability for science education, 4, 16–26
  ASA initiative, 7, 8, 110
  failure to graduate problem, 18
  and labor competitiveness, 7
"highly qualified" teacher status, 6
Himangshu, Sumitra, 110–17
Hjalmarson, Margret A., 307–12
HMM (Hidden Markov Modeling), 120, 121, 124
Hoese, Bill, 209–17
Holme, Thomas, 118–27
Houtman, Anne, 209–17
Hudesman, John, 37–44
hypothetical-deductive learning model, 119

IB (International Baccalaureate) program, 7
IE (interactive engagement) method (see interactive engagement (IE) method)
IHMC Concept Map Software (CMAP), 278–81
IMMEX (Interactive MultiMedia Exercises) system, 118–27, 354
in-course assessment tools, 173–83, 340–41, 343
information age, training for employment in, 20
innovation skills, 13
Inquiry-Item-Criteria approach to assessment design, 250, 252–55
inquiry science teaching, 247–65, 336, 356–57
in-service teachers, inquiry science teaching for, 256, 257
intellectual capital, 18
interactive engagement (IE) method
  application to physics students, 173–78
  and concept inventories, 60, 61–62, 64–65, 130
  effectiveness of, 141
  historical perspective, 180–83
interactive lecture demonstration technique, 177
Interactive MultiMedia Exercises (IMMEX) system, 118–27, 354
International Baccalaureate (IB) program, 7
international education, emphasis on science and engineering, 20
international students, 18
investigation experiment type, 328
iPeer system, 202
Item Response Theory (IRT)
  ChemQuery, 227–28, 234–35
  collaborative learning effects, 125
  GCI, 154
  IMMEX, 120, 121
  inquiry science teaching, 253–55
Iuli, Richard J., 110–17

Johnstone, A. H., 160
just-in-time teaching, 142, 177

K-12 education perspective, 5, 18, 213, 247–65
Kadlowec, Jennifer A., 46–52
Kandel, Judith, 209–17
Karelina, Anna, 68–77
Karplus, Robert, 181, 248
Kaufman, D. B., 202
Keith, Sandra, 167
Klymkowsky, Michael W., 130–39
Knowing What Students Know, 65, 220
knowledge
  students' prior, 160, 300–301, 335–36
  transfer of, 340–41
  understanding vs. memorization, 46–47, 160, 176, 181
  (see also conceptual understanding)
Kraus, Pam, 176

Laboratory Program Variables Inventory, 328
Labov, Jay B., 3–9
large-enrollment lecture classes, 173–83, 204, 209–10, 343
Lawrenz, F., 141
Laws, Priscilla, 62
Layne, Jean, 266–73
Layton, Richard, 204
learning, student
  approaches to, 111–12, 113, 116
  ARTIST project, 93–99, 339
  definition, 160
  and disillusionment with STEM courses, 159–64
  dissemination of programs to faculty, 167–72
  engineering student engagement study, 82–84, 85, 353
  inquiry science teaching, 247–65, 356–57
  limitations of federal government's initiative, 24
  Project Kaleidoscope assessment methods, 13–15
  rapid assessment methods, 46–52, 141–46, 176, 344–45, 349–50
  scientific abilities assessment, 68–77, 348
  SRL assessment approach effects, 41–42, 43–44
  team projects, 240–45
  (see also conceptual understanding; critical thinking; peer instruction; problem solving; self-assessment, student)
Learning and Studying Questionnaire (LSQ), 111
learning cycle, 181, 248, 350
Lee, Kevin M., 141–46
Leonard, Jeannie Brown, 240–45
Libarkin, Julie C., 148–56
liberal arts colleges and efficiency in higher education, 19
liberal education, value of, 21–23, 25
Likert-scale vs. BARS rating instruments, 204
literacy, math and science, 18, 24, 93–99, 314–23, 345–46
LON CAPA network, 347
Long, Tammy, 219–26
Looking at Student Work, 213
Loughry, Misty L., 202–7
LSQ (Learning and Studying Questionnaire), 111
Luttrell, Vickie, 314–23

MAA (Mathematical Association of America), 166–67
Madison, Bernard, 166–72
majors in subject area
  decision process for leaving STEM majors, 159–64
  education majors, 14, 247–65
  and valuation of science and math education, 318, 345–46
Marion, William, 167
Maryland Physics EXpectations (MPEX) survey, 160–61, 350
Mathematical Association of America (MAA), 166–67
mathematics
  CCI, 60–66, 352–53
  dissemination of learning programs, 166–72
  self-assessment tools, 102–3
  statistics, 47, 93–99, 338, 339
  student value decisions on, 314–23, 345–46
Math Value Inventory (MVI), 314–23
matter, change, and energy, ChemQuery tool, 228–37
matter and energy, problematic thinking challenges in, 222–26
Mazur, Eric, 62, 142, 177
MBR (model-based reasoning), 357
McClanahan, Elaine, 209–17
McDermott, Lillian, 181–82
McKenzie, J., 46
media computation method, 195–96
Mehta, S. I., 48
Meltzer, David E., 173–83
memorization vs. understanding, 46–47, 160, 176, 181
  (see also conceptual understanding)
Merrill, John, 219–26
Merritt, Brett, 219–26
Millard, Don Lewis, 190–93
Miller, Ronald L., 82, 300–305
misconceptions, student
  and BCI, 349
  definitional issues, 308

rapid-feedback systems to dispel, 141–46, 349–50
SSCI measurement of, 307–12, 354–55
thermal and transport sciences, 335
TTCI measurement of, 300–305
model-based reasoning (MBR), 357
Modell, H., 308
Morton, Ed, 37–44
Moskal, Barbara M., 82, 194–200
Moylan, Adam, 37–44
MPEX (Maryland Physics EXpectations) survey, 160–61, 350
multiple-choice testing instrument, 131–32, 252
multi-trait, multi-method (MTMM) approach, 252
Murthy, Sahana, 68–77
MVI (Math Value Inventory), 314–23

NAICU (National Association of Independent Colleges and Universities), 18
Narum, Jeanne, 12–15
Nasr, R., 308
National Assessment of Adult Literacy, 18, 24
National Education Standards, 6–7
National Research Council, 5, 6
National Survey of Student Engagement (NSSE), 24, 81–91, 294, 353
national tests, Spellings Commission focus on, 24
NCLB (No Child Left Behind) Act, 5
need-based financial aid, 19
Nelson, Mary A., 300–305
neural networks (ANN), 120, 121, 122, 124
No Child Left Behind (NCLB) Act, 5
normalized gain, definition, 61
Novak, G. M., 142, 177
Novak, J. D., 160
Noyes, Sara, 27–36
NSSE (National Survey of Student Engagement), 24, 81–91, 294, 353

observational experiment type, 326–27
Ohland, Matthew W., 202–7
Olds, Barbara M., 82, 300–305
online tools
availability during course progress, 340–41
CATME, 204–5
CPR, 190–93, 266–73, 295, 345, 351
Ed's Tools, 130–39
e-portfolios, 27–36, 100–109, 346–47, 350
IMMEX, 118–27, 354
real-time multi-dimensional assessment, 332–33
for sharing teaching solutions, 26
statistics literacy testing materials, 94–98
Web ARTIST, 339
Ono, Joyce K., 209–17
Ooms, Ann, 93–99, 97
Open Source Portfolio (OSP), 27–30
opportunistic engineering design, 276–77
outsourcing of jobs and higher education's teaching focus, 20–21
Oxtoby, David, 16–26

Parker, Joyce, 219–26
PBL (Problem-Based Learning), 251–52
PDAs (personal digital assistants) as rapid assessment tools, 46, 47, 49, 50
pedagogical content knowledge (PCK), 248, 255–56, 356–57
Pedagogy of Science Inquiry Teaching Test (POSITT), 247–65, 356–57
peer instruction
concept inventories for, 344–45
CPR, 190–93, 266–73, 295, 345, 351
diagnostic question clusters, 347
engineering design process, 276
EPR, 295–96
future research on, 342
large-enrollment classes, 174, 176, 177
structure of, 142–46
team-member effectiveness, 202–7, 344
Peer Instruction (Mazur), 62
PER (physics education research) community, 68
personal capacity assessment, 55–56, 58
personal digital assistants (PDAs) as rapid assessment tools, 46, 47, 49, 50
Perspectives of Chemists, 227–37
Physical Science Study Committee project, 180–81
physics
astronomy assessment, 141–46
experiment design, 326–30
FCI, 60, 61, 130–31, 141, 148
in-class formative assessments for, 173–83, 343
scientific abilities assessment, 68–77, 348
student expectations analysis, 160–61, 350
physics education research (PER) community, 68
Piaget, J., 181
Poklop, Laurie, 27–36
policy issues (see federal government)
politics and support for private vs. public higher education, 17
Porter, Veronica L., 27–36
portfolios, electronic, 27–36, 100–109, 346–47, 350
POSITT (Pedagogy of Science Inquiry Teaching Test), 247–65, 356–57
pre-service teachers, inquiry science teaching for, 247–65
prior knowledge, students', 160, 300–301, 335–36
privacy issue and unit record system proposal, 18–19
Problem-Based Learning (PBL), 251–52
problem solving
e-portfolio tool, 100–109, 350
faculty collaborative groups to assess, 209–17
IMMEX system, 118–27, 354
and interactive engagement, 175–76
RWP, 110–17
scientific abilities assessment, 68–77, 348
(see also critical thinking; design)
procedural vs. declarative knowledge, 118–19, 284, 334
process-based assessments
earth sciences, 148–56, 335–36
engineering design, 53–57, 274–85, 334, 348, 351
scientific abilities, 68–77, 348
Professional Developer, 202
professional development
for engineering students, 53
faculty, 94, 166–72
RWPs, 337
standards for teachers', 6, 7
Project Kaleidoscope, 12–15, 23, 274
Project Leap (Liberal Education and America's Promise), 22
project teams, student learning in, 240–45, 342–43
proprietary vs. public schools, funding and support for, 17
psychometrics, definition, 63
public vs. proprietary schools, funding and support for, 17
Pustejovsky, Susan, 100–109

quality vs. efficiency of education, 20

race and ethnicity variable, 244, 345
ranked response testing instrument, 252
rapid-assessment methods, 46–52, 141–46, 176, 344–45, 349–50
Rasch (IRT) analysis, 148, 154, 155
ratings of colleges, 17
real-time multi-dimensional assessment, 332–33
Redding, Michael, 290–97
Redish, E. F., 160, 176
Regional Workshops Project (RWP), 110–17, 337
research-based instruction, 176–79, 182
reviewer competency index (RPI), 268, 269
Richmond, Gail, 219–26
Rising Above the Gathering Storm, 7, 16
Rokeach, M., 314
Rotter, J. B., 314
RPI (reviewer competency index), 268, 269
Rue, Loyal, 13
Russell, Arlene, 266–73
RWP (Regional Workshops Project), 110–17, 337

SALG (Student Assessment and Learning Gains), 269
SAMPI (Science and Mathematics Improvement), 257
SaM-VI (Science and Math Values Inventory), 345–46
SATS (Survey of Attitudes Toward Statistics), 338
Saul, J. M., 176
SAUM (Supporting Assessment in Undergraduate Mathematics), 166–72
Scalise, Kathleen, 227–37
Scherr, R. E., 308
Schmidt, Edward G., 141–46
Schmidt, Janet A., 240–45
Schmidt, Linda C., 240–45
Schuster, David, 247–65
Schwartz, Renée, 247–65
Science and Mathematics Improvement (SAMPI), 257
Science and Math Values Inventory (SaM-VI), 345–46
Science Perceptions Survey, 332
Science Value Inventory (SVI), 314–23
scientific abilities assessment, 68–77, 348
SCI (Statistical Concepts Inventory), 338
self-assessment, student
  CATME, 203, 204
  CPR, 190
  electromechanical engineering technology, 37–44, 355–56
  e-portfolios as, 35, 100–109, 346
  real-time multi-dimensional tool for, 332–33
  scientific abilities assessment, 69, 71–72
  team effectiveness assessment, 358
Self-regulated Learning (SRL) assessment system, 37–44, 355–56
service economy, educating for, 21
Seymour, Elaine, 159
Sibley, Duncan, 219–26
Sibley, James, 202
Signals and Systems Concept Inventory (SSCI), 307–12, 354–55
Simpson, Nancy, 266–73
Sims-Knight, Judith E., 274–85
simulation task, design process, 274, 281–83, 284
Slater, Timothy F., 141–46
Smith, Paige E., 240–45
society and culture
  and engineering instruction, 81
  mistake of focusing blame on, 18
  value of liberal education for, 21
socioeconomic status and higher education opportunity, 18, 22, 37
Sokoloff, D. R., 177
solution requirements assessment, 55, 56–58
Sonnentag, S., 276
specialization in higher education, 20, 21–22, 244–45
Spellings, Margaret, 4, 17
Spellings Commission, 17–26, 24
SRL (Self-regulated Learning) assessment system, 37–44, 355–56
SSCI (Signals and Systems Concept Inventory), 307–12, 354–55
Stacy, Angelica, 227–37
standardized testing, 6–7, 24
state governments, education testing responsibilities, 5–6
Statistical Concepts Inventory (SCI), 338
statistics, 47, 93–99, 338, 339
Statistics Teaching Inventory (STI), 99
Steen, Lynn, 168
Stein, Barry, 290–97
Steinberg, R. N., 176
STEM (science, technology, engineering or mathematics) community
  and education policy, 3–9
  leadership role in education improvement, 22–23
  learning insights, 12–15
  role in assessment improvement, 25
  (see also learning, student)
Stevens, Ron, 118–27
STI (Statistics Teaching Inventory), 99
Streveler, Ruth A., 300–305
Student Assessment and Learning Gains (SALG), 269
student perspective
  on CAT, 292
  on CPR, 269–70
  decision process for leaving STEM majors, 159–64
  as evaluators of assessment tools, 31
  on Inquiry-Item-Criteria approach, 253–54
  on rapid feedback value, 50
  valuation of STEM education, 314–23, 345–46
  value of e-portfolio for, 35
  views of teamwork roles, 241–45
  (see also attitudes, student; learning, student)
Supporting Assessment in Undergraduate Mathematics (SAUM), 166–72
surface vs. deep approaches to learning, 111–12, 113
Survey of Attitudes Toward Statistics (SATS), 338
SVI (Science Value Inventory), 314–23

Talking About Leaving: Why Undergraduates Leave the Sciences (Seymour and Hewitt), 159
task work for teams, 240
Tavris, Carol, 12
teachers
  inquiry science approach for elementary school, 247–65
  introductory science students as future, 141
  need for advanced program, 7
  need for quality, 9, 19
  professional development standards, 6, 7
  on SRL assessment approach effects, 42–43
  (see also faculty)
Team-Maker, 204, 207
teams
  dynamics of, 240
  effectiveness measurement, 202–7, 344, 358
  processes assessment in, 55, 56, 58
  student learning in, 240–45, 342–43

technical colleges, student lack of preparation for, 37
technology in classroom and costs to colleges, 19
testing
  vs. concept inventories, 131–32
  constructed response, 252
  critical thinking, 290–97, 351–52
  DPKT, 274, 277–78, 284
  multiple-choice instrument, 131–32, 252
  under NCLB, 5
  POSITT, 247–65, 356–57
  ranked response instrument, 252
  standardized, 6–7, 24
  state government responsibilities, 5–6
  (see also assessment)
testing experiment type, 327
A Test of Leadership: Charting the Future of U.S. Higher Education, 17
Tew, Allison Elliott, 194–200
Thermal and Transport Concept Inventory (TTCI), 300–305
thermal and transport sciences, 300–305, 335
They're Not Dumb, They're Different (Tobias), 159–60
thinking level and problem solving in groups, 125–26
think-pair-share technique, 142
  (see also peer instruction)
Thompson, Phillip, 53–57
Thompson, Tracy, 100–109
Thorton, R. K., 177
TIDEE (Transferable Integrated Design Engineering Education) consortium, 348
Tobias, S., 159–60
top-down engineering design, 276
traditional lecture (TL) method
  ineffectiveness of, 60, 61–62, 64–65, 130, 141
  large-enrollment classes, 173–83, 204, 209–10, 343
Transferable Integrated Design Engineering Education (TIDEE) consortium, 348
Transfer Environment and Student Readiness Instrument, 341
transfer of knowledge and skills, 340–41
Treisman, Uri, 64
Trevisan, Michael, 53–57
Truchan, Leona, 100–109
TTCI (Thermal and Transport Concept Inventory), 300–305
two-year colleges, student lack of preparation for, 37

understanding vs. memorization, 46–47, 160, 176, 181
  (see also conceptual understanding)
Undreiu, Adriana, 247–65
unit record system, 18
Upchurch, Richard, 274–85
Urban-Lurain, Mark, 219–26

values and student decisions on math and science, 314–23, 345–46
Van Heuvelen, Alan, 68–77, 177
Vellom, Paul, 247–65
Viennot, L., 181–82
Villasenhor, Maria Ruibal, 68–77
vocational vs. liberal education, 22

Wade, Carole, 12
Wage, Kathleen, 307–12
weaving of ideas, 175
web-based assessment materials (see online tools)
Weiss, Iris, 248
White, Nelsha, 37–44
Whittinghill, Dexter C., 46–52
Wilson, Christopher, 219–26
Wilson, Mark, 227–37
Wood, Mark, 314–23
Woolfson, Judith F., 27–36
Workshop Physics, 62
writing skills
  and CPR, 190–93, 266–73, 295, 345, 351
  importance of, 35

Young, Chris, 100–109
Young, Todd S., 141–46

Zacharias, Jerrold, 180–81
Zimmerman, Barry, 37–44
Zou, Xueli, 326–30