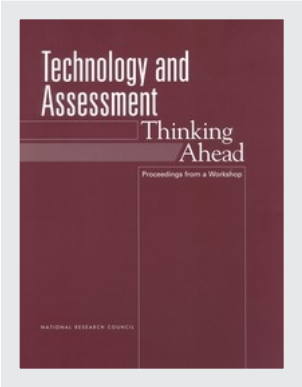


Technology and Assessment: Thinking Ahead -- Proceedings from a Workshop

DETAILS

104 pages | 8.5 x 11 | PAPERBACK
ISBN 978-0-309-08320-1 | DOI 10.17226/10297

AUTHORS

Board on Testing and Assessment, National Research Council

Technology and Assessment: Thinking Ahead

Proceedings from a Workshop

Board on Testing and Assessment
Center for Education
Division of Behavioral and Social Sciences and Education

National Research Council
Washington, DC 2002

NATIONAL ACADEMY PRESS 2101 Constitution Avenue, N.W. Washington, DC 20418

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Contract/Grant No.# 2001-6884 between the National Academy of Sciences and the Hewlett Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

Library of Congress Cataloging-in-Publication Data
or
International Standard Book Number 0-309-08320-6

Suggested citation: National Research Council. (2002). *Technology and assessment: Thinking ahead: Proceedings of a workshop*. Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Additional copies of this report are available from National Academy Press, 2101 Constitution Avenue, N.W., Lockbox 285, Washington, D.C. 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu

Printed in the United States of America

THE NATIONAL ACADEMIES

National Academy of Sciences
National Academy of Engineering
Institute of Medicine
National Research Council

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Wm. A. Wulf is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Kenneth I. Shine is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. Wm. A. Wulf are chairman and vice chairman, respectively, of the National Research Council.

Preface

The papers in this collection were commissioned by the Board on Testing and Assessment (BOTA) of the National Research Council (NRC) for a workshop held on November 14, 2001, with support from the William and Flora Hewlett Foundation. Goals for the workshop were twofold. One was to share the major messages of the recently released NRC committee report, *Knowing What Students Know: The Science and Design of Educational Assessment* (2001), which synthesizes advances in the cognitive sciences and methods of measurement, and considers their implications for improving educational assessment. The second goal was to delve more deeply into one of the major themes of that report—the role that technology could play in bringing those advances together, which is the focus of these papers. For the workshop, selected researchers working in the intersection of technology and assessment were asked to write about some of the challenges and opportunities for more fully capitalizing on the power of information technologies to improve assessment, to illustrate those issues with examples from their own research, and to identify priorities for research and development in this area.

BACKGROUND

In recent years, BOTA has explored pressing and complex issues in educational assessment, including the role and the appropriate uses of assessment in standards-based reform; how well current assessments are fulfilling the various demands placed on them; and concerns about fairness and equity in testing. In 1998, BOTA decided the time was right to address a long-standing issue noted by numerous researchers interested in problems of educational assessment: the need to bring together scientific understanding of how people learn with methods for assessing what they have learned. An NRC committee was formed to review advances in the cognitive and measurement sciences, as well as initial, promising work done in the intersection between the two disciplines, and to consider the implications for reshaping educational assessment. The National Science Foundation (NSF) recognized the importance and timeliness of such a study and agreed to sponsor the effort.

The resulting committee report, *Knowing What Students Know: The Science and Design of Educational Assessment*, was released in 2001. The underlying premise of the report is that new forms of classroom and large-scale assessments are needed that help all students learn and succeed in school, by making as clear as possible to them, their teachers, and other education stakeholders the nature of their accomplishments and the progress of their learning. Advances in the cognitive sciences have broadened the conception of those aspects of learning that are most important to assess, and advances in measurement have expanded the capability to interpret more

complex forms of evidence derived from student performance. A merger of these two sets of advances could lead to a significant leap forward in the science and practice of assessment, and technology is playing an important role in making such a merger feasible.

INFORMATION TECHNOLOGIES: OPPORTUNITIES FOR ADVANCING EDUCATIONAL ASSESSMENT

Some of the main conclusions from *Knowing What Students Know* about the nature of assessment and the role of technology are summarized below. They provide a set of framing ideas for the papers that follow.

Assessments are used in both classroom and large-scale contexts for three broad purposes: to assist learning, to measure individual achievement, and to evaluate programs. One type of assessment does not fit all purposes, and assessments used in various contexts often look quite different. But every assessment, regardless of its purpose, rests on three pillars: (1) a model of how students represent knowledge and develop competence in the subject domain; (2) tasks or situations that allow one to observe students' performance; and (3) an interpretation method for drawing inferences from the performance evidence thus obtained.

These three elements—cognition, observation, and interpretation—must be explicitly connected and designed as a coordinated whole. If the three elements are not in synchrony, the meaningfulness of inferences drawn from the assessment will be compromised. The three elements and the connections among them are referred to here as the *assessment triangle*. Improved methods of assessment require a design process that connects the three elements of the assessment triangle to ensure that the theory of cognition, the observations, and the interpretation process work together to support the intended inferences. Fortunately, there are multiple examples of technology tools and applications that enhance specific linkages among cognition, observation, and interpretation, as well as more general linkages among curriculum, instruction, and assessment. A few of these enhancements are mentioned here as a frame of reference for the examples described in the workshop papers.

Among the most intriguing applications of technology are those that extend the nature of the problems that can be presented and the knowledge and cognitive processes that can be assessed. When task environments are enriched through the use of multimedia, interactivity, and control over the stimulus display, it is possible to assess a much wider array of cognitive competencies than has heretofore been feasible. New capabilities enabled by technology include directly assessing problem-solving skills, making visible sequences of actions taken by learners to solve problems, and modeling and simulating complex reasoning tasks. Technology also makes possible the collection of data on concept organization and other aspects of students' knowledge structures, as well as representations of their participation in discussions and group projects.

A significant contribution of technology has been to the design of systems for implementing sophisticated classroom-based formative assessment practices. Technology-based systems have been developed to support individualized instruction by extracting key features of learners' responses, analyzing patterns of correct and incorrect reasoning, and providing rapid

and informative feedback to both student and teacher. Often such approaches are embedded in complex teaching-learning environments supported by technology, as noted below. *A major change in education has resulted from the influence of technology on what is taught and how it is taught. Schools are placing more emphasis on teaching critical content in greater depth.* Examples include the teaching of advanced thinking and reasoning skills within a discipline through the use of technology-mediated projects that involve long-term inquiry. Such projects often integrate content and learning across disciplines, as well as integrating assessment with curriculum and instruction in powerful ways.

A possibility for the future arises from the projected growth across curricular areas of technology-based assessment embedded in instructional settings. Increased availability of such systems could make it possible to pursue balanced designs representing a more coordinated and coherent assessment system. Information from such assessments could possibly be used for multiple purposes, including the audit function associated with many existing external assessments.

Finally, while technology holds great promise for enhancing educational assessment at multiple levels of practice, its use for this purpose also raises issues of utility, practicality, cost, equity, and privacy. These issues will need to be addressed as technology applications in education and assessment continue to expand, evolve, and converge.

PAPERS IN THIS VOLUME

The papers that follow address some of the opportunities and challenges just noted and provide specific examples of the linkages mentioned above.

The first two papers focus on the role of technology in improving assessments used for summative or external evaluation purposes. In the first paper, Gitomer and Bennett illustrate how computer technologies are being used by researchers at the Educational Testing Service to address a long-standing criticism of standardized tests: that they tend to consist of certain types of traditional test items and have lost sight of the underlying constructs, or cognitive competencies, that are the targets of assessment. The authors provide several examples of the use of computer technologies to "unmask" the constructs underlying traditional assessments, such as the PSAT, and to make the constructs more visible and explicit in the design of new assessments. In the second paper, Means and Haertel describe an effort to develop computer-based, quality assessments of scientific inquiry. Despite the visibility of inquiry skills as some of the most highly emphasized skills in science curriculum standards, they are the least likely to be adequately assessed in large-scale accountability systems. In contrast, technology has been widely used over the last decade to develop simulations and complex learning environments that support inquiry processes and the assessment of inquiry skills. Many such environments include embedded assessments that are closely tied to the learning activities. Means and Haertel describe efforts to capitalize on such work and take the next steps toward developing science assessments with broader applicability that can be disentangled from specific instructional contexts and used across various curricula for program evaluation purposes.

The focus of the next three papers shifts to uses of technology to improve classroom assessment that is aimed at monitoring students' understanding and guiding the next steps for instruction. Fletcher's paper reviews evidence of substantial learning gains when instruction is tailored to the needs and capabilities of individual learners, as in one-on-one tutoring situations. While tutoring is the most effective form of instruction, it is also the most expensive. Computer technologies, such as intelligent tutoring systems, are making it feasible to provide some of the advantages of human tutoring on a more widespread basis. Fletcher reviews a variety of examples of computerized instructional programs from the military and other arenas that have assessment components to tailor the pace, content, difficulty, and sequencing of instructional material to the needs of individual learners. In the fourth paper, Williams describes in more depth an example of such a system in the area of early, basic reading skills. Finding time to provide individual feedback during children's reading practice is difficult for teachers who often have 20 or more students in their classes. But recent developments in speech recognition technology are making it possible to increase opportunities for individual reading practice with feedback, as well as to collect assessment information for instructional decision making. In the fifth paper, Corbett provides an example of an intelligent tutoring system in the more complex content domain of algebra. Based on a cognitive model of how students learn algebra, the intelligent tutoring system provides students with rich problem-solving environments and adaptive student support.

In the sixth paper, Russell raises ideas about how technologies might not only make assessment more efficient but also more fundamentally "disrupt" current assessment practices. For instance, testing experts have argued for a long time that a single test should be applied to meet a single purpose, but technology may make it feasible for a single source of data to meet multiple purposes. Russell describes several examples of computer-based, complex learning systems that have the potential to capture rich information about student learning and performance during the actual learning process. He argues that information collected in this way could be used for formative assessment purposes to guide next steps for instruction; it could be accumulated and mined for summative assessment purposes, such as program evaluation, and eventually eliminate or reduce the need for separate on-demand, external exams.

One of the challenges emphasized in several of the papers is that the development of assessments based on advances in cognitive theory, measurement, and technology is a difficult and time-consuming task; it will be a huge undertaking to develop the kinds of assessments discussed in these papers for multiple areas of the curriculum. In the seventh paper, Baker describes the substantial research and development effort that is needed to make test design more systematic and efficient through the use of technology. This final paper lays out a set of goals and components for automated authoring systems for tests and describes some early work in this area.

ACKNOWLEDGMENTS

The Board on Testing and Assessment is grateful to Marshall (Mike) Smith of the William and Flora Hewlett Foundation for sponsoring this workshop. Special thanks also go to the steering group that helped plan and facilitate this activity: Jim Pellegrino, University of Illinois, Chicago; Eva Baker, Center for the Study of Evaluation, University of California, Los

Angeles; George Madaus, Boston College; Lauren Resnick, University of Pittsburgh; and Lorrie Shepard, University of Colorado, Boulder. Thanks also to all of the workshop presenters, who contributed to the success of the workshop: Rich Lehrer, University of Wisconsin, Madison; Mark Wilson, University of California, Berkeley; Bob Glaser, University of Pittsburgh; Larry Suter, National Science Foundation; Jose Mestre, University of Massachusetts; Albert Corbett, Carnegie Mellon University; Drew Gitomer and Paul Holland, Educational Testing Service; Mike Russell, Boston College; Susan Williams, University of Texas, Austin; Barbara Means and Geneva Haertel, SRI International; and Dexter Fletcher, Institute for Defense Analyses. The papers included in this collection were written by a subset of these participants, and we greatly appreciate the authors' additional time and effort.

Several National Research Council (NRC) staff helped plan the workshop and produce this collection of papers, including Naomi Chudowsky, who directed the project, and Andrew Tompkins, who provided excellent administrative help. Stuart Elliott served as a consultant; Pasquale DeVito, Director of BOTA, and Michael Feuer, Director of the Center for Education, provided guidance to the project.

These papers have been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the NRC's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making this volume as sound as possible and to ensure that the publication meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of these papers: Alan Lesgold, University of Pittsburgh; and Jim Pellegrino, University of Illinois, Chicago.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations nor did they see the final draft of the papers before their release. The review of this publication was overseen by Milt Hakel, Bowling Green State University. Appointed by the National Research Council, he was responsible for making certain that an independent examination of the papers was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of the papers rests entirely with the individual authors.

Table of Contents

Chapter 1
Unmasking Constructs Through New Technology, Measurement Theory, and Cognitive Science

Drew H. Gitomer and Randy Elliot Bennett
Educational Testing Service

Knowing What Students Know provides us with a compelling view of the future of educational assessment, a future that includes better information about student learning and performance consistent with our understandings of cognitive domains and of how students learn. That future also promises a much tighter integration of instruction and assessment. Realizing these ambitions depends on progress in the fields of cognition, technology, and assessment, as well as significant changes in educational policy at local and national levels.

The challenges to attaining the vision should not be underestimated. Key examples of cognitive models go back a quarter of a century or more (e.g., Brown & Burton, 1978; Siegler, 1976). Similarly, technology research efforts have demonstrated complex tasks that appear to assess problem solving in particular domains much more authentically than traditional methods (Steinberg & Gitomer, 1996). And our psychometric models are clearly up to characterizing human performance on these more complex tasks (e.g., Almond & Mislevy, 1999). Why, then, are we still very much in the early formative stages of a new generation of educational assessment (Bennett, 1998)?

One of the major obstacles is scale. Representing cognition in large domains remains a mammoth undertaking. We do not yet have the technology to rapidly and cost-effectively map the structure of knowledge for broad cognitive domains like the K-12 curriculum, for example. Designing tasks closely linked to these cognitive-domain structures is still a time-intensive enterprise reserved for a relatively small cadre of experts. The interpretation of evidence does not appear to face the same scaling limitations. If we can adequately scale the cognition and observation legs of the assessment triangle, we believe that the interpretation leg will not provide as great an obstacle.

Even if we can build assessments that scale cost effectively, we are still left with important policy questions. Will there be the political support for more textured assessments, or is there a comfort and familiarity with single summary scores, no matter how oversimplifying they may be? Will there be the willingness to give greater time, and funding, for assessments that provide better information? Time and economic constraints have had a major influence on the kinds of assessments that we currently practice. And will policy makers and educators give adequate attention to more formative assessments as a way of describing both student learning and the conditions affecting that learning? The more revealing an assessment, the more

threatening it can be, for it can uncover issues around opportunities to learn that can be fairly well hidden with our traditional test structures.

In considering these significant challenges, at Educational Testing Service (ETS) we are trying to reconceptualize assessment at a number of levels. We'd like to share with you some of our colleagues' efforts that vary on a host of dimensions; some of these efforts represent incremental improvements in our most traditional assessments, while others involve radically new approaches to assessment consistent with the most ambitious visions of *Knowing What Students Know*. What these efforts have in common, though, is that they have used technology to help unmask the constructs that are the targets of assessment.

What do we mean by the unmasking of constructs and why is this important? Standardized assessments have often been characterized as irrelevant and arcane to the test taker. The recent characterizations of the Scholastic Aptitude Test (SAT) by Richard Atkinson, president of the University of California System, are a striking example. Atkinson argues that the SAT is problematic, in part, because task types such as analogies are puzzle-like, limited in scope, and not directly linked to any California curricular frameworks. Thus, he contends that preparing for such tests distracts students and teachers from focusing on the important learning goals articulated in the state's K-12 content standards. Atkinson also makes the point that access to the secrets of these tests is not equitably distributed in our society.

Such criticisms are not unique, and they point to a historical problem with traditional tests—the masking of constructs, that is, a lack of clarity of the meaning associated with performance. On high stakes tests, such ambiguity causes overwhelming attention to particular task types and to test questions themselves. In attending so nearsightedly to these test components, we lose sight of the constructs underlying the measures and why the original designers thought those components might be useful indicators of important knowledge and skills. For example, while some might argue that verbal analogy items are irrelevant to content standards, most educators, including cognitive scientists, would agree that analogical reasoning is critical to learning and performance in virtually any discipline. Similarly, although reading comprehension items might be criticized for a lack of surrounding context, few would argue that the comprehension of written text is anything but essential.

The kinds of assessments envisioned in *Knowing What Students Know* are clearly designed to unmask the constructs by making the link between learning goals and assessment practices much more explicit. It is worth noting that much of the emphasis in this report is on providing rich, instructionally relevant assessment feedback to students. We would argue that the unmasking must begin far earlier. Students and teachers should have a much clearer sense of what is valued (i.e., the construct) through engagement with tasks more tightly coupled with content standards and instructional activities. The assessment tasks should facilitate, rather than interfere with, an understanding of what is important.

We will briefly discuss three efforts that attempt to further unmask important constructs. Recognizing the dominance of standardized assessments and the important issues that must be addressed before the promise of a new generation of assessments is realized, we begin with two efforts focused on our more traditional tests. In these projects, we investigate how we can help

to make the constructs underlying standardized assessments more transparent to students and teachers, with the goal of altering the focus from the tasks themselves to the constructs they measure. Indeed, the unmasking of constructs was not the primary goal of either of these efforts but the unintended, and fortunate, consequence of attempts to improve traditional assessments. Our third example is a prototype that illustrates the kind of purposefully designed assessment/instruction system that we believe represents the future of educational measurement. All three efforts have been made possible through advances in technology and assessment, as well as attention to the cognitive aspects of performance.

Our first project focuses on the production of greater diagnostic information for a test that was never designed to be diagnostic but to provide a summative judgment of a student's overall academic preparedness for college-level work: the Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT). This project confronted two questions: (1) What skills are necessary for success on the PSAT/NMSQT (and in college)? and (2) How can we communicate these skills, and ways to improve them, to students, teachers, parents, and counselors? To answer the first question, ETS staff conducted cognitive analyses to identify the skills required to solve test items. For the second question, they assembled three panels of math and English teachers who refined the report language, provided suggested activities for skill development, and prioritized the skills.

The essence of the approach was to extract, via psychometric modeling, diagnostic information from the *pattern* of item responses provided by the examinee. Solving each item requires some small subset of the skills tapped by the test section. The psychometric modeling allows the skill information to be aggregated across items so that meaningful statements can be made from what is essentially an item-by-skill patchwork. Uncertainty in that response pattern is accounted for by generating a mastery probability for each of the skills represented in the test. The basic psychometric machinery used is derived from the rule-space method of Tatsuoka (1995).

For the verbal section, 31 skills were identified. Examples are understanding difficult vocabulary, recognizing a definition when it is presented in a sentence, comprehending long sentences, understanding negation in sentences, choosing an answer based on the meaning of the entire sentence, and understanding writing that deals with abstract ideas. Sixteen mathematical skills were defined, including using basic concepts in arithmetic problem solving; creating figures to help solve problems; recognizing patterns and equivalent forms; understanding geometry and coordinate geometry; using basic algebra; making connections among math topics; dealing with probability, basic statistics, charts, and graphs; and applying rules and algorithms in algebra and geometry. Finally, the writing section was thought to tap 10 skills, such as using verbs correctly; recognizing improper pronoun use; following the conventions of word choice, phrases, and sentence construction; understanding the structure of sentences that contain abstract ideas; and understanding complicated sentences.

As a result of each individual's pattern of item performance, an enhanced score report is generated. An example of such a report is given in Figure 1-1. The report lists the three most promising skills for the student to work on and gives suggestions for improvement. For a diagnosis of *understanding difficult vocabulary*, the suggestion is:

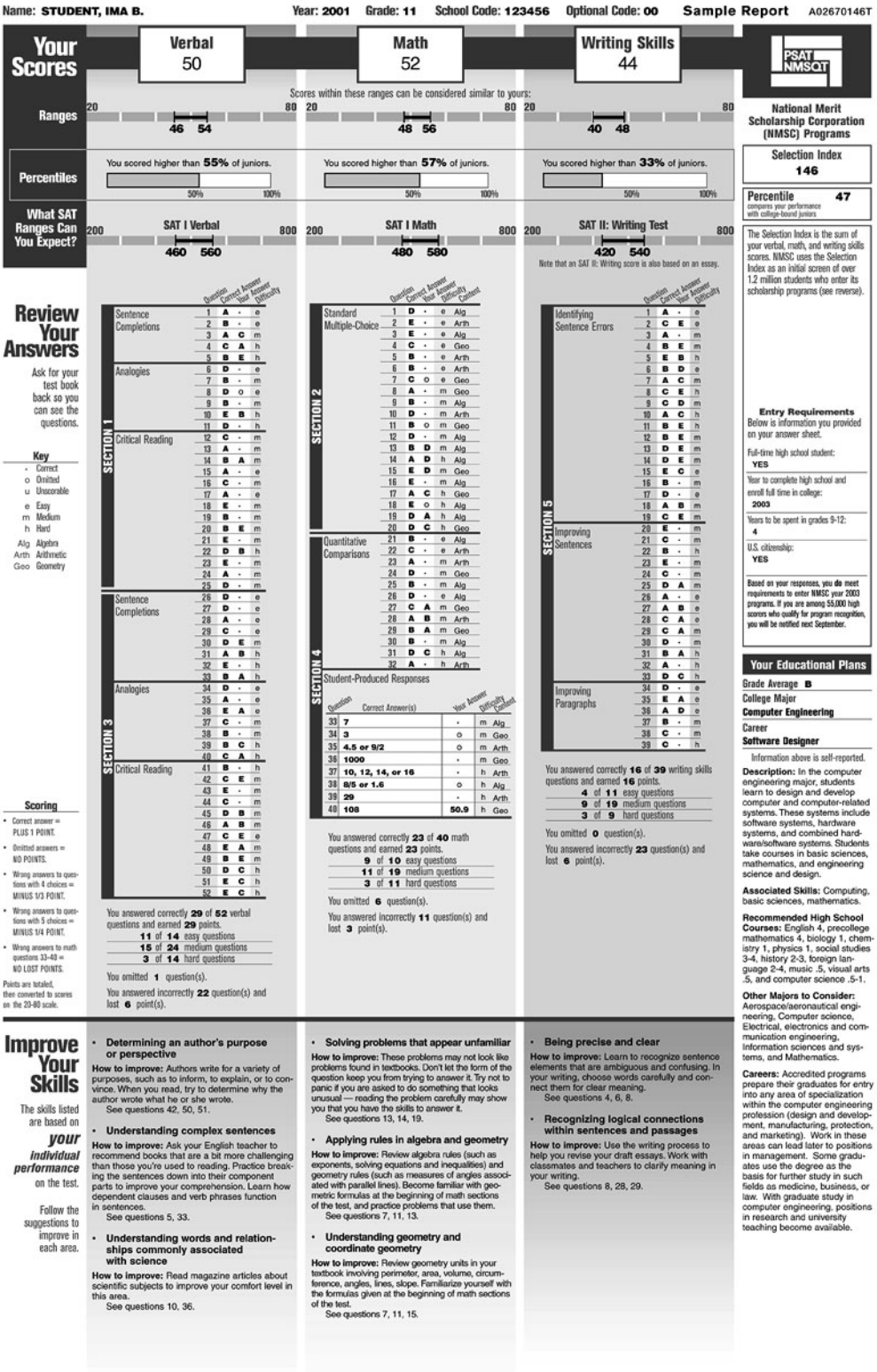


Figure 1-1 Sample enhanced score report for the PSAT. Note the bottom third of the report in which the specific instructional recommendations are provided.
SOURCE: http://www.collegeboard.com/psat/student/html/indx001.html [March 6, 2002]

Broaden your reading to include newspapers and magazines, as well as fiction and nonfiction from before the 1900s. Include reading material that is a bit outside your comfort zone. Improve your knowledge of word roots to help determine the meaning of unfamiliar words.

For a diagnosis of *applying rules and algorithms in algebra and geometry*, the suggestion is:

Review algebra rules (such as exponents, solving equations and inequalities) and geometry rules (such as angles associated with parallel lines). Become familiar with geometric formulas at the beginning of math sections, and practice problems that use them.

There are several issues associated with the provision of such diagnostic feedback that can be informed by empirical analysis. One key concern is whether the skills identified for students explain test performance. Regressing PSAT/NMSQT scaled scores on mastery probabilities is a preliminary means of exploring this question. Such regression produced multiple correlations of .82 for math and .92 for writing on one test form, and .97 for each section on a second form. This initial finding suggests that the probabilities do a reasonable job of explaining test scores and, thus, making more visible the constructs underlying the PSAT/NMSQT. Another issue is whether the same set of skills would be identified for an examinee as needing improvement on other forms of the same test. Preliminary analyses across two forms for the mathematical and writing sections suggest that the proportion of students who would receive the same "needs improvement/doesn't need improvement" designation exceeds chance levels (.50) for the vast majority of skills. However, these results also imply significant variability in the consistency of skill profiles. Such variability is to be expected because the PSAT/NMSQT was not designed with the requisite numbers of items to support fine-grained, highly reliable diagnostics. Some variability in this context may be acceptable, though, because the decisions based on the diagnostics—which concern what to study next—are relatively limited in import and easily reversible. What appears to be highly valued, though, is that the mystery of the PSAT/NMSQT (and SAT I) for many users is revealed by more effective communication of the underlying constructs and by reasonable guidance that moves from test preparation to more construct-relevant instruction. Ultimately, the value of this approach will be determined by the extent to which students successfully engage in learning activities that develop these competencies.

To be sure, the PSAT/NMSQT project represents only a first step. This test was neither designed from a construct definition that would be meaningful to examinees nor intended to be diagnostic. Given those facts, we are limited in how meaningful we can make the construct or how usefully we can guide instruction. The challenge for the future is to design tests *from inception* so that examinees can understand both what is being measured and how to improve their performance on that underlying construct.

Our second example derives from a pragmatic need to generate many assessment tasks efficiently and effectively, which we have begun doing through the use of Test Creation

Assistants (Singley & Bennett, 2002). Not only do we need to generate many assessment tasks, but we also want to be able to design tasks that have prespecified characteristics, including difficulty. To do this, we need to have a better understanding of the cognitive demands associated with particular tasks and task features. Again, the focus here is on our traditional assessments, though the basic approach can be generalized to other types of assessment tasks. The immediate goal is to automatically generate *calibrated* items so that costs can be reduced and validation is built into test development. Items are generated from templates that describe a content class. Each template contains both fixed and variable elements. The variable elements can be numeric or linguistic. Replacing the template's variables with values results in a new item.

The concept of automatic item generation goes back to the criterion-referenced testing movement of the 1960s-1970s, which introduced the notion of generating items to satisfy content specifications and psychometric requirements (Hively, Patterson, & Page, 1968). Further progress was made through research on intelligent tutoring in which generation proceeded from cognitive but *not* psychometric principles (e.g., Burton, 1982). More recent work has merged the cognitive and psychometric perspectives and demonstrated successful, though still experimental, applications (e.g., Bejar, 1993; Embretson, 1998).

The intent of these more recent efforts is to model both content and responses. This modeling can be done from strong or weak theory. Strong theory posits the cognitive mechanisms required to solve items and the features of items that cause difficulty. These approaches use design *principles* in manipulating item content to produce questions of desired difficulty levels. Variation in difficulty may be obtained by creating different templates, each intended to produce items in a particular target range, or by creating a single template to generate items spanning the desired range.

We use both weak and strong theories of performance within this general approach. Weak theory is used when strong theory does not exist, which is true especially in the broad domains covered by most admissions tests, where the intensive cognitive analysis needed to develop strong theory is not practical. Weak-theory approaches also attempt to generate calibrated items automatically, but do so from design *guidelines*. These guidelines constitute a theory of "invariance" which, in addition to indicating which features affect difficulty, suggests which ones do not. Empirically calibrated items spanning the target range are used as the basis for developing templates. Each template is then written to generate items of the *same* difficulty by varying the incidental features. Figure 1-2 is a template—essentially an abstracted representation—for a mathematics problem, while Figure 1-3 illustrates an item generated from that representation.

At ETS we have begun a research initiative to introduce automatic item generation into our large-scale testing programs. The studies cover the mathematical, analytical, verbal, and logical reasoning domains. The issues touch psychometrics (e.g., how does one calibrate items without empirical data?), security (e.g., at what point does a template become overexposed?), and operations (e.g., what tools might be constructed to help test developers create and test item templates?).

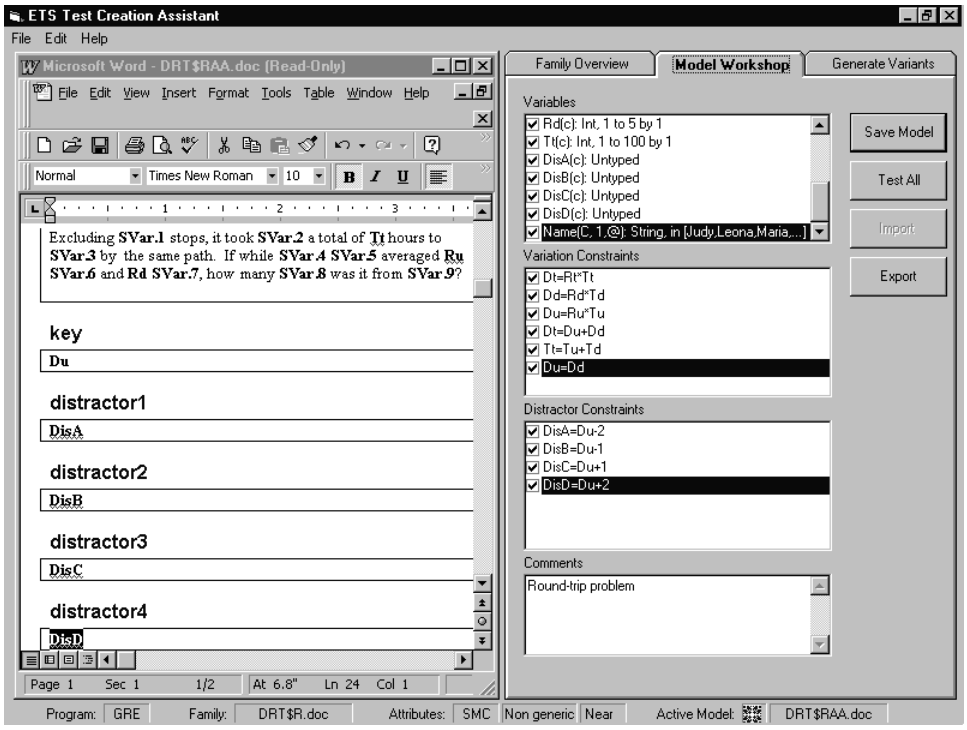


Figure 1-2 An abstracted representation of a mathematics task or *item template*.
SOURCE: ETS Mathematics Test Creation Assistant (TCA)

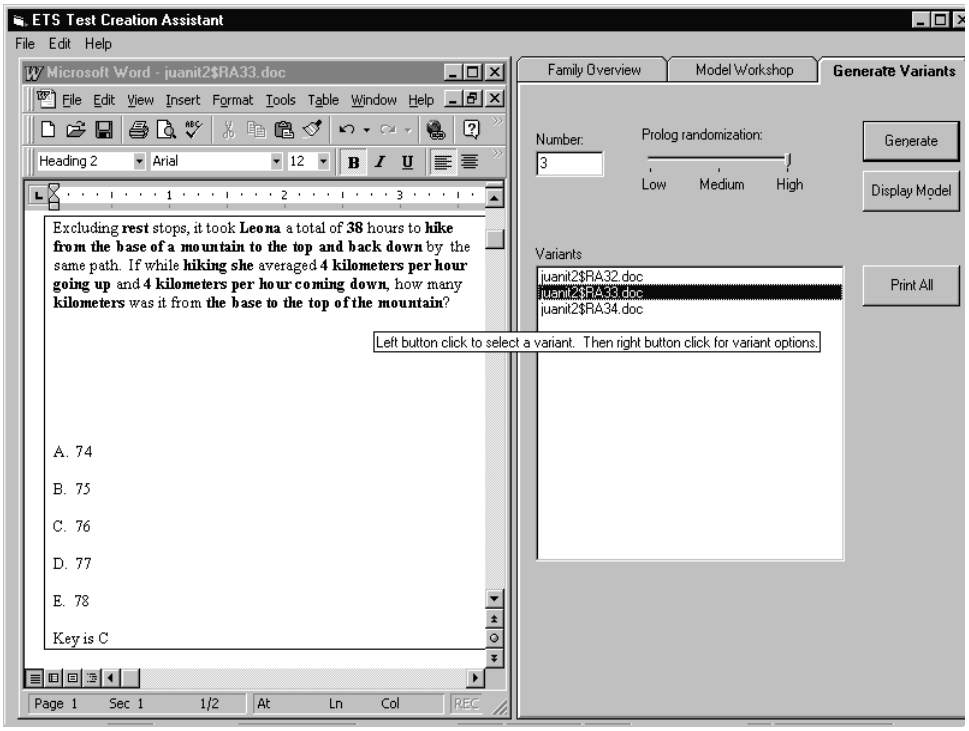


Figure 1-3 A specific task generated automatically from the template.
SOURCE: ETS Mathematics Test Creation Assistant (TCA)

> How does automatic item generation help to unmask the underlying construct? Generation from strong theory is most helpful in this regard because item content is modeled in terms of the demands it places on the cognitive apparatus abstracted from the particulars of any item. Thus, the structures and processes that underlie item performance must be made explicit. Otherwise, item parameters will not be accurately predicted, and the calibration goal will fail. But generation from weak theory may also be revealing because it allows tests to be described, designed, and implemented *not* as a large collection of unrelated problems but, rather, as a smaller set of more general problem *classes* with which we want students to be proficient. Designing tests in this way encourages instruction to focus on developing problem schemas that, according to cognitive theory, constitute the units into which all knowledge is packaged (Marshall, 1995; Rumelhart, 1980).

As an end state, what we would hope to do one day in the not too distant future is to make available to all assessment candidates an entire library of *task models* for all types of assessments. Based on the item templates, each task model would define in a more understandable way an important mathematical problem class. We would aspire to the goal that a full understanding of all task models constitutes a thorough understanding of the relevant domain. Thus, memorizing task models would not be seen as beating the test, but as a legitimate way of learning the domain. This, of course, implies that the set of task models must adequately represent the domain of interest.

Finally, we turn to our work that has the potential to help us develop a fundamentally new generation of assessments. The Evidence-Centered Design Framework (ECD) of Bob Mislevy, Linda Steinberg, Russell Almond, and others (e.g., Mislevy, Almond, Yan, & Steinberg, in press) provides tools and principles for developing assessments that, through every step of the design and delivery process, force a detailed thinking of the constructs to be assessed.

While the two previous examples involve some significant retrofitting and elaboration of existing tests, ECD pushes us into thinking of assessment development as an integrated design process. While ECD doesn't prescribe any particular cognitive-domain model, type of evidence, tasks, or scoring models, it does force designers into considering these aspects of assessment design very explicitly. We will illustrate our points by referring to BIOMASS, a prototype system developed by Mislevy, Almond, Yan, and Steinberg (in press) to assess understanding of transmission genetics. By adhering to a disciplined design process, the developer of an assessment must explicitly consider *and represent* the following:

The Domain—What concepts and skills constitute the domain, how are the various components related, and how are they represented? The domain representation becomes the vehicle to communicate, through the assessment process, the valued nature of understanding. One of the continuing criticisms of standardized assessments is that the domain representations that one would infer from looking at tests is often at odds with more robust conceptualizations of these domains. Therefore, if a domain is represented as a rich and integrated conceptual network, it would not be consistent to have an assessment that queried students about isolated facts. An abstracted representation of the science domain can be viewed in Figure 1-4. This representation highlights the interplay of domain-specific conceptual structures, unifying

concepts, and scientific inquiry understanding as all contributing to an integrated understanding of science.

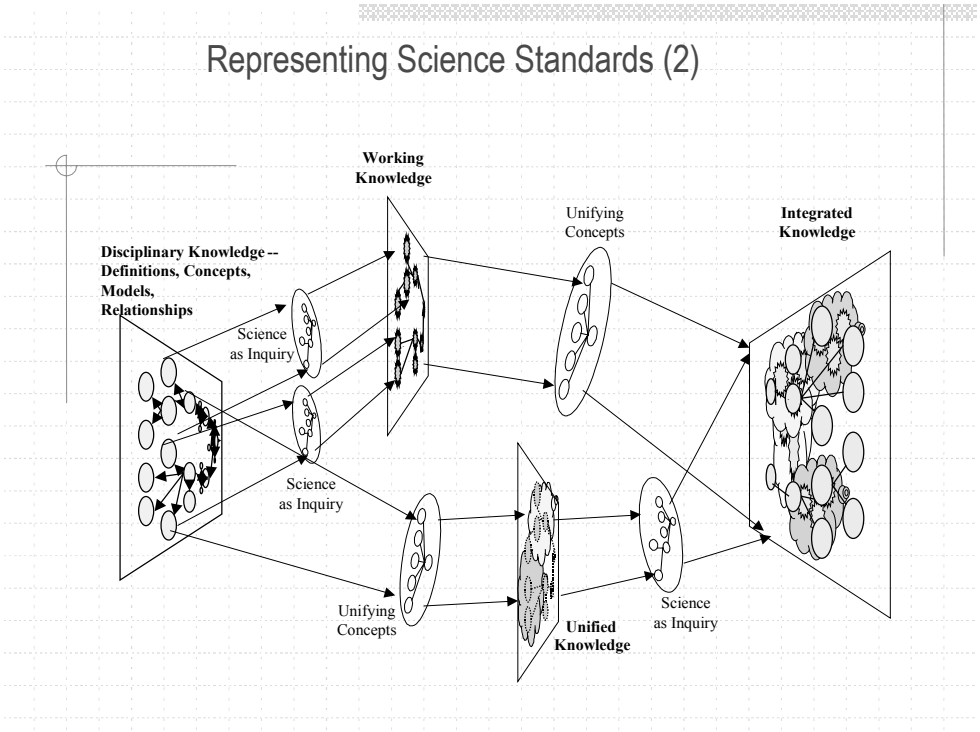


Figure 1-4 An abstracted representation of science understanding
SOURCE: Mislevey et al., in press.

It is also important to use the appropriate communicative methods and symbols for a given domain. Certainly, we wouldn't expect an assessment of musical skill that was strictly verbal, and we wouldn't expect an assessment of mathematics that did not require the use of numbers. Transmission genetics includes a complex conceptual structure as well as a set of domain-specific reasoning skills that are interleaved with genetics concepts. In addition, there are symbolic formalisms that scientists use to represent concepts within the domain.

The Evidence—What are the data that would lead one to believe that a student did, in fact, understand some portion of the domain model? What would a student have to demonstrate to show that he or she could perform at a designated level of accomplishment? Clarifying what the evidence should be is important, not only for the shaping of tasks but also to help students understand in very clear ways what is expected. For a richly represented domain, evidence would likely involve demonstrations of the ability to explain complex relationships. In the case of transmission genetics, evidence of understanding can be gauged, in part, by the ability to explain generational patterns for a variety of plausible conditions.

The Tasks—In light of domain and evidence requirements, assessment tasks can be developed. If the tasks are driven by such requirements, there is a much greater likelihood that the tasks will be focused, relevant, and representative. Note that the path of moving from domain, to evidence, to task is quite different from many traditional test-development practices

in which the availability and constraints of particular tasks shape the assessment development. Note, too, that with an ECD approach, the tasks are more visibly construed as vehicles to elicit evidence, *not* as the definition of the assessment itself. (It is this same conceptual hurdle that must occur among teachers and students generally if assessment tasks are not to be the overwhelming focus of instruction.) In BIOMASS, a small set of complex scenarios with multiple layers have been designed to elicit evidence about students' understanding of transmission genetics. These scenarios, quite compatible with effective biology instruction, provide pieces of evidence relevant to different aspects of science understanding, e.g., disciplinary knowledge, model revision, investigation, etc. For example, one scenario provides evidence of student understanding of investigations and disciplinary knowledge, a second offers evidence of both these aspects together with evidence of understanding of how students revise their working mental models of phenomena (model revision) with new data, and a third is designed to give evidence of model revision only.

ECD also considers the interplay between these and other assessment components. How are tasks selected from an array of potential tasks? How are tasks presented amidst a set of constraints, including delivery options and time available? How are complex responses evaluated? How are response evaluations aggregated so that we can make statements about student performance with respect to the larger domain? Each of these considerations, in conjunction with explicit representations of the domain, the evidence, and the tasks, can give students insight into what matters and how a person can demonstrate specific levels of accomplishment.

CONCLUSION

We believe that each of the three above efforts—enhanced score reporting, automatic item generation, and evidence-centered design—is consistent with the vision espoused in *Knowing What Students Know* of forging a tighter integration of assessment and instruction. Our particular tactic has been to unmask the constructs we measure so that students can more easily improve their standing on them. By forcing a clarification of the domain and a consistent set of representations that govern what students see and how they are evaluated, ECD gives us a methodology for doing exactly that. A logical extension to ECD, automatic item generation, permits us to efficiently instantiate ECD's domain representations in terms of higher order task classes, which can themselves become a legitimate way of learning the domain. Finally, the technology of enhanced score reporting can be used to make clear the specifics of what a student needs to work on to improve. Clearly, these design, item creation, and reporting tools do not guarantee good assessment. But they can help reduce, if not eventually eliminate, the mystery associated with traditional tests, as well as improve the outlook for future assessments.

ACKNOWLEDGMENTS

We are grateful to the following individuals for their reviews of this paper (although the authors are solely responsible for the contents): Russell Almond, Isaac Bejar, Lou DiBello, Dan Eignor, and Linda Steinberg.

REFERENCES

Almond, R.G., & Mislevy, R.J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*, 223-238.

Bejar, I.I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323-357). Hillsdale, NJ: Erlbaum.

Bennett, R.E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing*. Princeton, NJ: Policy Information Center, Educational Testing Service. Also as RR-97-14. Also available: (http://www.ets.org/research/pic/bennett.html).

Brown, J.S., & Burton, R.R. (1978). Diagnostic models for procedural bugs in basic mathematics skills. *Cognitive Science, 2*, 155-192.

Burton, R.R. (1982). Diagnosing bugs in a simple procedural skill. In D.H. Sleeman & J.S. Brown (Eds.), *Intelligent tutoring systems* (pp. 157-183). London: Academic Press.

Embretson, S.E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380-396.

Hively, W., Patterson, H.L., & Page, S. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement, 5*, 275-290.

Marshall, S.P. (1995). *Schemas in problem solving*. New York: Cambridge University Press.

Mislevy, R. J., Almond, R.G., Yan, D., & Steinberg, L.S. (in press). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.). *Generating items for cognitive tests: Theory and practice.* Hillsdale, NJ: Erlbaum.

Rumelhart, D.E. (1980). Schemata: The building blocks of cognition. In R.J. Spiro, B.C. Bruce, & W.F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33-58). Hillsdale, NJ: Erlbaum.

Siegler, R.S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8*, 481-520.

Singley, M.K., & Bennett, R.E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development*. Hillsdale, NJ: Erlbaum.

Steinberg, L.S., & Gitomer, D.H. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science*, *24*, 223-258

Tatsuoka, K.K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. Nichols, S.F. Chipman, et al. (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.

Chapter 2
Technology Supports for Assessing Science Inquiry

Barbara Means and Geneva Haertel
SRI International

The *National Science Education Standards* (National Research Council [NRC], 1996) place inquiry, applied to scientific content areas, at the core of what it means to be scientifically literate:

> Inquiry is central to science learning. When engaging in inquiry, students describe objects and events, ask questions, construct explanations, test those explanations against current scientific knowledge, and communicate their ideas to others. They identify their assumptions, use critical and logical thinking, and consider alternative explanations. In this way, students actively develop their understanding of science by combining scientific knowledge with reasoning and thinking skills. (p. 2)

The *Standards* (NSES) characterize these aspects of science inquiry as a set of "abilities" that all students should exhibit.

Box 2-1 Standards for Science Inquiry, Grades 5-8

- Identify questions that can be answered through scientific investigations.
- Design and conduct a scientific investigation.
- Use appropriate tools and techniques to gather, analyze, and interpret data.
- Develop descriptions, explanations, predictions, and models using evidence.
- Think critically and logically to make the relationships between evidence and explanations.
- Recognize and analyze alternative explanations and predictions.
- Communicate scientific procedures and explanations.
- Use mathematics in all aspects of inquiry.

SOURCE: National Research Council, 1996.

This listing of inquiry abilities should not be interpreted as promoting a framework of discrete, linear competencies. The various aspects of inquiry (e.g., "communicate scientific procedures and explanations" and "think critically and logically to make the relationships between evidence and explanations") get flexibly combined with each other in different permutations in the course of different kinds of science investigations (see Champagne, Kouba, & Hurley, 2000; NRC, 2000). Nor is science inquiry independent of content knowledge.

Inquiry includes examining what is already known in order to effectively plan, conduct, and interpret the results of investigations in specific content areas.

THE IMPORTANCE OF ASSESSMENT

It will not be possible to achieve the goal of this kind of scientific literacy for all—or even most—students without the use of assessments of scientific inquiry. Inquiry assessments are needed within classrooms to help teachers diagnose the nature of their students' understanding and to give students feedback about their performance. Science inquiry assessments are needed within the research and evaluation community to make it possible to compare the efficacy of alternative approaches to supporting science learning. Within accountability systems, science inquiry assessments are needed if teachers and systems are to be held accountable for the NRC standards, but more than that, if we are to avoid sending the wrong message to teachers, students, and parents about what it means to "learn science."

Despite the central importance of inquiry, both as a means for students to acquire a deep understanding of science and as a complex set of interrelated knowledge and processes which in and of themselves are targets of instruction, inquiry is the aspect of science that is least likely to be adequately assessed in large-scale accountability systems. Conventional assessment approaches are quite capable of measuring content knowledge and some of the process skills related to science inquiry (e.g., recognizing confounded variables) in a decontextualized manner. They are ill-suited, however, to capturing multifaceted inquiry processes in meaningful contexts. The conduct of complex, hands-on inquiry is missing from most state, national, and international assessments as well as standardized science achievement tests developed by commercial publishers. Instead, standardized assessments typically emphasize decontextualized factual knowledge (Quellmalz & Haertel, in press). Even when performance or hands-on tasks are administered on a broad scale, their structure and length, and the demand for coverage of a broad range of science content, significantly limit the aspects of inquiry that can be elicited (cf. Baxter & Glaser, 1998).

While classroom assessment practices do not have to conform to the time limits that constrain externally imposed standardized tests, many teachers mimic the format and focus of standardized tests when they are creating assessment tools for classroom use (NRC, 2001a). As a result, even teachers who incorporate extensive inquiry-oriented investigations in their science teaching often score the inquiry work mainly on the basis of "participation" or "completion" and base class grades on conventional tests of factual knowledge from the textbook (Young, Haertel, Ringstaff, & Means, 1998). When classroom assessments do not reflect adequately the engagement required to pursue a line of inquiry or solve a complex problem, the assessment activities are often perceived as dull and disconnected from the hands-on activities (Cognition and Technology Group at Vanderbilt, 1992).

TECHNOLOGY AS CATALYST AND SUPPORT

The NRC report *How People Learn* (1999) makes the point that technology can be used to help teachers understand student thinking and provide meaningful, timely feedback. Nowhere is there greater need and potential for this kind of contribution than in the area of science inquiry

(Brophy et al., 2000; Duschl & Gitomer, 1997; White & Frederiksen, 1998). Increasingly, technology plays a major role in science inquiry in all areas of science. If students use tools and data sets with some degree of authenticity when they engage in science investigations, they are using technology. Under such circumstances, it makes sense to think about capitalizing on the data capture capabilities of the technology to preserve student actions for the purposes of assessment.

Over the last decade, technology-based simulations and environments for science inquiry have been a rich area of research and development (with tools such as GenScope, the Knowledge Integration Environment, and ThinkerTools). Because of the importance of feedback in supporting learning, these software environments have incorporated activities with learner feedback that can be considered embedded assessments.

In contrast to standardized tests and the more conventional paper-and-pencil tests used in most science classrooms, these measures of learning embedded in technology-based learning environments reflect the richness and complexity of science inquiry. They provide examples of ways in which learner choices and the explanations developed within the course of inquiry can provide insights into students' thinking without the interruption of a "test-like" series of questions and answers. Multimedia environments offer opportunities to present students with complex, lifelike situations in which they can pursue a sustained investigation or inquiry. Because students can engage in multiple phases of inquiry (for example, planning an investigation into the quality of the water in a given watershed; collecting data within a simulated environment; organizing and analyzing the data they have collected; forming conclusions and communicating their procedure, findings, and explanations), we can tap not just the individual inquiry "abilities" as stipulated in the Standards, but also students' ability to orchestrate these abilities within a complex task. Technology environments have all kinds of capabilities for capturing the process of student inquiry during this sustained investigation (down to the level of the keystroke if we want that much information) and can accommodate the use of a range of approaches and tools, including collaborative problem solving. Table 2-1 summarizes these capabilities and contrasts them with the features of more conventional assessments.

Despite all this potential, in most cases the rich, technology-based inquiry assessments we can point to are so intertwined with the learning systems within which they are embedded as to be impractical for broader administration (Quellmalz, Haertel, Hoadley, Marshall, & Mishook, 2000). That is, they serve their intended assessment function within the system for which they were developed, but they do not solve the problem of how to assess inquiry activities that are not within that particular learning system.

NEED FOR INQUIRY ASSESSMENTS WITH BROADER FOCUS AND SCALE

NRC (2001a) present six different *purposes* for which educational assessments are used: improving learning, informing instruction, grading, placement, promotion, and accountability. We would add research and evaluation—our own focus—as a seventh purpose. As Atkin and colleagues point out, these different purposes involve different types of people making different

kinds of decisions, and therefore are best served by different (although, ideally, compatible) kinds of assessments.

We find it useful to augment this classification of assessments according to purpose with two related dimensions—the focus and the intended scale of application of the assessment procedure or instrument. Table 2-1 illustrates our framework. Focus refers to the breadth of student understanding or skill the assessment seeks to capture. Many attempts to get at students' thinking, either within the context of research or within the moment-by-moment assessment practices of teachers, are concerned with a very specific aspect of knowledge or skill—the

Table 2-1 Contrasts Between Innovative Technology-Supported Assessments and Traditional Tests

ASSESSMENT FEATURES	TRADITIONAL STANDARDIZED ACHIEVEMENT TESTS	INNOVATIVE TECHNOLOGY-SUPPORTED ASSESSMENTS
Administration	• Individual learners • No collaboration • One common setting • Standardized conditions and procedures	• Individual learners or small groups • Opportunities to demonstrate • social competencies and collaboration • Multiple, distributed settings • Documented but flexible procedures
Item/Task Content	• Typically measures knowledge and facts • Rarely measure inquiry and communication, other than brief writing samples and simple calculations on small data sets	• Measure all aspects of inquiry • Linked to content, inquiry, and performance standards
Item/Task Presentation, Format, and Scaffolding	• Discrete, brief problems • Decontextualized content • Mostly multiple-choice/ "fill-in-the bubble" format • Limited number of constructed response items • Usually no external resources can be used in problem solving	• Extended, performance tasks, including hands-on tasks with use of simulations, probeware, Web searches, visualizations, and multiple representations • Option for access to other resources, including software, the Internet, and remote experts
Scoring and Analysis	• Number and percent correct; percentiles; NCES • Competency-based categorical ratings sometimes used	• Qualitative and quantitative data • Use of scoring rubrics that characterize specific attributes of performance • Potential for automated scoring of natural-language responses (e.g., essays) and complex problem solving (e.g., diagnostic tasks)

Recording and Archiving of Responses	• Paper-pencil • Optical scan	• Mechanisms to reveal steps of problem solving (e.g., Internet trace strategies, electronic notebooks for annotations and describing rationale and documentation of steps) • Web pages • Screen shots • May accumulate responses over time

SOURCE: Adapted from Quellmalz and Haertel, In Press.

content of a single learning activity or even a fragment of an activity. A common example of a broader focus is measurement of understanding and skill at the level of a whole course curriculum. Competence and achievement are broader foci still. These different foci tend to be associated with different purposes (e.g., competence and achievement tend to be associated with accountability systems), but they are logically distinct dimensions, and different combinations do occur. Similarly, the *scale* of the assessment can vary; it can be used for individual diagnosis or for students within a single classroom, within a school or throughout a district or state, or within a given project or program (which could be very small or very large).

One of the things that strikes us in reading the NRC's recent publication *Knowing What Students Know* (2001b) is that, on the one hand, we have large-scale assessment practices that most teachers find wanting for the purpose of informing learning and, on the other hand, we have research-based assessments of very specific aspects of learning. These two types of assessments differ not only in purpose but also in focus and scale. While the research-based assessments provide guidance as to what is important to measure from a learning science perspective and are extremely useful sources of inspiration for new approaches to measurement, they are typically narrow in focus. When critics of embedded assessments and performance assessments deride them as "learning activities," these misgivings reflect a concern that the assessment is so entwined with one particular instructional activity that it could not be used for broader purposes or on a wider scale. Yet, for the purpose of informing learning within a particular instructional unit, it is all to the good if the assessment is seamlessly intertwined with the instructional content of the learning activity. It is when we want to focus on a broader picture of student understanding and skill, and to do so in classrooms where students have had a range of different learning experiences, that such close coupling becomes problematic.

At SRI's Center for Technology in Learning, we have been working to leverage the capabilities of technology and to adapt ideas from system-specific embedded learning assessments (designed for use with specific modules) to the development of assessments with broader applicability. One important impetus to this work is the need that arises within research and evaluation projects to have measures of learning that tap deeper understanding and inquiry skills, yet do so in a way that provides a "fair test" of learning in a reasonably large sample of classrooms, using a range of different software systems or textbooks. Many of the instructional interventions SRI researchers have studied involve the use of the Internet, and SRI has taken advantage of this infrastructure to develop engaging, complex multimedia assessments for delivery over the Web (Center for Technology in Learning, 1999; Coleman & Penuel, 2000;

Means, Penuel, & Quellmalz, 2001; Mislevy, Steinberg, Almond, Haertel, & Penuel, 2000; Quellmalz & Zalles, 1999).

TECHNOLOGY-SUPPORTED PERFORMANCE ASSESSMENTS

We can illustrate this kind of work with one of the assessment tasks we developed for use in evaluating the GLOBE environmental science education program. Students participating in GLOBE follow scientists' protocols and collect environmental data on a local study site. They submit their data to the project database over the Internet and have access to data contributed by 4,000 schools from countries around the world. In recent years, the program has attempted to reinforce aspects designed to promote students' use of the collective GLOBE database to explore questions of their own framing. For our evaluation, we wanted to be able to measure inquiry skills associated with the analysis and interpretation of climate data in both GLOBE and non-GLOBE classrooms.

Box 2-2 Sample Student Justifications for Site Selections

Flagstaff seems like the ideal place for the Winter Olympics to take place. There are about 11 days out of the month of February with sunshine. So the people wanting to watch at the base camp can watch outside with plenty of sunshine and warmth.... The average snowfall for Flagstaff is 1389 mm and it meets the requirements for the O.C. by 389 mm….

Since the temperature at the base level should be at least warm, and if possible, sunny, this proves that Salt Lake City is cool enough compared to Banff, which is too cold -3 degrees Celsius. And, Salt Lake City has up to 5 days of sunshine in February. This keeps the players and spectators more comfortable than if they were at Banff.

Flagstaff best met all of the requirements except in maximum peak temperature. Their temperature was so low, that with the aid of sunlight, their snow could melt.

Elevation, temperature, and the sunny days were all considered when making the choice between the five cities. Although all of the choices would be ideal sites for the winter games only one of the sites could be used. After comparing the data Canada was chosen.

SOURCE: Center for Technology in Learning, 2001.

We have developed several Web-based assessment tasks for our evaluation. One of these tasks, for example, presented students with a set of climate-related criteria for choosing a site for the next Winter Olympics. Given multiple types of climate data on a set of feasible candidate cities, students were asked to analyze the data in terms of the criteria, decide which candidate city best met the climate criteria overall, and prepare a persuasive presentation for the Olympic Committee, complete with graphs of relevant climate data contrasting the city they chose with the default candidate (Salt Lake City). From students' performance on this complex task, SRI researchers derived both measures of specific skills, such as the ability to comprehend quantitative information presented in graphic form, and measures of broader aspects of scientific inquiry, such as the ability to communicate and defend a scientific argument (Coleman & Penuel, 2000). The explanations students provided for their choices (see Table 2-2) revealed both

confusion concerning certain concepts (e.g., "Their temperature was so low, that with the aid of sunlight, their snow could melt") and wide variation in the ability to systematically apply a complex set of criteria. Students who identified the objectively "best" city according to the criteria, but did not provide a systematic data-based justification, could be distinguished from students who did both. (Both sets of students would have been similarly successful on a multiple-choice test.) There were also students who did not choose the "best" site but who approached the task systematically and presented an argument and a set of graphs with data consistent with their choice. Table 2-3 presents the scoring scheme for the Olympic task.

While students enjoyed completing the Web-based assessments, and the assessments served the purposes of our evaluation, such assessments, like those embedded within learning systems discussed above, have limitations and do not satisfy broader assessment needs. To date, much of SRI's effort in assessment has been devoted to finding ways to use technology tools to deliver and capture students' performance. As the challenges associated with technology are overcome, we have begun to turn our attention to other limitations of situation-specific,

Table 2-3 Scoring Scheme for GLOBE Olympic Task

FEATURE	CRITERIA	RANGE
▪ Selection of site meeting all 5 Olympic Committee criteria: ➢ Mountains at least 1000 tall ➢ 1000 mm of snow from Dec. to Feb. ➢ Warm, sunny base camp ➢ Mountain peaks with temperatures consistently below freezing ➢ Latitudes closer to equator (provided snow is adequate)	Flagstaff = 1 All Other Sites = 0	0 - 1
▪ Congruence between selected site and student explanation	If selected site excels all others on dimension(s) cited in student explanation, assign 1 point.	0 - 1
▪ Evidentiary value of graphs presented: Weight of evidence	Take the system-generated evidentiary value* for the selected site for each data graph presented and compute sum.	0 - 5
▪ Evidentiary value of graphs presented: Efficiency of evidence	Average evidentiary value for the graphs presented (score for #3 divided by the number of graphs).	0 - 1
▪ TOTAL Score	Sum of scores	0 - 8

* For each parameter that could be graphed, the system computes an "evidentiary value" for each site equal to the value of the site on the selected parameter relative to the value of the best possible site on that parameter.

embedded, Web-based assessment tasks that can impact the validity of the scores they generate and the applicability of the tasks and scores in varying classroom contexts. We note the following limitations of our own and others' work: Each assessment task covers only a narrow piece of curriculum, and a broad set of assessment tasks guided by an assessment framework of inquiry skills within content areas is generally absent. Scoring rubrics are developed for each task or set of tasks used within a given project; their relationship to rubrics used to score other, related tasks used in other projects is not explicit. In some cases, teachers were not involved in the design of the assessment tasks; and the tasks are labor-intensive to develop and score. In light of these limitations, we have concluded that a "one off" approach to assessment will not be sufficient to meet the needs for assessments with a broader focus and scope, as identified in Table 2-4.

At the same time, some of our SRI colleagues have been exploring the use of assessment templates in designing classroom assessment tools. They have implemented this approach to support the GLOBE environmental education program described above. The GLOBE database contains student-collected data from more than 20 protocols in four investigation areas (atmosphere, hydrology, land cover, and soil). SRI has developed templates for assessing

Table 2-4 Three Dimensions of Educational Assessment

PURPOSE*	FOCUS	SCOPE OF APPLICATION
Improving learning	Learning act	Nation
Informing instruction	Instructional module	State
Placement	Course	Project/program
Promotion	Competencies or achievement	District
Accountability		School/grade
Research & evaluation		Class
		Individual

SOURCE: Adopted from Natural Research Council, 2001a.

students' ability to plan, conduct, analyze, compare, interpret, and communicate investigations with environmental data (Quellmalz, Hinojosa, & Rosenquist, 2001). Teachers have access to a Web-based set of exemplar assessments and to tools for customizing the templates to create their own data inquiry assessments (i.e., they can choose the particular inquiry abilities and type of data with which they want students to work).

FUTURE DIRECTIONS: PRINCIPLED ASSESSMENT DESIGNS FOR INQUIRY

Our experiences developing technology-based assessment tasks for use within evaluation studies and by classroom teachers left us convinced of the potential contributions technology could make to assessment practices, but at the same time highly aware of the need for a more systematic approach to the enterprise. The work of Robert Mislevy and his colleagues (Mislevy,

Steinberg, Breyer, Almond, & Johnson, 1999; Mislevy, Almond, Yan, & Steinberg, 1999; Mislevy et al., 2000) offered a set of principles and a guiding conceptual framework for assessment design, as well as a demonstration that measurement models could be applied to complex assessments such as those needed to assess science inquiry. This "evidence-centered design" framework consists of: (1) a student model, explicating the relationships among the inferences the assessor wants to make about the student; (2) an evidence model, specifying what needs to be observed to provide evidence for those inferences; and (3) a task model, identifying features of the assessment situation that will make it possible for the student to produce that evidence.

Application of the evidence-centered assessment design model and associated statistical techniques has the potential to address many of the issues arising in more situation-specific science inquiry assessment work, such as that performed by SRI, Vanderbilt's Cognition and Technology Group (1992), White and Frederiksen (1998), and Duschl and Gitomer (1997).

Working with Mislevy on an Interagency Educational Research Initiative (IERI) planning grant, we conceived of Principled Assessment Designs for Inquiry (PADI) as an approach to creating assessments for classroom and research use that would cover a broader spectrum of the science curriculum; incorporate cognitive research on learning in specific science domains and in areas of inquiry; build on a robust measurement model; and demonstrate the power of technology to support assessment design, development, delivery, and interpretation.

The essential PADI concept is a system for developing reusable assessment-task templates, organized around schemas of inquiry that are based on research from cognitive psychology and science education. The completed system will have multiple components, including: generally stated rubrics for recognizing and evaluating evidence of inquiry skills; an organized set of assessment development resources; and an initial collection of schemas, exemplar templates, and assessment tasks.

In planning for this project, we quickly realized that if we wanted to develop templates and assessment development tools that would support the work of curriculum developers, we should involve curriculum developers in both the design and the evaluation of the templates and tools. The team for the recently funded PADI implementation project complements SRI's expertise in science inquiry and technology development and Mislevy's assessment design and psychometric expertise with the science education and curriculum development knowledge of Nancy Songer, principal investigator for the University of Michigan's IERI-funded BioKIDS Project, and Kathy Long, who leads the Full Option Science System (FOSS) project at the Lawrence Hall of Science, University of California, Berkeley. The BioKIDS curriculum consists of eight weeks of inquiry-fostering activities focusing on biodiversity. While the program will be used by tens of thousands of learners nationwide in upcoming years, the primary focus is on 5th and 6th grade students in high-poverty urban classrooms within the Detroit public schools. The BioKIDS curricular sequence includes activities to build students' ownership and control of inquiry thinking over time. Students begin their exploration of biodiversity through focused fall and winter monthly observations of their local schoolyard. Data are collected on animal distribution and seasonal changes across city regions. Students systematically explore data and organize their understandings in the form of species accounts that are compiled in an electronic

field guide. Students' own questions focusing on animal distribution, interdependence, and the impact of humans on animal diversity are explored through the comparison of city and national park data on similar species. The PADI assessment will allow BioKIDS to systematically characterize students' understandings over time, as their inquiry understandings develop across the various curriculum units. FOSS middle school courses, each of which requires 9-12 weeks, cover the content areas of earth/space, life, and physical sciences/technology. Lawrence Hall of Science estimates that 60 teachers and 10,000 students have participated in the development and testing of these curriculum units. FOSS focuses on supporting student learning in three areas: understanding science content, conducting investigations, and building explanations. FOSS developers have had great success in developing assessments for the science content and building explanations variables, but have found assessment of the inquiry skills entailed in conducting investigations more of a challenge. The PADI project is expected to provide a theoretical and practical framework that can advance the FOSS assessment system and provide teachers with critical tools to improve student learning.

In addition to these partnerships with curriculum development projects, the PADI team will be strengthened by the participation of Mark Wilson, professor at the University of California, Berkeley and an expert in the psychometric modeling of cognitive structures. Wilson brings his experience modeling cognitive structures in the area of science inquiry (Roberts, Wilson, & Draney, 1997; Wilson & Sloane, 2000) and his M2RCMI measurement model (Adams, Wilson, & Wang, 1997), which will be used to support the scoring of the assessment tasks.

PADI will have multiple components, including:

- a classification of different types of science inquiry tasks, each of which can become the basis for an assessment "template";
- generally stated rubrics for recognizing and evaluating evidence of inquiry skills within each developed template;
- an organized set of assessment development resources;
- an initial collection of schemas, exemplar templates, and assessment tasks produced in the context of the BioKIDS and FOSS projects; and
- a statistical model that will support rigorous analyses of student learning.

In addition, we will be exploring a multidisciplinary, multi-institutional, co-development process in which knowledge engineers, software developers, psychometricians, content experts, curriculum developers, and teachers form a networked improvement community (NIC) around the design and evaluation of PADI assessment tasks. NIC members will both contribute to and take from the pooled resources of the community.

PROGRESS TO DATE

During the past year's planning grant effort, we applied the PADI conceptual framework to existing assessment tasks from two SRI projects (the GLOBE classroom assessments described above and a computer-based environment for learning chemistry). Working with the individuals who designed the original assessment tasks, we applied the evidence-centered design

framework to produce prototype reusable task templates, built around inquiry schemas in the environmental and physical sciences.

SRI staff who were very familiar with the GLOBE and chemistry curricula but less familiar with the PADI framework, completed the retrofitting process, which involved specifying the student, evidence, and task models for each of the investigation phases included in the science curricula. For the *student model*, we identified those science inquiry concepts and skills that students would be expected to know. In specifying the *evidence model*, we first identified the concepts on which each observable variable would depend. Second, we developed a generalized rubric to score the observations conducted within each investigation phase. The *task model* included the specification of representational forms that student work products would take. An example of a work product is "an ordered list of free-form phrases describing the steps in an investigation plan." The task model also included the presentation materials and properties that students would use in creating their work products (e.g., tools, technology affordances, and materials). For example, students might be asked to create a drawing or animation that illustrates a phenomenon or to record data in a log. By retrofitting these assessment tasks to the PADI framework, we were able to demonstrate the usability of the PADI design processes with individuals who were new to the approach, but whose skills and backgrounds were similar to those of our curriculum development partners.

RESEARCH ON QUALITY OF EVIDENCE OF STUDENT LEARNING AND SCALABILITY

The PADI project will conduct research on whether the assessments that are generated provide better evidence about students' inquiry skills and whether the PADI design process is scalable. Working with our curriculum development partners, we will conduct an evaluative study to examine the quality of evidence yielded by the PADI assessments. We will compare the evidence of student inquiry from three sources: cognitive analyses (think-alouds) of inquiry problems, inquiry tasks used as part of large-scale reference exams (e.g., NAEP, TIMSS, or New Standards), and the newly developed PADI assessment tasks.

To better understand the scalability of the PADI process, we will study the assessment design and implementation process. To achieve scalability, we seek to develop our conceptual framework, implementation framework, templates, and design supports at a level of generality that can be applied in different science content areas. The FOSS and BioKIDS implementation sites will provide access to hundreds of middle school students and teachers in diverse settings. The contributions of FOSS and BioKIDS curriculum developers and teachers, and the problems they encounter with our tools and assessments, will be documented in a qualitative study. We will also describe the use of the assessments in different classroom contexts, including urban schools and schools with considerable experience implementing inquiry science, as well as those with less experience.

THE ROLE OF TECHNOLOGY IN PADI

Technology will support almost every component of PADI. The various categories of science inquiry tasks will be realized as reusable software templates that allow curriculum developers to "fill in the slots" in much the way GLOBE teachers customize the classroom assessments described above. SRI staff will work with BioKIDS and FOSS to develop Web-based exemplar assessment tasks using these templates. A software instantiation of Wilson's M2RCMI measurement model will be applied to the tasks. Electronic communication and online repositories of resources will support the networked improvement community (NIC). Thus, technology will play an important role in the design, dissemination, presentation, and scoring of PADI assessment tasks.

CONCLUSION

The recent National Research Council publication *Knowing What Students Know* (2001b) asserts that "Developers of educational curricula and classroom assessments should create tools that will enable teachers to implement high-quality instructional and assessment practices, consistent with modern understanding of how students learn and how such learning can be measured" (p. 306). In complex domains, such as science inquiry, where the knowledge and skills being assessed are numerous, interdependent, and executed over an extended timeframe, it is unlikely that this goal can be attained without the use of technology supports.

REFERENCES

Adams, R.J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1-23.

Baxter, G.P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessment. *Educational Measurement: Issues and Practices*, *17*(3), 37-45.

Brophy, S., Elder, S., Pfaffman, J., Martin, T., Mayfield, C., Vye, N., & Zech, L. (2000). Expanding new methods of technology-embedded assessment and instruction. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Center for Technology in Learning. (1999). *GLOBE year 4 evaluation: Evolving implementation practices*. Menlo Park, CA: SRI International.

Center for Technology in Learning. (2001). *GLOBE year 5 evaluation*. Menlo Park, CA: SRI International.

Champagne, A.B., Kouba, V.L., & Hurley, M. (2000). Assessing inquiry. In J. Minstrell & E. van Zee (Eds.), Inquiring into inquiry teaching in science. Washington, DC: American Association for the Advancement of Science.

Cognition and Technology Group at Vanderbilt. (1992). The Jasper series as an example of anchored instruction: Theory, program description, and assessment data. *Educational Psychologist*, *27*, 291-315.

Coleman, E., & Penuel, W.R. (2000). Web-based student assessment for program evaluation. *Journal of Science Education and Technology, 9*, 327-342.

Duschl, R.A., & Gitomer, D.H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment, 4*, 37-73.

Means, B., Penuel, W., & Quellmalz, E. (2001). Developing assessments for tomorrow's classrooms. In W. Heinecke & L. Blasi (Eds.), Research Methods for Educational Technology. Volume One: Methods of Evaluating Educational Technology. Greenwich, CT: Information Age Press.

Mislevy, R.J., Almond, R.G., Yan, D., & Steinberg, L.S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K.B. Laskey & H. Prade (Eds.), Proceedings of the fifteenth conference on uncertainty in artificial intelligence (pp. 437-446). San Francisco: Morgan Kaufmann.

Mislevy, R.J., Steinberg, L.S., Almond, R.G., Haertel, G.D., & Penuel, W.R. (2000, February). Leverage points for improving educational assessment. Paper prepared for the Technology Design Workshop, SRI International, Menlo Park, CA.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (1999, April). A cognitive task analysis, with implications for designing a simulation-based performance assessment. Presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

National Research Council. (1999). *How people learn: Brain, mind, experience, and school*. Committee on Developments in the Science of Learning. J.D. Bransford, A.L. Brown, & R.R. Cocking (Eds.). Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2000). *Inquiry and the National Science Education Standards: A guide for teaching and learning*. Washington, DC: National Academy Press.

National Research Council. (2001a). *Classroom assessment and the National Science Education Standards*. Atkin, M.J., Black, P., & Coffey, J. (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2001b). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Quellmalz, E., & Haertel, G. (in press). Breaking the mold: Technology-based science assessment in the 21st century.

Quellmalz, E., Haertel, G.D., Hoadley, C., Marshall, S., & Mishook, J. (2000). *21st century assessment planning grant: Final report* (PDU-99-086). Menlo Park, CA: SRI International.

Quellmalz, E., Hinojosa, T., & Rosenquist, A. (2001). Design of student assessment tools for the Global Learning and Observations to Benefit the Environment (GLOBE) program. Presentation at the annual GLOBE International Conference, Blaine, WA.

Quellmalz, E., & Zalles, D. (1999). *World student assessment report: 1998-99*. Menlo Park, CA: SRI International.

Roberts, L., Wilson, M., & Draney, K. (1997, June). *The SEPUP assessment system: An overview (BEAR Report Series, SA-97-1)*. University of California, Berkeley.

White, B.Y., & Frederiksen, J.R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*, 3-117.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181-208.

Young, V.M., Haertel, G., Ringstaff, C., & Means, B. (1998). *Evaluating global lab curriculum: Impacts and issues of implementing a project-based science curriculum*. Menlo Park, CA: SRI International.

Chapter 3
Is It Worth It? Some Comments on Research and Technology in Assessment and Instruction

J.D. Fletcher
Institute for Defense Analyses

As is true of many things, technology, specifically computer technology, offers both challenges and opportunities. Computer technology is becoming increasingly powerful, ubiquitous, and affordable. Computers are turning up in our automobiles, refrigerators, and hair-dryers, and their effects on our lives and daily routines may have only begun. The challenges this technology presents include rapidly changing work procedures and priorities, which in turn affect what our education and training institutions must do. Computer technology influences not only what we do but also what we choose to do and aspire to accomplish. It affects the structure and organization of our established institutions, as well as the way they go about their business. These issues are as real and challenging for educators concerned with assessment as they are for every other sector of human activity. The effort required to meet these challenges naturally raises questions about whether the promised opportunities outweigh the resources needed to bring them about. In short, is it (the effort) worth it (the new capabilities computer technology offers)? This paper discusses the opportunities and capabilities promised by computer technology for assessing and ensuring human competence, and it suggests some research directions that will help bring these opportunities and capabilities to fruition. It particularly concerns technology used to perform the assessments needed to tailor instruction to the needs of individual students, thereby helping to ensure that the instruction reliably produces its intended outcomes for all. Discussion of these issues, then, may best begin with a perspective on the promise of technology for instruction.

THE THIRD REVOLUTION IN INSTRUCTION

Among other things arising from the ubiquity of computer technology may be a third revolution in instruction—"instruction" being a catch-all term for education, training, and tutoring. From this viewpoint, the first revolution was the development of writing about 7,000 years ago. Writing allowed the content of advanced ideas and instruction to transcend time and place and thereby effect a revolution in instruction. In addition to reviewing trade accounts pressed into mud tablets, people with enough time and resources could study the thoughts of the sages without having to rely on face-to-face interaction or the vagaries of human memory.

The introduction of books produced from moveable type was the second major revolution in instruction. Printed books were first produced in China around 1000 A.D. and in Europe in the mid-1400s (Kilgour, 1998). As with writing, books provided access to learning content that

was available anytime, anywhere, but they also increased accessibility to learning by reducing costs. Books effected major changes in both the techniques and, notably, the objectives of instruction. Curriculum and syllabi were altered to take advantage of the availability of the learning content in books. Moreover, books contributed to the rise of a middle class that, in turn, increased the demand for more access to learning content through more books.

Computer technology may now be effecting a third revolution in instruction. This technology makes both the content and the interactions, the tutorial give-and-take, of learning widely and inexpensively accessible. Computer-based instructional materials are available anytime and anywhere, but they also provide relevant and appropriate instructional interactions. They can be designed to adapt and respond to the needs and intentions of individual learners on a microsecond to microsecond basis. They may foment a third revolution in instruction that is at least as significant as the previous two. We might, therefore, ask if there is any evidence that this revolution is occurring and what role technology-based assessment has played in this activity.

WHAT ARE THE CONTRIBUTIONS OF TECHNOLOGY TO INSTRUCTION?

Computer technology has from the beginning been used interactively to tailor the pace, content, difficulty, and sequencing of instructional material to the needs of individuals. Research, development, use, and assessment of computer applications in instruction began in the mid-1950s. Relevant research and development were well underway by the late 1950s and early 1960s in universities (Holland, 1959; Porter, 1959; Bitzer, Braunfeld, & Lichtenberger, 1962; Suppes, 1964), industry (Uttal, 1962), and the military (Fletcher & Rockway, 1986).

We know that substantial improvements in instructional effectiveness may be obtained by tailoring instruction to the needs and capabilities of individual learners. One widely cited discussion was based on studies performed by Benjamin Bloom and his students (Bloom, 1984), who compared the achievement of individually tutored students (one instructor for each student) with that of classroom students (one instructor for every 28-32 students). It is not surprising to find that individual tutoring in these studies increased the achievement of students. What is surprising is the magnitude of the increase. Bloom reported that the overall difference in achievement across three studies was about two standard deviations, which means, roughly, that tutoring improved the achievement of 50th percentile students to that of 98th percentile students. Two standard deviations is a large difference. Bloom posed it to educators as a 2-sigma challenge.

Why is this 2-sigma difference such a challenge? Why don't we simply provide one-on-one tutoring for all our students? The answer is straightforward and obvious: We can't afford it. The provision of one instructor for each student is, in most cases, prohibitively expensive. Individualized, tutorial instruction seems both an instructional imperative and an economic impossibility.

We may now have the means to break out of this dilemma. Gordon Moore's (famous) law states that the power and memory of computers double about every 18 months (Brenner,

1997). The increasing power and affordability of computer technology, combined with its ability to adapt its interactions in real time and on demand, should help solve the problem for us. Its promise for assessment and instruction has not been lost on researchers and developers.

TECHNOLOGY AND ASSESSMENT IN INSTRUCTION

How might assessment best be used to achieve this promise? One way concerns the speed, or "pace," at which students learn in classrooms. Classroom teachers regularly report on the differences in the time different students need to achieve instructional objectives. These reports are supported by empirical findings like the following:

- Ratio of time needed by individual kindergarten students to build words from letters: 13 to 1 (Suppes, 1964);
- Ratio of time needed by individual hearing-impaired and Native American students to reach mathematics objectives: 4 to 1 (Suppes, Fletcher, & Zanotti, 1975);
- Overall ratio of time needed by individual students to learn in grades K-8: 5 to 1 (Gettinger, 1984); and
- Ratio of time needed by undergraduates in a major research university to learn features of the LISP programming language: 7 to 1 (private communication, Corbett, 1998).

That these differences exist should come as no surprise. As with Bloom's findings, what is surprising is their magnitude. Doubtless these differences are due in part to ability, but as Tobias (1982) and others have found, prior knowledge appears to be a major factor, one that quickly overtakes ability in accounting for the speed of learning.

These differences can be accommodated by instruction that takes into account both ability and prior knowledge. Such instruction can take advantage of what students know and concentrate on what they have yet to learn, but tailoring instruction in this way represents a difficult, almost impossible, challenge to classroom teachers working with 20-30 (or more) students. However, technology-based instruction has been tailoring or individualizing instruction practically from its beginning. The benefits of doing so are verified by empirical studies. "Meta-analyses" that compare the time students take to reach a threshold of achievement under technology-based and classroom instruction find an overall time savings of about 30 percent for technology-based instruction (National Research Council [NRC], 1997). These savings matter. For instance, they could reduce by about a fourth the $4 billion the Department of Defense (DoD) spends annually on specialized skill training.

These savings also matter in our K-12 classrooms. Aside from the obvious motivational issues of keeping students interested and involved in educational material, using their time well will profit both the students and any society that will eventually depend on their competency and achievement. The time-savings offered by technology-based instruction in K-12 education could be more significant and of greater value than those obtained in post-education training.

Often the assessments needed to support this approach are accomplished, even in technology-based instruction, by the use of explicit tests such as we find in Keller's Personalized

System of Instruction (Keller, 1968). We may now be in a position to progress beyond explicit assessment to something less visible, less obtrusive, and, notably, continuous. Specifically, we may begin to employ the kinds of transparent assessments found in "intelligent" tutoring systems. True systems of this sort are generative—they produce instructional interactions on demand and in real time as needed by individual students. They accomplish this in what has become a commonly accepted practice of maintaining a model of the subject matter, a model of what the student knows or does not know about the subject, and a collection of procedures intended to bring about targeted instructional objectives.

In these applications, the student model is created by analyzing a student's responses in interactions as they occur and inferring from these what the student knows and does not know by mapping his or her responses onto the "expert" model (represented by the model of the subject matter). Or the student model can consist of a parallel model of the subject matter that accounts for the student's misconceptions (e.g., Fletcher, 1975; Brown & Burton, 1978; Corbett, Koedinger, & Anderson 1997; VanLehn & Niu, in press). The assessment is accomplished continuously and transparently. This is a promising line of development.

WHAT ARE THE BENEFITS OF ASSESSMENT?

Before investing in such a line of development, we might want to know something about its benefits. Payoffs from assessment transcend instructional applications and extend beyond education to military and industrial applications for screening, classifying, and ranking individuals. These latter applications tend to separate out personnel actions, such as selecting individuals for accession or hiring and classifying them into occupational categories. False positives in these cases can be costly. For example, it costs about $4 million to fully train an Air Force F-16 pilot and about $8 million to fully train an F-15 pilot (F15s have two engines and F16s have only one, which accounts for most of this cost difference). It is an expensive matter to select an individual for this type of training if he or she will not be able to complete it successfully.

Aircraft operation is not the only expensive training performed by the military and industry. There are other examples of instruction involving operation, maintenance, and deployment of complex equipment. These costs are increasing because of the continuing infusion of technology into military and industrial operations, and attrition from training is a serious and expensive matter for both sectors. More reliable, valid, and precise assessment to select, classify, and/or certify individuals is at an increasing premium in both sector.

What is the value of our current efforts to select individuals for accession? Within the military, the impact of personnel assessment research has been substantial. Zeidner and Johnson (1989) estimated that savings for the first tour of duty resulting from the Army's use of personnel selection, classification, and assignment procedures compared to random selection, classification, and assignment are about $414 million annually and that savings could be increased to $1 billion annually through simple adjustments in policies and procedures. Improved classification procedures for clerical, surveillance, and communications jobs have been estimated to save the Army $25 million per year compared to previous methods (Grafton, 1990).

The cost-benefits of some future improvements have also been estimated. An increase of 3 percent in the validity of the current test battery used by the Navy for personnel classification could result in an annual savings of $83 million in performance improvement (Schmidt, Hunter, & Dunn, 1987). Using the recently developed Enlisted Personnel Allocation System to supplement the current system of classifying soldiers for jobs would save the Army nearly $480 million per year (Grafton, 1990). The impact of personnel assessment research and development on sectors of the economy outside the military was estimated by Hunter and Schmidt (1982) to be equally substantial. Hunter and Schmidt suggest that the productivity improvement likely to result from replacing univariate selection models with multivariate ones would amount to $43-54 billion a year. Whatever the actual amounts may be, beneficial results from the continued development and use of personnel assessment procedures on the operational costs of military and civilian organizations are likely.

WHERE DOES TECHNOLOGY COME IN?

How might we improve our personnel assessment procedures? How might we develop precision classification that can identify "aces" for at least some occupation classifications before we begin training or at least very early in the training process? We would like to determine those unique, measurable indicators that characterize a Mozart or a Shakespeare and invest our education and training resources appropriately. Computer technology may make this feasible.

With this technology we may have in hand devices that are capable of opening up and measuring whole new areas of cognition, the significance of which we are now only dimly aware, if at all. More could and should be done to use the unique, multimedia display, timing, and data-recording capabilities of computers to assess knowledge, skills, and abilities of individuals. We may be in a position like that of a person with a telescope not yet turned to the stars or a microscope not yet used to examine a drop of water. We need to look beyond our hard-won, well-wrought psychometric techniques based on paper-and-pencil testing and begin to use our new computer-based tools to full advantage.

Most research and development strategies are built around the concept that scientific principles guide design. This concept is both desirable and feasible, but its opposite is more common. Practice begets principle. We built many bridges before we abstracted bridge-building techniques and principles. In the assessment realm, it may well be time to begin systematic experimentation with many types of new item formats intended to assess the specific, innate capabilities possessed by aces, maestros, and star performers of all sorts. These item formats will produce new conceptions of cognition, which in turn will suggest improved, more targeted item formats. It seems past time to pursue programs intended to promote and encourage such spiral development.

Brown and Burton (1978) embedded such considerations in their "Buggy" computer-assisted instruction program. An entire issue of the *International Journal of Man-Machine Studies* (1982) was devoted to papers on automated psychological testing, many of which involved presentations other than our well-worn multiple-choice items. Hunt and Pellegrino (1984) suggested such an approach as a means to expand our notions of intelligence. A first-rate Air Force laboratory was devoted to exploring these notions until it was disbanded in 1998, when

it was just beginning to document what it was learning about human cognition (e.g., the temporal processing assessment discussed by Chaiken, Kyllonen, & Tirre [2000]). More needs to be done.

ADAPTIVE TESTING

The possibility of adaptive, or "stradadaptive," testing was studied extensively at the University of Minnesota under a multiyear effort sponsored by the three DoD personnel research and development laboratories and orchestrated by the Office of Naval Research. This work focused on the use of technology to select, in real time, specific multiple-choice test items to be presented to examinees based on their responses to earlier items. Overall, the results of this work showed that tests using adaptive techniques could be shorter, more precise, and reliable (Weiss, 1983).

Adaptive testing might also reduce costs for personnel assessment by using computers to administer and score tests and by requiring fewer test items to accurately assess individuals, but costs were not directly investigated in this effort. Further, only one (Church & Weiss, 1980) of the 16 technical reports produced by this effort concerned the use of non-multiple-choice items and instead investigated items that could only be presented through the unique display capabilities of computers. Nonetheless, adaptive testing using adaptive techniques for presenting and scoring items is a significant advance and has been implemented by the DoD in some high-profile areas. For instance, with more than 270,000 potential recruits taking the Armed Services Vocational Aptitude Battery each year at a cost of about $20 per administration, the military has a considerable stake in efficient personnel assessment. The Armed Services are now turning to computer technology to provide both the economic benefits of group testing and the precision and flexibility of individual testing. A computerized version of the Armed Services Vocational Ability Test (ASVAB) has been administered to thousands of recruits since 1998. In this case, technology is making an assessment imperative economically feasible.

SIMULATION

Rather than marching individuals through a series of test items, assessments might immerse them in situations like the ones for which they are being selected or prepared. Simulation has been a prominent, long-established technique for both conducting training and assessing the readiness of individuals, crews, teams, groups, and units to perform military operations. Today, it is supported by devices ranging from plastic mock-ups to laptop computers to full-motion aircraft simulators costing more than the aircraft they simulate. Applications range from the operation of oscilloscopes to the repair of computer printers to the deployment of armies. All sectors, educational, industrial, and the military, use techniques ranging from simulated device operation to role-playing in order to prepare and assess personnel. With its current emphasis on "situated learning," shared mental models, problem solving, and higher-order cognitive processes, instructional use of simulation is becoming as familiar to elementary school children as it is to Air Force pilots and business executives.

But the promise and growth of simulation techniques have masked measurement issues that are now being articulated by psychologists, military commanders, industry leaders, and others who are professionally concerned with assessment. We are just beginning to consider

such psychometric properties of simulation as reliability, validity, and precision, as can be seen in empirical forays into this area by O'Neil and his colleagues (e.g., O'Neil, Allred, & Dennis, 1997a; O'Neil, Chung, & Brown, 1997b). In the free and unscripted flow of simulations, correct decisions can lead to wrong outcomes, and incorrect decisions can lead to success. How do we assess capability under these conditions? Is one pass through a simulation sufficient for assessment or are ten needed? Is one scenario (with its single set of initial conditions) needed or many? Along which dimensions should scenarios be varied? In brief, how should simulated environments be designed to support assessments of individual and group performance?

The realism, or "fidelity," needed by simulations to perform successful assessment is a perennial topic of discussion (e.g., Hays & Singer, 1989; Detterman & Sternberg, 1993). Much of this discussion responds to the intuitive appeal of Thorndike and Woodworth's early argument (1901) for the presence and necessity of identical elements to ensure successful transfer of what is learned in training to what is needed on the job.

Thorndike and Woodworth suggested that such transfer is always specific, never general, and keyed to either substance or procedure. This point of view is echoed in more recent studies of transfer, such as the widely noted paper by Gray and Orasanu (1987) who remark on the "surprising specificity of transfer." As Holding (1991) points out, the identical elements theory is hard to argue with—it seems reasonable to expect task elements mastered in simulation to be performed with some appreciable degree of success on the job.

For dynamic pursuits such as combat where unique situations are frequent and expected, the focus on identical elements often leads to an insistence on maximum fidelity in simulations used for assessment. Because we do not know precisely what will happen, we assume that we must provide as many identical elements as we can. This prescription would suggest a viable approach if fidelity came free, but it does not. As fidelity rises, so do costs. High costs can be borne, but they will also reduce the number, availability, and accessibility of valuable resources that can be routinely provided. We must therefore reduce costs by selecting just the fidelity we need to achieve our objectives. These reductions are as necessary for assessment as they are for training.

There is another issue worth mentioning that involves fidelity, simulation, and assessment. Simulated environments permit an assessment of performance and competence that cannot or should not be attempted without simulation. Aircraft can be crashed, expensive equipment ruined, and lives hazarded in simulated environments in ways that range from impractical to unthinkable without them. Simulated environments provide other benefits for assessment. They can make the invisible visible, compress or expand time, and reproduce events, situations, and decision points over and over. Simulation-based assessment is not a degraded reflection of the real environment we would prefer to use. It allows us to assess aspects of performance that would otherwise be inaccessible.

ASSESSMENT AND NETWORKED SIMULATION

One use of simulation for assessment is receiving increasing and perhaps overdue attention. It concerns the learning and capabilities of collectives (crews, teams, groups, and

organizational units). Concern with collective performance is pervasive and by no means limited to military operations (Cannon-Bowers, Oser, & Flanagan, 1992; Huey & Wickens, 1993). However, in the military, the stakes for collective proficiency are high, and interest in assessing collective behavior is intense. Much current interest in the assessment of collective behavior has centered on the military's development and use of networked simulation.

Networked simulation was originally developed for training applications and was intended to improve the performance of crews, teams, and units (Alluisi, 1991). The individual members of crews, teams, and units who use networked simulation are assumed to be already proficient in their individual skill specialties—they are expected to know how to drive tanks, read maps, fly airplanes, fire weapons, and so on at some acceptable threshold of proficiency before they begin networked simulation exercises. Moreover, the commanders of these crews, teams, and units are expected to possess some basic academic knowledge and practical skills in the command and control of their collectives—they are expected to know at some rudimentary level how to maneuver, use terrain in a tactically appropriate manner, fly helicopters, create and overcome engineered obstacles, etc. The focus in networked simulation is on team rather than individual performance.

Networked simulation consists of modular objects intended to simulate combat entities. Typical entities are vehicles such as tanks, helicopters, and aircraft. During simulation exercises, these vehicles are mostly operated by human crews located in the devices that simulate them. These entities, these simulators, may be located anywhere because they are modular and autonomous and because they all share a common model of the battlefield and its terrain. In a networked simulation exercise conducted on simulated California terrain, a tank crew sitting in a simulated tank in Germany can call for air support from simulated aircraft in Nevada because they are being attacked by a simulated helicopter located in Alabama.

Each entity, along with many others, is connected to the network. If the simulated vehicles encounter allied vehicles on the digital terrain, they can join together to form a larger team and undertake a mission with all the problems of command, control, communications, coordination, timing, and so on that such activity presents. If they encounter enemy vehicles, they can engage in force-on-force engagements in which the outcome is determined solely by the performance of the individuals, crews, teams, and units involved. No umpires, battlemasters, or other outside influences are expected or permitted to affect the outcome of a networked simulation engagement once it begins.

All the digital communication packets used to control networked simulation may be recorded. Generally, each entity issues 3-5 packets per second. Actions undertaken in networked simulation may be recorded in extensive detail for later analyses and replay during After Action Reviews (Meliza, Bessemer, & Hiller, 1994; Morrison & Meliza, 1999). The scene from any vantage point (friendly or enemy, inside or outside vehicles, ground level or "God's eye") can be recorded at almost any level of detail and then replayed for the purposes of assessment. Packets have even been created and used to replay entire battles, such as the 73 Easting combat engagement during the Gulf War (Orlansky & Thorpe, 1992).

Use of networked simulation in assessment has been discussed by Fletcher (1994, 1999) and O'Neil et al. (1997b). The paper by O'Neil and his colleagues is particularly interesting because of its presentation of empirical data on the validity of networked simulation used to assess performance on negotiation tasks. Empirical evaluations concerning the training value of networked simulation used by the military have been summarized by Fletcher (1999) and Orlansky, Taylor, Levine, & Honig (1997).

The report by Orlansky et al. is notable for its careful examination of the cost benefits of networked simulation. These researchers compared the costs of a 5-day close air support (aircraft and ground forces operating together) exercise using linked simulators located in Arizona, Kentucky, and Maryland with a "live" simulation performed in the field using actual equipment. The simulation exercise involved 75 people; a similar exercise in the field with actual equipment would have required 245 people. It cost $267,000 to support the simulation exercise; the field exercise would have cost $2,897,000. Cost per person trained and assessed in the simulation exercise was $3,600; cost per person trained in the field would have been $11,800. As is typical for combat exercises, it was not possible to validate the results of the exercise with real experience (a situation for which we may all be grateful), but steady improvements in combat-relevant tasks were found in the simulation exercise, and its cost benefits for both training and assessment were clearly evident.

Civilian applications of networked simulation for training and education were identified and discussed by Fitzsimmons and Fletcher (1995). These applications were both potential and real. They included two demonstrations involving high school students in DoD schools in Germany, Kentucky, and Korea who collaborated in playing music together ("The World Band") and in designing and flying aircraft using materials available in the early 1900s ("The Wright Flyer"). Although the emphasis in these demonstrations was on education, assessment of such collective issues as teamwork, communication, leadership, interpersonal skills, etc., could easily have been carried out in these demonstrations.

WHERE ARE WE HEADED?

When we consider the possibilities for the use of technology in assessment, it seems reasonable to ask, what will be next? Technology-based instruction appears to be headed for distributed (anytime, anywhere) lifelong learning. It may even be object-oriented, using instructional objects available on the World Wide Web or whatever the global ether will be in the future. These objects will be assembled, on-demand, in real time, in some granular, perhaps item-by-item basis, and tailored to the needs, capabilities, and intentions of individual users, who may be learners, users seeking decision aids, or individuals needing certification for some set of knowledge and skills. The challenges presented by this future are being addressed by the Advanced Distributed Learning (ADL) initiative, which is led by the Department of Defense in coordination with other federal agencies such as the Departments of Agriculture, Education, Labor, Interior, and Health and Human Services; National Aeronautics and Space Administration; National Institute for Standards and Technology; and the White House Office of Science and Technology Policy (http://www.adlnet.org).

The Department of Defense is coordinating development with industry of a Sharable Content Objects Reference Model (SCORM) to ensure accessibility, durability, portability, and reusability of instructional objects and to provide guidelines concerning the creation, archiving, and assembly of instructional objects into relevant instructional presentations. Benefits in terms of saved or avoided personnel and training costs are very close to those identified for technology-based instruction (discussed earlier in this paper). Benefits in terms of improved productivity and effectiveness are more difficult to assess, but they are expected to exceed the monetary value of the ADL initiative. The benefits of allowing assessment to take place at any time, any place, and as needed seem likely but have yet to be systematically determined. That such assessment capabilities will be developed seems equally likely.

In any case, assessment can take advantage of sharable objects. Much, however, remains to be done. How, for instance, can we assemble, aggregate, and sequence different objects at different times to produce assessments that are both fair and comprehensive? Should psychometric data be included in the "meta-data" in which objects are packaged? What do we need to do to certify the quality of these objects? These questions, among others, remain as challenges to those who are concerned with what might be described as object-oriented, technology-based assessment.

FINAL WORD

The above comments suggest a number of areas for research. Four that might be emphasized here are:

- *Transparent, continuous assessment*. How do we, or should we, extract assessment information from the interactions between a student and a teacher, whether human or computer? Master teachers know some of the techniques for doing this, and others have been developed for intelligent tutoring systems. More could and should be done. Our current processes of extracting assessment information once every few years, once a year, or even once a month are insufficient if we hope to use instructional and student time well. The hallmark of good management is continuous assessment. We should develop it.
- *Precision classification*. Every human being should have the assessment tools to develop to its fullest extent whatever package of abilities he or she has been handed at birth. We need more comprehensive models of cognition to do this. These models will have to be keyed to our ability to measure them. Through computer technology, we may have in hand the capabilities to devise new item formats and to pursue, in a spiral of development, both the measures and the models of cognition we need. It seems past time to begin this work in earnest.
- *Assessment based on simulation*. Simulation is widely used by industry and the military to assess the capabilities and preparation of individuals, crews, teams, and units. Given the current emphasis (which despite its rhetorical fluff seems sensible) on approaches involving situated, problem- or project-based learning in (more or less) authentic environments—which are very close to, if not the same thing as, what the military calls simulations—the need to determine what students are learning from these simulated environments seems likely to grow. But how many simulations using what scenarios are

needed to ensure reliable, valid, fair assessment? What are the measurement properties of simulations, and how should we develop them further? There is a great need in both education and military and industrial training for answers to these questions—answers that again must come from vigorous, targeted programs of instruction

- *Object-oriented assessment*. The vision of a World Wide Web heavily populated with objects that are accessible, portable, durable, and reusable seems very likely to occur. These objects are likely to include assessment as well as instructional objects. How should we use these objects to assemble assessments in real time and on-demand as needed by individuals? How would we develop the measurement properties of such presentations to ensure reliability, validity, and fairness? Given the advances made by such efforts as the ADL initiative, we are in a good position to begin the necessary research and development. Again, it seems the time is ripe to begin doing so.

All of these areas present challenges to assessment. As suggested, technology will change not only the way we do assessment but our objectives and expectations for assessment as well. The object of assessment is, of course, not better measurement, although that is clearly an enabling capability. What we seek are better (more reliable, valid, and precise) inferences and decisions based on our assessment. Technology will allow access to areas of human cognition and performance we have been unable to consider with our paper-based techniques, and this, in turn, will necessitate new notions of human cognition and potential. It may enable us to identify human capabilities that might otherwise remain latent and undeveloped. The challenges presented include great opportunities.

In the area of human cognition, we may well seek to identify something that might be called (and has been so called by CRESST) a "learnome." The human genome lists all the micro-components needed for reproduction or replication; the learnome might list all the micro-components needed to reproduce or replicate areas of knowledge or skills. First we need to identify—and measure—these components. If we are successful, we will have made significant progress toward new concepts of cognition and our ability to assess performance of very complex tasks, which seem to be growing increasingly common in both industry and the military (NRC, 1997).

Finally, e-learning is increasing emphasis on learner, as opposed to teacher, classroom, or school, productivity. Learners are expected to be self-motivated, self-guided, and self-regulating in the Webbed world of lifelong learning. Such activity benefits the individual seeking to achieve his or her potential, the organizations depending for their success on human competence, and the nations competing in the global marketplace. All these ends are likely to be well served by tools placed in learners' hands to help them assess progress toward their goals. Technology seems key in developing these assessment tools and making them available anytime and anywhere to those who need them.

REFERENCES

Alluisi, E.A. (1991). The development of technology for collective training: SIMNET, a case history. *Human Factors, 33*, 343-362.

Bitzer, D.L., Braunfeld, P.G., & Lichtenberger, W.W. (1962). Plato II: A multiple-student, computer-controlled, automatic teaching device. In J.E. Coulson (Ed.), *Programmed learning and computer-based instruction* (pp. 205-216). New York: John Wiley.

Bloom, B.S. (1984). The 2-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4-16.

Brenner, A.E. (1997). Moore's law. *Science, 275*, 1551.

Brown, J.S., & Burton, R.B. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*, 155-192.

Cannon-Bowers, J.A., Oser, R., & Flanagan, D.L. (1992). Work teams in industry: A selected review and proposed framework. In R.W. Swezey & E. Salas (Eds.), *Teams: Their training and performance* (pp. 355-377). Norwood, NJ: Ablex.

Chaiken, S.R., Kyllonen, P.C., & Tirre, W.C. (2000). Organization and components of psychomotorability. *Cognitive Psychology, 40*, 198-226.

Church, A.T., & Weiss, D.J. (1980). *Interactive computer administration of a spatial reasoning test* (Research Report 80-2). Minneapolis, MN: Computerized Adaptive Testing Laboratory, University of Minnesota.

Corbett, A.T., Koedinger, K.R., & Anderson, J.R. (1997). Intelligent tutoring systems. In M.G. Helander, T.K. Landauer, & P.V. Prabhu (Eds.), *Handbook of human-computer interaction* (pp. 849-874). Amsterdam: Elsevier Science.

Detterman, D.K., & Sternberg, R.J. (Eds.). (1993). *Transfer on trial: Intelligence, cognition, and instruction*. Norwood, NJ: Ablex.

Fitzsimmons, E.A., & Fletcher, J.D. (1995). Beyond DoD: Non-Defense training and education applications of DIS. *Proceedings of the IEEE, 83*, 1179-1187.

Fletcher, J.D., & Rockway, M.R. (1986). Computer-based training in the military. In J.A. Ellis (Ed.), *Military contributions to instructional technology* (pp. 171-222). New York: Praeger.

Fletcher, J.D. (1975). Modeling the learner in computer-assisted instruction. *Journal of Computer-Based Instruction, 1*, 118-126.

Fletcher, J.D. (1994). What networked simulation offers to the assessment of collectives. In H.F. O'Neil, Jr. & E.L. Baker (Eds.), *Technology assessment in software applications* (pp. 255-272). Hillsdale, NJ: Lawrence Erlbaum.

Fletcher, J.D. (1999). Using networked simulation to assess problem solving by tactical teams. *Computers in Human Behavior, 15*, 375-402.

Gettinger, M. (1984). Individual differences in time needed for learning: A review of the literature. *Educational Psychologist, 19*, 15-29.

Grafton, F. (1990). Improving the selection, classification, and utilization of Army enlisted personnel. In J. Orlansky, F. Grafton, C.J. Martin, W. Alley, & B. Bloxom (Eds.), *The*

current status of research and development on selection and classification of enlisted personnel (IDA Document D-715). Alexandria, VA: Institute for Defense Analyses.

Gray, W.D., & Orasanu, J.M. (1987). Transfer of cognitive skills. In S.M. Cormier & J.D. Hagman (Eds.), *Transfer of learning* (pp. 183-215). New York: Academic Press.

Hays, R.T., & Singer, M.J. (1989*). Simulation fidelity in training system design: Bridging the gap between reality and training*. New York: Springer-Verlag.

Holding, D.H. (1991). Transfer of training. In J.E. Morrison (Ed.), *Training for performance: Principles of applied human learning* (pp. 93-125). New York: John Wiley.

Holland, J. (1959). A teaching machine program in psychology. In E. Galanter (Ed.), *Automatic teaching: The state of the art* (pp. 69-82). New York: John Wiley.

Huey, B.M., & Wickens, C.D. (1993). *Workload transition: Implications for individual and team performance*. Washington, DC: National Academy Press.

Hunt, E., & Pellegrino, J. (1984). *Using interactive computing to expand intelligence testing. A critique and prospectus* (Report 84-2).

Hunter, J., & Schmidt, F. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In M.D. Dunnette & E.A. Fleishman (Eds.), *Human performance and productivity: Human capability assessment*. Hillsdale, NJ: Lawrence Erlbaum.

Keller, F.S. (1968). Goodbye, teacher … . *Journal of Applied Behavior Analysis, 1*, 79-89.

Kilgour, F.G. (1998). *The Evolution of the Book*. New York, NY: Oxford University Press.

Meliza, L.L., Bessemer, D.W., & Hiller, J.A. (1994). Providing unit training feedback in the distributed interactive simulation environment. In R.F. Holtz, J.A. Hiller, & H.H.

McFann (Eds.), *Determinants of effective unit performance* (pp. 257-280). Alexandria, VA: U.S. Army Research Institute.

Morrison, J.E., & Meliza, L.L. (1999). *Foundations of the After Action Review Process* (IDA Document 2332). Alexandria, VA: Institute for Defense Analyses. (DTIC/NTIS AD-A368 651)

National Research Council (1997). *Technology for the United States Navy and Marine Corps, 2000-2035: Becoming a 21st century force vol. 4 Human Resources*. Committee on Technology for Future National Forces. Washington, DC: National Academy Press.

O'Neil, H.F., Allred, K., & Dennis, R.A. (1997). Validation of a computer simulation for assessment of interpersonal skill. In H.F. O'Neil (Ed.), *Workplace readiness: Competencies and assessment* (pp. 229-254). Mahwah, NJ: Lawrence Erlbaum.

O'Neil, H.F., Chung, G.K.W.K, & Brown, R.S. (1997). Use of networked simulations as a context to measure team competencies. In H.F. O'Neil (Ed.), *Workplace readiness: Competencies and assessment* (pp. 411-452). Mahwah, NJ: Lawrence Erlbaum.

Orlansky, J., Taylor, H.L., Levine, D.B., & Honig, J.G. (1997). *The cost and effectiveness of the Multi-Service Distributed Training Testbed (MDT2) for training close air support* (IDA Paper P-3284). Alexandria, VA: Institute for Defense Analyses.

Orlanksy, J., & Thorpe, J. (1992). *Proceedings of conference on 73 Easting: Lessons from Desert Storm via advanced simulation technology held in Alexandria, Virginia on 27-29 August 1991* (IDA Document 1110). Alexandria, VA: Institute for Defense Analyses. (DTIC/NTIS AD-A253 991)

Porter, D. (1959). Some effects of year long teaching machine instruction. In E. Galanter (Ed.), *Automatic teaching: The state of the art* (pp. 85-90). New York: John Wiley.

Schmidt, F.L., Hunter, J.E., & Dunn, W.L. (1987). *Potential utility increases from adding new tests to the Armed Services Vocational Aptitude Battery (ASVAB)*. Unpublished manuscript (Battelle Contract Delivery Order 53). San Diego, CA: Navy Personnel Research and Development Center.

Suppes, P. (1964). Modern learning theory and the elementary-school curriculum. *American Educational Research Journal, 1*, 79-93.

Suppes, P., Fletcher, J.D., & Zanotti, M. (1975). Performance models of American Indian students on computer-assisted instruction in elementary mathematics. *Instructional Science, 4*, 303-313.

Thorndike, E.L., & Woodworth, R.S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review, 8*, 247-262.

Tobias, S. (1982). When do instructional methods make a difference? *Educational Researcher, 11*(4), 4-9.

Uttal, W.R. (1962). On conversational interaction. In J.E. Coulson (Ed.), *Programmed learning and computer-based instruction* (pp. 171-190). New York: John Wiley.

VanLehn, K., & Niu, Z. (in press). Bayesian student modeling, user interfaces and feedback: A sensitivity analysis. *International Journal of Artificial Intelligence in Education, 12*.

Weiss, D.J. (1983). *Final report: Computer-based measurement of intellectual capabilities*. Minneapolis, MN: Computerized Adaptive Testing Laboratory, University of Minnesota.

Zeidner, J., & Johnson, C.D. (1989). *The economic benefits of predicting job performance* (IDA Paper P-2241). Alexandria, VA: Institute for Defense Analyses.

Chapter 4
Speech Recognition Technology and the Assessment of Beginning Readers

Susan M. Williams
The University of Texas at Austin

At the beginning of first grade, most children are entering the initial stage of reading development during which they acquire basic decoding knowledge. Until they become fluent, they must rely on more able readers, such as parents or teachers, to listen to them read and provide assistance when they falter. Technology such as talking books with synthetic or digitized speech can also provide support by reading large blocks of text aloud and giving oral pronunciation and definitions of unfamiliar words. However, talking books provide only passive support. Students must monitor their own decoding and comprehension and request assistance when needed.

Recent research in speech recognition technology has made it possible to develop computer-based reading coaches that listen to students, assess their performance, and provide immediate customized feedback (Mostow & Aist, 1999; Nix, Fairweather, & Adams, 1998). This active support allows students with less knowledge and fewer learning strategies to read independently.

This paper provides an overview of speech recognition technology and how it is being used by computer-based reading coaches to assess the performance of beginning readers. I begin by outlining research related to the importance of frequent practice with feedback for beginning readers. Next, I describe *Watch Me! Read*, an example of a computer-based reading coach. I then provide an overview of speech recognition technology and how it is adapted for use with children and beginning readers. Finally, I discuss issues and possibilities for using speech recognition as an assessment tool.

THE IMPORTANCE OF FREQUENT READING PRACTICE

Research has shown a positive correlation between frequent reading and reading achievement: Frequent reading improves the speed at which words are recognized which, in turn, leads to fewer disruptions in the comprehension process (Perfetti, 1985). Extensive reading leads to enhanced phonemic awareness (Stanovich, 1986). Extensive reading promotes the acquisition of new vocabulary and grammatical constructions (Stanovich & Cunningham, 1982). Children who read more frequently have higher test scores (Cipielewski & Stanovich, 1992; Greany & Hegarty, 1987).

Furthermore, the relationship between frequent reading and reading achievement is reciprocal (Stanovich, 1986), i.e., frequent reading leads to higher achievement which leads to more frequent reading. This means that the gap between more and less frequent readers will grow over time. Stanovich (1986) dubbed this phenomenon the "Matthew effect" as a reference to the biblical passage about the rich getting richer and the poor getting poorer (Matt. 13:12).

This research on frequent reading is consistent with cognitive theories suggesting that regular extensive practice in a skill promotes proficiency (Anderson, 1995; Ericsson & Smith, 1991). Practice at earlier stages of learning is thought to be more beneficial. As a learner increases in skill, additional practice is likely to bring diminishing returns (Anderson, 1995). Thus, additional reading practice in the early grades when many children are learning to read could be especially important. It is noteworthy that most studies of the effects of frequent reading have been done with older children, presumably because younger children are not yet fluent readers. Because of the Matthew effect, the potential value of frequent reading for younger children could be even greater than the results of these studies suggest.

Translating these findings into classroom practice is not straightforward (Byrnes, 2000). While the research seems to suggest that providing more time for reading, especially in the early grades, would lead to increased reading achievement, there is also evidence that certain conditions of practice may be more effective in promoting achievement. For example, at early stages of acquisition, learners often need expert advice to help them understand how they are doing. Formative assessment (and instruction based on that assessment) is especially important for struggling readers who benefit more from scaffolded tutoring than from attempts to read literature on their own (Juel, 1996).

Guthrie (1980) also makes a distinction between the time allocated for reading and the time that students are actually engaged in this task. Teachers differ in their instruction and classroom management strategies and in their ability to keep children "on task." Independent reading attempts by unsuccessful beginning readers can lead to frustration and lack of engagement (Williams, 2000). Thus, allocating time for independent reading is not enough to improve reading performance, especially for beginning and less successful readers. These readers also require feedback and instruction to make the additional time beneficial.

Finding time to provide individual feedback during children's reading practice is difficult for teachers who often have 20 or more students in their class in the early grades. Thus, class size is likely to be a constraint on students' opportunities for the type of reading practice that might be most beneficial.

It is possible that new developments in speech recognition technology could increase opportunities for individual reading practice with feedback, as well as collecting assessment data to inform instructional decision making. In the next section of this paper, I describe Watch Me! Read, a computer-based reading environment developed by IBM's T.J. Watson Research Center and currently being tested in the Houston Independent School District as part of IBM's Reinventing Education program.

WATCH ME! READ

Commercially available reading software often seems somewhat alien to the actual process of learning to read. Such software must resort to clever schemes to compensate for its inability to react directly to youngsters as they read aloud. One common strategy calls upon the child to perform tasks that presume to exercise the same skills that reading requires, with directions like "Find the word on the screen that rhymes with this picture." These types of activities fail to give children much of a sense of the experience of reading—not surprising, given their orientation toward isolated word recognition and their reliance on picture interpretation. They also fail to provide students and teachers with valid, meaningful assessment information because they bear little similarity to the cognitive demands of "real" reading.

Watch Me! Read (WM!R) software is designed to give a young child a sense of being a reader (Nix et al., 1998). Specifically, the designers' goals are to provide reading practice, comprehension awareness, and a sense of reading as communication. The software uses speech recognition to assess a child's performance and provide individualized feedback. It works in much the same way as an adult who listens to the child, provides help with the pronunciation of words when the child falters, and asks questions to probe the child's understanding of what he or she is reading.

In the WM!R environment, books appear on the computer screen much as they do in traditional form, i.e., text and illustrations are displayed on two facing "pages" of a graphic book (Figure 4-1). A small, animated Panda acts as a guide, walking across the surface of the book, pointing to the current reading location, and providing feedback and encouragement. Students are asked to read the text one phrase at a time. For students who are just beginning to read, the Panda reads each phrase first and then students read only the last word when the phrase is repeated. At the most advanced level, students read the entire phrase without assistance. The phrase being read is marked in color: The text read by the Panda is blue; the text read by the student is red. If a student does not know a word, he or she can click on the word and hear the Panda pronounce it. The student's voice is recorded as he or she reads the book. This recording is used later in the performance section of the program.

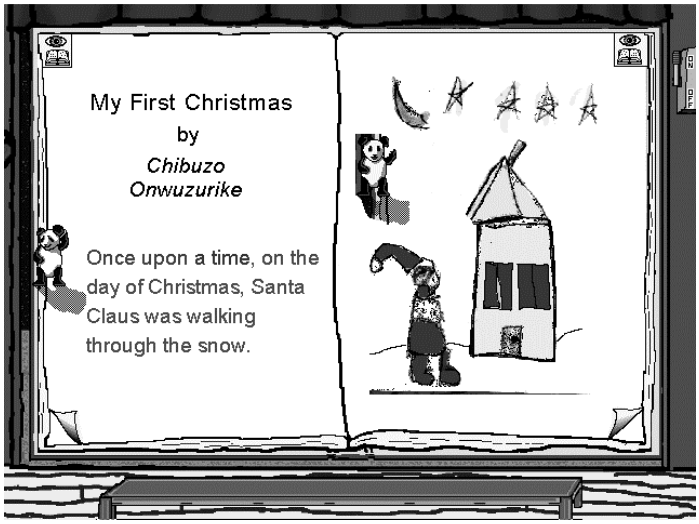


Figure 4-1 Reading view of Watch Me! Read showing a book written and illustrated by a student
SOURCE: Williams, Nix, & Fairweather, 2000, p. 116

At the beginning of each page, the student can choose to hear an overview of what he or she is about to read. At the end of each page, the Panda asks a comprehension or prediction question based on the contents of the page. These questions are customized for each book. The student uses a graphical "boom box" tool displayed at the bottom of the page to record the

answers. The student can listen to his or her answer and re-record it if desired. WM!R does not provide immediate feedback for answers to comprehension questions, but the teacher can review recorded answers at a later time.

After the student finishes reading, WM!R presents a performance of the book with the words highlighted as they are read in the student's voice. If a camera is attached to the computer, the student can create a video introduction to the performance.

Information collected about the interaction includes a recording of the student's reading of the book, answers to comprehension questions, and an assessment of his or her word recognition performance. A discussion of this information and how it might be used is included in the final section of this paper.

OVERVIEW OF SPEECH RECOGNITION TECHNOLOGY

In order to better evaluate the potential of speech recognition technology for the assessment of beginning readers, it is helpful to have at least a rudimentary understanding of how this technology works. What follows is a highly simplified description of a very complex process. This explanation is intended to highlight how speech recognition systems deal with variations among speakers and domain vocabulary that have an impact on the reliability of the technology.

Computers use speech recognition[1] technology to capture human speech and translate it to a written format. This translation process is based on two underlying models: an acoustic model containing representations of the phonemes in English and a language model representing typical sequences of words for a specific target population or domain.

The acoustic model is created by an analysis of actual human speech. A set of words is chosen that contains all the phonemes in the English language. During data collection a recording is made of a person saying each of these words. Next, an acoustic model of each phoneme is created from all words having this phoneme, e.g., the tee sound in "tree" is created from all tee sounds in all words having such a sound (e.g., toad, tree, sit).[2] To represent the natural variability in human speech, samples are collected from a large number of speakers and then blended to form a single acoustic model. These blends enable the system to recognize speakers with a wide variety of regional and second language accents.

Word forms, representations of the actual vocabulary for an application, are constructed by concatenating the appropriate phonemes from the acoustic model. Thus, a word that was not one of the words spoken during data collection should still be recognizable because a new word form can be created from phonemes in the original sample. Multiple word forms are another way to represent naturally occurring variability. For example, in some dialects the word "get" might be pronounced "git." Acoustic models can be constructed that include both "get" and "git"

[1] The terms 'speech recognition' and 'voice recognition' are sometimes used interchangeably; however, voice recognition is primarily the task of determining the identity of a speaker, rather than the content of his or her speech.

[2] This simplification could be misleading. The system doesn't actually model true phonemes. Instead, the acoustic input is divided into small time slices, and Markov modeling is used to create statistically coherent probabilistic clusters using about 200 sound templates (rather than the 50 or so phonemes).

in order to recognize this difference in pronunciation and still support a translation that accurately captures the intended meaning.

Until recently, research on speech recognition for children used standard acoustic models based on a blend of adult voices. In order to improve the accuracy of recognition for children, researchers at IBM's T.J. Watson Research Center have created a children's acoustic model based on data collected from 800 children interviewed at multiple locations across the United States.

When a user speaks a word, the speech recognition system converts the recorded word to sets of phonemic representations and searches the system's stored word forms for a "close enough" match. The smaller the set of word forms, the faster and more accurate the matching process.

Applications that must recognize large numbers of words (100,000 in IBM's ViaVoice product) also benefit from the addition of a language model to speed up the recognition process and improve accuracy. Language models are constructed by scanning millions of lines of running text and calculating the statistical probabilities of three-word transitions. For example, a speech recognition system for a business application might be based on samples from business letters, *Wall Street Journal* articles, etc. When a person dictates a business letter, the speech recognition system takes the phonemic representation of the words the person says and, instead of trying to match them to the 100,000 word forms, trims the search as it progresses by excluding paths of lower probability. The probability of the next word depends on the history of the words that have been spoken so far.

Speech recognition technology is used in two modes: command and dictation. Command mode uses only the acoustic model along with a limited set of word forms. The captured speech is typically used to trigger an action such as dialing a phone or launching a computer application. Software such as Watch Me! Read falls into this category because the passages presented to the reader represent a limited vocabulary that is known in advance.

In dictation applications, the captured speech data are transcribed and stored as a text file, edited by the user, and used like any other word-processing file. The vocabulary to be recognized for dictation is comprehensive; in order to translate speech efficiently and accurately, both an acoustic and a language model are used. (See previous section.) Dictation applications are popular with adult users; however, they are beyond the scope of this short paper, which focuses only on reading (command) applications.

GENERAL DESIGN ISSUES AND OPPORTUNITIES FOR RESEARCH

Tradeoffs are made in the design of any system. In the case of speech recognition technology, the tradeoffs are typically a balance among accuracy of recognition, speed of recognition, and ease of use. While computing power and research may eventually lessen the impact of these design decisions, they highlight important considerations for those thinking about the use of this technology with children or for assessment. Here is a partial list of issues relevant to reading applications.

Speaking Rate

In order to achieve an acceptable accuracy in recognition, some existing systems only recognize discrete speech, i.e., a user must pause slightly between words. Research on continuous speech has recently created systems that allow users to speak naturally. Without continuous speech recognition, children must learn to speak slowly and deliberately so that their reading can be reliably assessed.

The speech of children who are learning to read contains numerous pauses, repetitions, omissions, and partial words as they sound out unfamiliar words. While a human tutor may be able to follow this process and identify a child's current position in the text, this tracking is difficult for a computer system. Interfaces are needed that aid in this tracking without slowing the reader. If the computer loses its place, the feedback provided will be incorrect.

Speaker Dependency

Some speech recognition systems improve accuracy by having users train the computer system to understand individual variations in their speech. To do this, a user reads prescribed passages into the computer so that the system can personalize its acoustic model. But this training is time-consuming, and the logistics of implementing this in a classroom would be difficult. In addition, alternative training procedures would have to be created for children who cannot read.

Microphones and Headsets

Recognition is improved by the use of high-quality microphones. To facilitate proper placement and optimum distance from the user's mouth, microphones are often incorporated into a headset. These microphones are delicate, expensive instruments. Headsets have not been developed in a size appropriate for children and sturdy enough for the wear and tear of the classroom. Some research has been done with microphones strategically placed around the room, but this has not been tested in classrooms.

ASSESSMENT AND SPEECH RECOGNITION

Reading environments such as Watch Me! Read are based on oral reading as a measure of students' competence. Oral reading assessments such as running records (Clay, 1993), informal reading inventories (IRI) (Farr & Carey, 1986), and miscue analysis (Goodman, 1982) take into consideration contextual factors such as passage length and complexity and the reader's reliance on pictures and prior knowledge. They provide rich, detailed data about students' performance in the context of real reading. This information is related to students' fluency, oral reading accuracy, and decoding, as well as strategies such as rereading and self-correction.

WM!R's assessment is based on fluency and accurate word identification. The system tracks the child's progress as it compares its model of each word in the text with the word spoken by the child. The data the system provides for the teacher include a copy of the text that was read, a recording of each word spoken by the child, and an indication of whether or not the word

spoken was accepted as a match for the comparison text. If a word is accepted as a match, the system provides positive oral feedback and moves onto the next word/phrase. If the word does not appear to be a match, the system asks the student to repeat the word. If a second attempt also does not match, the system supplies the correct pronunciation. Failed matches receive feedback such as "I did not hear you." Failed matches are not labeled as errors because the technology is not capable of making this judgment with accuracy.

WM!R does not attempt to interpret data in order to assign a score or reading level. Instead it provides the detailed data to the teacher for his or her interpretation. The decision not to summarize the data is based in part on the reliability of speech recognition. While the average accuracy of recognition for the WM!R system is above 95 percent, some children are not as easily recognized as others because of characteristics of their speech, reading rate, regional or second language accents, etc. Equally important, contextual factors such as background noise and microphone adjustments can vary from session to session, even for the same reader. Summarizing data across one or more sessions can mask this variation and make the assessment appear more reliable.

A second reason for not summarizing the data stems from the belief that fluent reading sometimes means that a reader makes meaningful substitutions. Goodman (1982) called such substitutions miscues rather than errors because meaning may be maintained even when the text is not read as written.[3] Thus, valid inferences about whether or not a mismatch represents an error or an appropriate substitution need to be made by the teacher.

From a practical point of view, the amount of data produced by WM!R can be overwhelming to monitor on a regular basis. Listening to students' oral reading, whether live or recorded, requires a great amount of time. The designers of WM!R are currently exploring the generation of alert messages to the teacher to identify children who might be having difficulty. These alerts are triggered by a "matching" rate that falls below a prescribed threshold. Additionally, the system identifies potential problem words and creates a list of practice words for students to study. These reporting strategies allow efficient use of the feedback while leaving the final interpretation of the data up to the teacher.

The data provided by WM!R also offer interesting instructional possibilities. First, students can review their own reading performance by listening to their recorded voice as text is highlighted word-by-word on the screen. This can encourage and enable self-assessment. Second, student and teacher could review the recorded reading together and discuss appropriate and inappropriate substitutions and other strategies.[4] These types of reflective activities would be impossible without the assistance of technology to create a representation pairing each word in the text with the child's attempt to read it.

[3] It is possible to represent common miscues as additional word forms; however, it is not possible to include all substitutions.

[4] Thanks to my colleague, David Schwarzer, for this suggestion.

EVALUATION OF WM!R

Preliminary studies suggest that WM!R can enhance literacy instruction by supporting independent reading practice (Williams, 2000). In these studies, first graders using WM!R as a part of their regular instruction read similar stories without assistance and with WM!R. While using WM!R, all the students were significantly more engaged in the reading task, even those who did not need the support that it provided. When they were asked to reread or retell the stories, their word recognition and retelling scores were significantly higher after they had used WM!R.

Interviews with classroom teachers provided further insight into the benefits of WM!R. Teachers reported that the main benefit for their students was regular reading practice with individualized feedback. Other benefits of WM!R practice varied according to the reading level of the student and the way the student used the software. For beginning readers who had not yet learned to sound out words for themselves, the software greatly increased the likelihood that they could get prompt help with words they did not recognize. The software also helped new readers mark the word they were currently reading and track their progress across the page.

For more advanced readers, teachers reported that the software provided structure. Without WM!R, these students were likely to rush through books without getting the details of what they were reading or taking time to monitor their own comprehension. The pacing provided by the software helped them attend to details, and the comprehension questions at the end of the page encouraged them to reflect on the book's meaning. When asked about their advanced students, all the teachers said that these students were very interested in using the program and were benefiting from it.

Special needs students also benefited from WM!R. One teacher described a hearing-impaired student who was not fitted for a hearing aid until spring of the school year. WM!R was his best opportunity for getting feedback on his pronunciation because the volume could be adjusted so that he could hear well. Students who exhibited symptoms of attention deficit disorder were more engaged in reading practice with WM!R than with reading on paper.

All teachers mentioned Limited English Proficiency students as benefiting from the program. In the Houston Independent School District, classes for bilingual students are conducted mostly in Spanish, and Spanish is the primary language in their homes and communities. Thus, these students have few opportunities to work on their developing English skills, and they can be very insecure about trying to say the words out loud. Using WM!R allows these students to hear their own voice and compare their pronunciation with that of the system.

It is interesting that not a single teacher mentioned problems with speech recognition as a barrier to use of the software by their students. When teachers were asked about failures (false positives or false negatives) in the speech recognition, they indicated that the positive aspects of WM!R far outweighed occasional problems with recognition.

CONCLUSION AND CAUTIONS

Preliminary research indicates that speech recognition technology has developed to the point that it is useful as a scaffold for early reading because of the feedback it provides and the engagement in reading that it encourages. The potential benefits of WM!R for students depend on many factors, such as the design of the technology itself, the interaction of the technology with students and classroom instruction (Williams, Nix, & Fairweather, 2000), etc. Tools such as WM!R provide reading practice but not reading instruction. They monitor word recognition but not comprehension. Therefore, the teacher must ensure that students have the instruction they need in order to make the best use possible of time spent with WM!R.

The benefits of frequent reading also depend on the availability of extensive reading material at appropriate reading levels. Although WM!R includes an authoring tool to enter new books into the system easily, getting permission to use trade books is almost impossible. Writing and illustrating engaging books at appropriate levels require skills and time that many teachers do not have.

Speech recognition technology requires top-of-the-line hardware. Most school systems that purchase powerful hardware focus on supplying older students. Thus, it may be difficult to get enough high-performance hardware into 1st and 2nd grade classrooms to make a difference in practice time.

It is important to be cautious about relying on data provided by speech recognition technology as a summative assessment. These data must be considered in context to be useful for making instructional decisions.

Despite these issues, the promise of this technology is very real and very exciting: It can be a valuable aid in supporting practice for beginning readers and providing assessment information for their teachers.

REFERENCES

Anderson, J.R. (1995). *Learning and memory: An integrated approach*. New York: Wiley.

Byrnes, J.P. (2000). Using instructional time effectively. In L. Baker, M.J. Dreher, J.T. Guthrie (Eds.), Engaging young readers: Promoting achievement and motivation (pp. 188-208). New York: Guilford Press.

Cipielewski, J., & Stanovich, K.E. (1992). Predicting growth in reading ability from children's exposure to print. *Journal of Experimental Child Psychology, 54*, 74-89.

Clay, M. (1993). *An observation survey of early literacy achievement*. Portsmouth, NH: Heinemann.

Ericsson, & Smith, J. (1991). *Toward a general theory of expertise: Prospects and limits*. Cambridge, UK: Cambridge University Press.

Farr, R., & Carey, R. (1986). *Reading: What can be measured?* Newark, DE: International Reading Association.

Goodman, K.S. (1982). A linguistic study of cues and miscues in reading. In F.V. Gollasch (Ed.), Language and literacy—The selected writings of Kenneth S. Goodman: Vol. 1. Process, theory, and research (pp. 115-120). Boston: Routledge & Kegan Paul.

Greany, V., & Hegarty, M. (1987). Correlates of leisure time reading. *Reading Research Quarterly, 15*, 337-357.

Guthrie, J.T. (1980). Time in reading programs. *The Reading Teacher, 34*, 500-502.

Juel, C. (1996). What makes literacy tutoring effective? *Reading Research Quarterly, 31*, 268-289.

Mostow, J., & Aist, G. (1999). Giving help and praise in a reading tutor with imperfect listening—because automated speech recognition means never being able to say you're certain. *CALICO Journal, 16*(3), 407-424.

Nix, D., Fairweather, P., & Adams, W. (1998). *Speech recognition, children and reading. Human factors in computing systems*. New York: Association for Computing Machinery.

Perfetti, C.A. (1985). *Reading ability*. New York: Oxford University Press.

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-407.

Stanovich, K.E., & Cunningham, A.E. (1992). Studying the consequences of literacy within a literate society. The cognitive correlates of print exposure. *Memory and Cognition, 21*, 51-68.

Williams, S.M. (2000). What children learn from using Watch Me! Read. A report prepared for IBM.

Williams, S.M., Nix, D., & Fairweather, P. (2000). Using speech recognition technology to enhance literacy instruction for emerging readers. Proceedings of the International Conference for the Learning Sciences. Ann Arbor, MI.

Chapter 5
Cognitive Tutor Algebra I: Adaptive Student Modeling in Widespread Classroom Use

Albert Corbett
Human-Computer Interaction Institute
Carnegie Mellon University

Individual human tutoring is perhaps the oldest form of instruction, and countless millennia since its introduction, it remains the most effective and most expensive form of instruction. Studies show that students working with the best human tutors attain achievement levels that are two standard deviations higher than students in conventional classroom instruction (Bloom, 1984; Cohen, Kulik, & Kulik, 1982). Over the past 15 years the Carnegie Mellon Pittsburgh Area Cognitive Tutor (PACT) Center has been developing an educational technology called *cognitive tutors* that provides some of the advantages of individual human tutoring. Cognitive tutors are rich problem-solving environments. Each cognitive tutor is constructed around a cognitive model of the problem-solving knowledge students are acquiring. The cognitive model is employed to provide two types of adaptive student support. In *model tracing* the cognitive model is used to interpret and respond to each of the student's problem-solving actions. In *knowledge tracing*, the tutor monitors the student's growing problem-solving knowledge and individualizes the problem sequencing accordingly.

This paper briefly describes three related topics. First, it introduces the Algebra I Cognitive Tutor and describes the project that brought this successful educational technology out of the research lab and into widespread classroom use. Next, the paper describes the cognitive theory underlying cognitive tutors and reports studies that assess the validity of knowledge tracing and its effectiveness in individualizing each student's problem-solving sequence. The paper concludes by suggesting future research directions.

Figure 5-1 displays the Algebra I Cognitive Tutor near the completion of a problem. The problem situation is presented in the scenario window in the upper left corner of the screen:

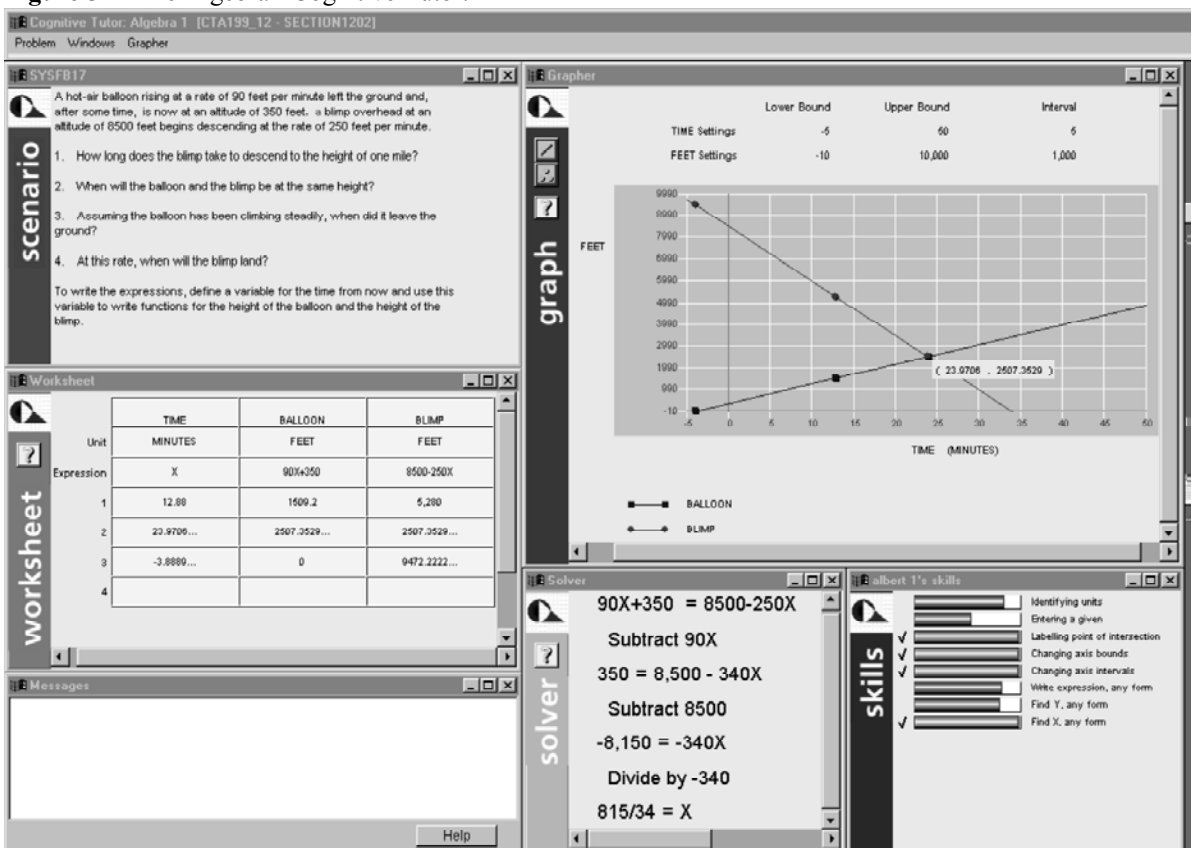
> *A hot-air balloon rising at a rate of 90 feet per minute left the ground and, after some time, is now at an altitude of 350 feet. A blimp overhead at an altitude of 8500 begins descending at the rate of 250 feet per minute.*

Four questions are posed for the student to answer:

1) How long does the blimp take to descend to the height of one mile?
2) When will the balloon and the blimp be at the same height?
3) Assuming the balloon has been climbing steadily, when did it leave the ground?
4) At this rate, when will the blimp land?

The student answers the questions by filling in the worksheet immediately below the scenario window. The cells in the worksheet are blank initially. The student analyzes the problem situation, identifies the relevant quantities that are varying (in this situation, time and height of the airships), and labels the worksheet columns accordingly ("TIME," "BALLOON," and "BLIMP"). The student enters the appropriate units for measuring these quantities in the second row of the table and enters a symbolic model relating the three quantities in the third row. In the figure, the student has represented the quantity of time with the variable *X*, related the height of the balloon to time with the algebraic expression $90X + 350$, and similarly related the blimp's height to time with the expression $8500 - 250X$. Early in the curriculum, the student enters algebraic expressions near the end of the problem, and the individual questions are intended to help scaffold this algebraic modeling. By the time the student has reached the linear systems unit represented by the current problem, he or she tends to enter the algebraic expressions early in the problem. The emphasis in these problems is on using the expressions as problem-solving tools, both in symbol manipulation and to automatically generate values in the worksheet (as described below).

Figure 5-1 The Algebra I Cognitive Tutor.



SOURCE: Carnegie Learning, Inc., 2000

The student answers the questions by filling in the corresponding rows in the worksheet. To answer the first question, "How long does the blimp take to descend to the height of one mile?" the student needs to perform a unit conversion on the given value, "one mile," and enter 5280 in the question-1 cell of the BLIMP column, immediately below the formula. To compute

the solution, the student can set up the equation 5280 = 8500 - 250*X* in the solver window in the lower center of the screen, solve it in the window, and type the answer, 12.88, into the question-1 cell of the TIME column. Or the student can type an arithmetic expression that unwinds the equation, (5280-8500)/ -250, directly into the question-1 cell of the TIME column, where it will be converted to 12.88. Once the value of X has been typed in this cell, the worksheet automatically generates the corresponding height of the balloon, 1509.2 feet at 12.88 minutes, employing the algebraic expression for balloon height typed by the student.

In completing the problem, the student also graphs the two linear functions in the graph window in the upper right corner of Figure 5-1. The student labels the axes, adjusts the upper and lower bounds on the axes, and sets the scales so that the data points in the table can be displayed. Note that in question 2, "When will the balloon and the blimp be at the same height?" the student is asked to solve for the intersection of the two functions, and in questions 3 and 4, the student is asked to find the x-intercept of the two functions. The student can answer these questions by finding the relevant points on the graph or by setting up and solving equations in the solver window. In Figure 5-1, the student has set up an equation to solve for the intersection of the two functions, $90X + 350 = 8500 - 250X$, and proceeded to solve the equation by isolating *X*.

ADAPTIVE STUDENT MODELING

A central claim of *Knowing What Students Know*, the recent National Research Council report on educational assessment (NRC, 2001), is that three essential pillars support scientific assessment: a general model of student cognition, tasks in which to observe student behavior, and a method for drawing inferences about student knowledge from students' behaviors. Cognitive tutors embody this framework. Each cognitive tutor is constructed around a cognitive model of the knowledge students are acquiring. As a student performs problem-solving tasks such as that displayed in Figure 5-1, the cognitive model is employed to interpret the student's behavior, and simple learning and performance assumptions are incorporated to draw inferences about the student's growing knowledge.

Model Tracing

In *model tracing*, the underlying cognitive model is employed to interpret each student action and follow the student's individual solution path through the problem space, providing just the support necessary for the student to complete the problem successfully. The cognitive model is run forward step-by-step along with the student, and each student action is matched to the actions that the model can generate in the same context. As with effective human tutors, the cognitive tutor's feedback is brief and focused on the student's problem-solving context. If the student's action is correct, it is simply accepted. If the student makes a mistake, it is rejected and flagged (in red font). If the student's mistake matches a common misconception, the tutor also displays a brief just-in-time error message (in the window in the lower left corner of Figure 5-1). The tutor does not provide detailed explanations of mistakes, but instead allows the student to reflect on mistakes. Finally, the student can ask for problem-solving advice at any step. The tutor generally provides three levels of advice. The first level advises on a goal to be

accomplished, the second level provides general advice on achieving the goal, and the third level provides concrete advice on how to solve the goal in the current context.

Knowledge Tracing

The cognitive model is also employed to monitor the student's growing knowledge in problem solving, in a process we call *knowledge tracing*. At each opportunity for the student to employ a cognitive rule in problem solving, simple learning and performance assumptions are employed to calculate an updated estimate of the probability that the student has learned the rule (Corbett & Anderson, 1995). These probability estimates are displayed in the skillmeter in the lower right corner of Figure 5-1. Each bar represents a rule, and the shading reflects the probability that the student knows the rule. As advocated by NRC (2001), the goal of knowledge tracing is to improve learning outcomes. It is employed to implement *cognitive mastery*. Within each curriculum section, successive problems are selected to provide the student the greatest opportunity to apply rules that he or she has not yet mastered. The tutor continues presenting problems in a section until the student has "mastered" each of the applicable rules in the curriculum section. (Mastery is indicated by a checkmark in the skill meter).

TRANSFORMING EDUCATIONAL PRACTICE

I believe that cognitive tutors for mathematics are the first intelligent tutoring systems that are beginning to have a widespread impact on educational practice. In the 2001-2002 school year, Cognitive Tutor Algebra and Geometry courses are in use at about 700 sites and by more than 125,000 students in 38 states. This includes urban, suburban, and rural middle and high schools, both public and private. This success in moving from the research lab into widespread classroom use depends on several factors, including project design, research-based development, demonstrated impact, and classroom support.

Project Design

Several project design features were essential to the success of the dissemination project (Corbett, Koedinger, & Hadley, 2001).

Opportunity: Targeting a National Need

National assessments such as the NAEP and international assessments such as the TIMSS have raised awareness of the need to improve mathematics education. Cities and states have increasingly mandated that all students need to master academic mathematics, and virtually every state has defined high-stakes academic mathematics assessments that are employed to evaluate schools and/or govern student graduation. For more than a decade, the National Council of Teachers of Mathematics (NCTM, 1989) has been recommending that academic mathematics for all students should place a greater emphasis on problem solving, reasoning among different mathematical representations, and communication of mathematical results. As a result of these trends, school districts actively look for, and are open to trying, new solutions to mathematics education, and Cognitive Tutor Algebra I aligns well with the NCTM-recommended objectives.

Integrating Technology into a Comprehensive Solution

Teachers face a major "usability" challenge when they are trying to integrate educational technology into a course. It may be difficult to align course curriculum objectives and technology curriculum objectives and to make time for the technology activities. In our cognitive tutor mathematics project, cognitive tutors are fully integrated into yearlong courses. We develop both the paper text that is employed in 60 percent of class periods and the cognitive tutor that is employed in 40 percent of class periods. This coordinated development helps ensure that the problem-solving activities presented two days a week by the cognitive tutors address and develop the same curriculum objectives that students explore in small-group problem solving and whole-class instruction the other three days a week.

Interdisciplinary Research Team

The research team is a collaboration of cognitive psychologists, computer scientists, and practicing classroom teachers throughout the process of developing, piloting, evaluating, and disseminating a cognitive tutor course.

Research-Based Development

Cognitive tutor design is guided by multiple research strands, including cognitive psychology of student thinking (Heffernan & Koedinger, 1997; 1998), research in student learning (Koedinger & Anderson, 1998), and research in effective interactive learning support (Aleven, Koedinger, & Cross, 1999; Corbett & Trask, 2000; Corbett & Anderson, 2001). Formative evaluations of tutor lessons are employed to guide iterative design improvements, including studies of learning rate, validity of the underlying student model, and pre-test to post-test learning gains (Corbett, McLaughlin, & Scarpinatto, 2000).

Demonstrated Impact

Cognitive tutor courses have a demonstrable impact on the classroom, student motivation, and student achievement.

Substantial Achievement Gains

Beginning with our two earliest cognitive tutors, the ACT Programming Tutor (APT) and the Geometry Proof Tutor (GPT), cognitive tutor technology has an established history of yielding substantial achievement gains compared to conventional learning environments (Anderson, Corbett, Koedinger & Pelletier, 1995). College students working with APT completed a problem set three times faster and scored 25 percent higher on tests than students completing the same problems in a conventional programming environment. High school students in geometry classes that employed GPT for in-class problem solving scored about a letter grade higher on a subsequent test than students in other geometry classes who engaged in conventional classroom problem-solving activities. Koedinger, Anderson, Hadley, and Mark (1997) demonstrated that the Cognitive Tutor Algebra I course yields similar achievement gains.

High school students in the Cognitive Tutor Algebra I course scored about 100 percent higher on tests of algebra problem solving and reasoning among multiple representations, and about 15 percent higher on standardized assessments than similar students in traditional Algebra I classes.

Student-Centered Learning-by-Doing

In cognitive tutor courses, students actively learn-by-doing, both in the cognitive tutor lab and in small-group problem solving in the classroom. Schofield (1995) formally documented the impact of cognitive tutors on the student-teacher relationship in a field study of the Geometry Proof Tutor in the mid-1980s. This was the first cognitive tutor deployed in high school classrooms. She found that teachers in the cognitive tutor lab serve as collaborators in learning. Teachers shift their attention to the students who need more help, and they can engage in more extended interactions with an individual student while other students in the class make substantial progress as they work with the cognitive tutor.

Increased Student Motivation

In the same study, Schofield (1995) documented that students are highly motivated and highly engaged in mathematics in the cognitive tutor lab. Teachers are excited about student attitudes and about engaging students in individualized discussions of mathematics. Letters from some of our Cognitive Tutor Algebra I teachers include such comments as "Gone are the phrases 'this is too hard—I can't do this,' instead I hear 'how do you do this? why is this wrong?'" and "Students now love coming to class. They also spend time during their study halls, lunch, before and after school working on the computers. Self-confidence in mathematics is at an all time high."

Classroom Support

When a school adopts a cognitive tutor mathematics course, we also provide comprehensive classroom support that includes pre-service and in-service professional development, both on the cognitive tutor technology and on small-group problem solving in class. We also provide software installation, hotline support (both email and telephone) for pedagogical questions and technical problems, and email user groups and teacher focus group meetings.

U.S. Department of Education Exemplary Curriculum

In 1999 Cognitive Tutor Algebra I was designated an "exemplary" curriculum by the U.S. Department of Education. Sixty-one K-12 mathematics curricula were reviewed on three criteria: the program's quality, usefulness to others, and educational significance. Of these 61 curricula, five were awarded the highest, "exemplary," designation.

Comments on Continued Scaling Up

Perhaps two key issues arise in considering the future growth in impact of cognitive tutor courses: the cost of developing new cognitive tutor courses and the need to provide high-quality site support.

Cognitive Tutor Course Development

Our best current estimate of the development cost for a cognitive tutor course comes from our current Cognitive Tutor Middle School Mathematics Project in which we are developing three full-year courses. In this project, just over 100 hours of effort yields one hour of classroom activity. It should be emphasized that this estimate includes all aspects of design, development, piloting support, and evaluation including: developing the cognitive task analysis that underlies text and cognitive tutor design: writing the text: programming the cognitive model; programming the tutor interface; writing and coding the tutor problems; installing and maintaining the tutor; conducting teacher training; and designing, conducting, and analyzing formative and summative evaluations.

We believe that this cost level is already economically competitive, given the substantial impact of cognitive tutor courses on achievement outcomes and the demonstrable impact on student motivation. Evaluations of the mathematics and programming cognitive tutors indicate that model tracing alone can yield a one-standard deviation effect size, which is about half the benefit of the best human tutors (Anderson et al., 1995). Our estimations, based on evaluations of knowledge tracing and cognitive mastery in the programming tutor, suggest that this method of dynamic assessment and curriculum individualization can add as much as another half-standard deviation effect size (Corbett, 2001). We believe that the cost/benefit ratio will continue to improve as new research leads to improvements in tutor effectiveness. Perhaps the more important limiting factor in cognitive tutor course development is not the cost, but the availability of trained professionals to conduct cognitive task analyses and develop cognitive models.

Site Support

The single greatest challenge in site support is teacher training. As the Algebra and Geometry Cognitive Tutors become more robust, technical support is not a problematic issue. Teachers and students need little training in the use of the cognitive tutor software. Instead, the greatest need is to help students become not just “active problem solvers” but “active learners,” who view problem solving not as an end in itself, but as a vehicle for learning. Teachers need professional development to help students make use of the learning opportunities that arise in problem solving, not just in the cognitive tutor lab but also in small-group problem-solving activities during other class periods. As the deployment of cognitive tutor mathematics courses has grown, we have relied on a growing number of experienced cognitive tutor mathematics teachers to offer professional development. But this need for effective teacher development is not limited to our project. Research is needed to define effective teaching methods that support active student learning, and this knowledge needs to become integrated into pre-service teacher education.

COGNITIVE THEORY AND DYNAMIC ASSESSMENT

Cognitive tutors are grounded in cognitive psychology. The cognitive model underlying each tutor reflects the ACT-R theory of skill knowledge (Anderson, 1993). ACT-R assumes a fundamental distinction between declarative knowledge and procedural knowledge. Declarative

knowledge is factual or experiential and goal-independent, while procedural knowledge is goal-related. For example, the following sentence and example in an algebra text would be encoded declaratively:

> *If the same amount is subtracted from the quantities on both sides of an equation, the resulting quantities are equal.*
> *For example, if we have the equation X + 4 = 20, then we can subtract 4 from both sides of the equation and the two resulting expressions X and 16 are equal, X = 16.*

ACT-R assumes that skill knowledge is initially encoded in declarative form when the student reads or listens to a lecture. Initially the student employs general problem-solving rules to apply this declarative knowledge in problem solving, but with practice, domain-specific procedural knowledge is formed. ACT-R assumes that procedural knowledge can be represented as production rules—if-then rules that associate problem-solving goals and problem states with actions and consequent state changes. The following production rule may emerge when the student applies the declarative knowledge above to equation-solving problems:

> *If the goal is to solve an equation of the form X + a = b for the variable X,*
> *Then subtract a from both sides of the equation.*

Substantial cognitive tutor research has validated production rules as the unit of procedural knowledge (Anderson, Conrad, & Corbett, 1989; Anderson, 1993).

Evaluating Knowledge Tracing and Cognitive Mastery

As the student works, the tutor estimates the probability that he or she has learned each of the rules in the cognitive model. The tutor makes some simple learning and performance assumptions for this purpose (Corbett & Anderson, 1995). At each opportunity to apply a problem-solving rule, the tutor

- uses a Bayesian computational procedure to update the probability that the student *already knew* the rule, given the evidence provided by the student's response (whether the student's action is correct or incorrect), and
- adds to this updated estimate the probability that the student *learns* the rule at this opportunity if it has not already been learned.

The goal of knowledge tracing is to promote efficient learning and enable cognitive mastery of the problem-solving knowledge introduced in the curriculum. Within each curriculum section, the tutor presents an individualized set of problems to each student, until the student has "mastered" the rule (typically defined as a 0.95 probability of knowing the rule).

Validating Knowledge Tracing: Predicting Tutor Performance

The same learning and performance assumptions that allow us to infer the student's knowledge state from his or her performance also allow us to predict student performance from the student's hypothesized knowledge state. A series of studies validated knowledge tracing in

the ACT Programming Tutor by predicting student problem-solving performance, both in the tutor environment and in subsequent tests (Corbett & Anderson, 1995; Corbett, Anderson, & O'Brien, 1995). Figure 5-2 displays the mean learning curve, both actual and predicted, for a set of problem-solving rules in an early section of the ACT Programming Tutor. The first point in the empirical curve indicates that students had an average error rate of 42 percent in applying each of the rules in the set for the first time. Average error rate declined to under 30 percent across the second application of all the rules, and it continued to decline monotonically over successive applications of the rules. As can be seen, the knowledge-tracing model very accurately predicts students' mean production-application error rate in solving tutor problems.

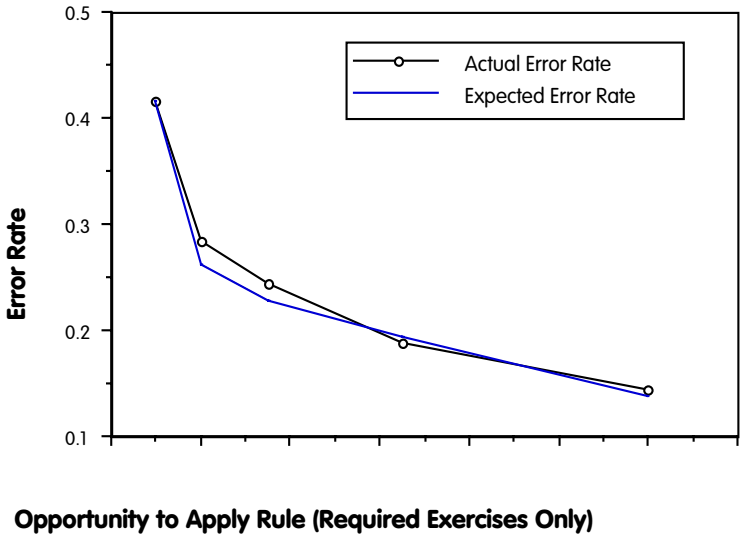


Figure 5-2 Actual error rate and predicted error rate for successive applications of problem-solving rules in the ACT Programming Tutor. SOURCE: Corbett, Anderson, & O'Brien, 1995, p. 26.

Validating Knowledge Tracing: Individual Differences in Post-test Performance

The more important issue is whether knowledge tracing accurately predicts students' performance when they are working on their own. A sequence of studies (Corbett & Anderson, 1995) examined the accuracy of the knowledge-tracing model in predicting students' post-test performance after they had completed work in the ACT Programming Tutor. Figure 5-3 displays quiz results for 25 students in the final study of the series. The figure displays each student's actual post-test accuracy (proportion of problems completed correctly), plotted as a function of the knowledge-tracing model's accuracy prediction for the student. As can be seen, the model predicted individual differences in test performance quite accurately. The model predicted that this group of students would average 86 percent correct, and they actually averaged 81 percent. The correlation of actual and expected performance across the 25 students is 0.66.

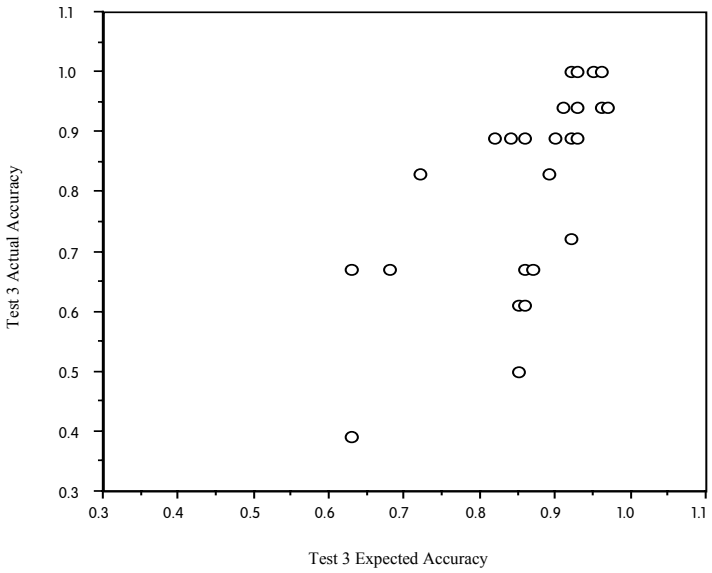


Figure 5-3 Student post-test accuracy plotted as a function of accuracy predicted by the ACT Programming Tutor knowledge-tracing model. SOURCE: Corbett & Anderson, 1995, p. 274.

Cognitive Mastery Effectiveness

A recent study examined the efficiency of cognitive mastery learning (Corbett, 2001). In this study 10 students in a fixed-curriculum condition worked through a set of 30 ACT Programming Tutor problems. Twelve students in a cognitive mastery condition completed the fixed set of 30 problems and an additional, individually tailored sequence of problems as needed to reach mastery. On a subsequent test, students in the mastery condition averaged 85 percent correct on the test, while students in the fixed-curriculum condition averaged 68 percent correct. This difference is reliable, $t(20) = 2.31$, $p < .05$. Of the cognitive mastery students, 67 percent reached a high mastery criterion on the test (90 percent correct), while only 10 percent of students in the fixed-curriculum condition reached this high level of performance. Students in the cognitive-mastery condition completed an average of 42 tutor problems—40 percent more problems than students in the fixed-curriculum condition—and they only required 15 percent more time to do so. This investment of 15 percent more time yielded a high payoff in achievement gains.

Future Research

Three lines of research can be identified to enhance the educational effectiveness and broaden the impact of cognitive tutors in classrooms around the country:

- We need to develop cognitive tutor interventions that will help students become more active learners and develop a deeper, conceptual knowledge of the problem-solving domain.
- We also need to better understand how teacher interventions can help students become more active learners.
- We need to develop authoring systems that can make cognitive tutor development faster.

Students working with cognitive tutors are active problem solvers. The principal strength of cognitive tutors is that they expose learning opportunities in detail. They reveal students' missing knowledge and misconceptions step-by-step and afford students the opportunity to construct knowledge. However, the tutors' feedback and advice capabilities are limited. Both take the form of short written messages, with multiple levels of help available upon request at each problem-solving step. Studies show that students do not always make effective use of the assistance available. Eye-tracking studies of students working with the Algebra I cognitive tutor show that they often do not read or even notice the error feedback message (Gluck, 1999). Other studies with the Geometry Cognitive Tutor show that students often make poor use of the help that is available (Aleven & Koedinger, 2000). The knowledge-tracing validation research in the ACT Programming Tutor reveals a related point. The knowledge-tracing model consistently overestimates students' test performance by a small amount, about 10 percent. Studies suggest that this happens because some students are learning some shallow rules in the tutor that do not transfer to the test (Corbett & Knapp, 1996; Corbett & Bhatnagar, 1997).

Cognitive tutors are already at least half as effective as the best human tutors and two or three times as successful as conventional computer-based instruction (Corbett, 2001). They can become far more effective if they provide scaffolding to help students become not just active problem solvers, but active learners when learning opportunities are exposed. We have already had some success in engendering deeper learning with graphical feedback (Corbett & Trask, 2000) and student explanations of problem-solving steps (Aleven & Koedinger, in press), but we need to develop a more general framework for understanding effective tutorial scaffolding for student knowledge. Recent research is continuing to develop our understanding of effective human tutor tactics (e.g., Chi, Siler, Jeong, Yamauchi & Hausmann, 2001). We need to integrate these results into a general theory and to implement effective scaffolding in cognitive tutors.

Similarly, we need to better understand how the teacher in a cognitive tutor class can effectively scaffold active student learning. In the cognitive tutor lab, the teacher has the opportunity to interact with individual student tutors on an extended basis (Schofield, 1995), and there is preliminary evidence that the benefits of cognitive tutors can depend on the teacher's activities in the lab (Koedinger & Anderson, 1993). Research on effective human tutor tactics is relevant, but the classroom teacher needs some additional skills: recognizing when a "teachable moment" arises for one student in a classroom of 20-30 students and being able to jump into the student's problem-solving context to provide effective scaffolding. We also need to understand how teachers can best support small-group problem solving in cognitive tutor courses and integrate these group-paced classroom activities with the individually paced cognitive tutor activities. And we need to develop effective professional development based on this research in effective teacher strategies.

Finally, to broaden the impact of cognitive tutor technology, we need to develop authoring tools that can speed its design and implementation. These tools need to make cognitive tutor development more accessible to domain experts who do not have computer science or cognitive science backgrounds. At minimum, these tools should facilitate curriculum (problem situation) authoring. At best, these can be intelligent tools that make cognitive modeling more accessible to domain experts. In conjunction with this tool development, we need to begin designing cognitive tutors for other domains to examine how well the lessons

learned in mathematics generalize. Middle school science would be an opportune domain, both for research and for purposes of educational impact.

In 1984 Bloom issued a challenge to develop educational interventions that are as effective as human tutors, but affordable enough for widespread dissemination. We believe that the research outlined here can make it possible to meet and even exceed Bloom's goal.

REFERENCES

Aleven, V., & Koedinger, K.R. (in press). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*.

Aleven, V., & Koedinger, K.R. (2000). Limitations of student control: Do students know when they need help? In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent tutoring systems: Fifth International Conference, ITS 2000* (pp. 292-303). New York: Springer.

Aleven, V., Koedinger, K.R., and Cross, K. (1999). Tutoring answer explanation fosters learning with understanding. In S. Lajoie & M. Vivet (Eds.), *Proceedings of the Artificial Intelligence and Education 1999 Conference* (pp. 199-206). Washington, DC: IOS Press.

Anderson, J.R. (1993). *Rules of the mind*. Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, J.R., Conrad, F., & Corbett, A.T. (1989). Skill acquisition and the LISP Tutor. *Cognitive Science, 13*, 467-505.

Anderson, J.R., Corbett, A.T., Koedinger, K.R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, *4*, 167-207.

Bloom, B.S. (1984). The 2-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*, 4-15.

Chi, M.T.H., Siler, S.A., Jeong, H., Yamauchi, T., and Hausmann, R.G. (2001). Learning from human tutoring. *Cognitive Science*, *25*, 471-533.

Cohen, P.A., Kulik, J.A., & Kulik, C.C. (1982). Educational outcomes of tutoring: A meta analysis of findings. *American Educational Research Journal*, *19*, 237-248.

Corbett, A.T. (2001). Cognitive computer tutors: Solving the two-sigma problem. *User Modeling: Proceedings of the Eighth International Conference, UM 2001*, 137-147.

Corbett, A.T., & Anderson, J.R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253-278.

Corbett, A.T, & Anderson, J.R. (2001). Locus of feedback control in computer-based tutoring: Impact of learning rate, achievement and attitudes. In *Proceedings of ACM CHI'2001 Conference on Human Factors in Computing Systems* (pp. 245-252).

Corbett, A.T., Anderson, J.R., & O'Brien, A.T. (1995). Student modeling in the ACT Programming Tutor. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 19-41). Mahwah, NJ: Lawrence Erlbaum Associates.

Corbett, A.T., & Bhatnagar, A. (1997). Student modeling in the ACT Programming Tutor: Adjusting a procedural learning model with declarative knowledge. *Proceedings of the Sixth International Conference on User Modeling*. New York: Springer-Verlag Wein.

Corbett, A.T., & Knapp, S. (1996). Plan scaffolding: Impact on the process and product of learning. In C. Frasson, G. Gauthier, & A. Lesgold (Eds.), *Intelligent tutoring systems: Third international conference , ITS '96.* New York: Springer.

Corbett, A.T., Koedinger, K.R., & Hadley, W. S. (2001). Cognitive Tutors: From the research classroom to all classrooms. In P. Goodman (Ed.), *Technology enhanced learning: Opportunities for change* (pp. 235-263). Mahwah, NJ: Lawrence Erlbaum Associates.

Corbett, A.T., McLaughlin, M., & Scarpinatto, K.C. (2000). Modeling student knowledge: Cognitive Tutors in High School and College. *User Modeling and User-Adapted Interaction*, *10*, 81-108.

Corbett, A.T., & Trask, H. (2000). Instructional interventions in computer-based tutoring: Differential impact on learning time and accuracy.

Gluck, K. (1999). Eye movements and algebra tutoring. Doctoral dissertation. Psychology Department, Carnegie Mellon University

Heffernan, N.T., & Koedinger, K.R. (1997). The composition effect in symbolizing: The role of symbol production vs. text comprehension. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 307-312). Mahwah, NJ: Lawrence Erlbaum Associates.

Heffernan, N.T., & Koedinger, K.R. (1998). A developmental model for algebra symbolization: The results of a difficulty factors assessment. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 484-489). Mahwah, NJ: Lawrence Erlbaum Associates.

Koedinger, K.R., & Anderson, J.R. (1993). Effective use of intelligent software in high school math classrooms. In P. Brna, S. Ohlsson, & H. Pain (Eds.), *Proceedings of AIED 93 World Conference on Artificial Intelligence in Education* (pp. 241-248). Charlottesville, VA: Association for the Advancement of Computing in Education.

Koedinger, K.R., & Anderson, J.R. (1998). Illustrating principled design: The early evolution of a cognitive tutor for algebra symbolization. *Interactive Learning Environments*, *5*, 161-180.

Koedinger, K.R., Anderson, J.R., Hadley, W.H., & Mark, M.A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*, 30-43.

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Research Council. (2001*). Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser (Eds). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Schofield, J.W. (1995). *Computers and classroom culture*. Cambridge, England: Cambridge University Press.

Chapter 6
How Computer-Based Technology Can Disrupt the Technology of Testing and Assessment

Michael Russell
National Board on Educational Testing and Public Policy
Center for the Study of Testing, Evaluation, and Educational Policy
Boston College

Over the past decade, both the presence of and the access to computer-based technology in K-12 schools have increased rapidly. In turn, computer-based technologies are changing the tools with which teachers teach and students learn. As computer-based tools continue to evolve and become more prevalent in K-12 classrooms, their use provides challenges to and opportunities for assessment. In some cases, the challenges result from pressure applied on testing programs as a result of classroom uses of technology. In other cases, the technology itself can increase the efficiency of testing. And in still other cases, computer-based technology provides opportunities to radically transform testing and assessment. In this paper, I briefly discuss how classroom uses of technology and the efficiency it affords impact testing. The bulk of this paper, however, focuses on disruptive applications of computer-based technology to educational assessment.

PRESSURE FROM THE CLASSROOM UP

As the use of computer-based technologies gradually becomes a regular component of classroom teaching and learning, the tools with which students solve problems and produce work are evolving from paper-and-pencil-based to computer-based. As students become increasingly accustomed to learning and working with these computer-based tools, a misalignment develops between the tools students regularly use to learn and the tools they are allowed to use while their achievement is tested. In turn, students, teachers, and educational systems begin to pressure testing programs to allow the use of these instructional tools during testing. For example, students are increasingly using calculators during mathematics instruction and word-processors for writing.

During the mid-1990s, there was much debate over whether students should be provided access to calculators during testing (Dion et al., 2000; Dunham & Dick, 1994; Kenelly, 1990). When the debate over calculators first arose, several concerns were raised. These concerns related to:

- equity issues: Do all students have access to calculators, both in the classroom and during testing?
- standardization: Should all students use the same type of calculator during testing?

- construct validity: Is the construct measured the same when students use calculators and when they do not?

While test developers and those who use test results are still concerned about these issues, the widespread and regular use of calculators in the classroom has led many testing programs to allow students to use calculators on items that do not specifically measure students' arithmetic skills.

Similarly, the widespread use of word processors for developing and refining written work in K-12 schools poses similar challenges to current paper-and-pencil testing practices. As a series of studies has shown, the writing ability of students accustomed to writing with computers is seriously underestimated by paper-and-pencil tests of writing (Russell & Haney, 1997; Russell, 1999; Russell & Plati, 2001). In a series of randomized experiments, this mode of administration effect has ranged from an effect size of about .4 to just over 1.0. In practical terms, the mode of administration found in the first study indicated that when students accustomed to writing on computer were forced to use paper and pencil, only 30 percent performed at a "passing" level; when they wrote on computer, 67 percent "passed." In a second study, the difference in performance on paper versus on computer for students who could keyboard approximately 20 words a minute was larger than the amount students' scores typically change between grade 7 and grade 8 on standardized tests. However, for students who were not accustomed to writing on computer and could only keyboard at relatively low levels, taking the tests on computer diminished performance. Finally, a third study, which focused on the Massachusetts Comprehensive Assessment Systems (MCAS) Language Arts Tests, demonstrated that removing the mode of administration effect for writing items would have a dramatic impact on the study district's results. Figure 6-1 shows the implications of the 1999 MCAS results: 19 percent of the 4th graders classified as "Needs Improvement" would move up to the "Proficient" performance level, and an additional 5 percent of students who were classified as "Proficient" would be deemed "Advanced".

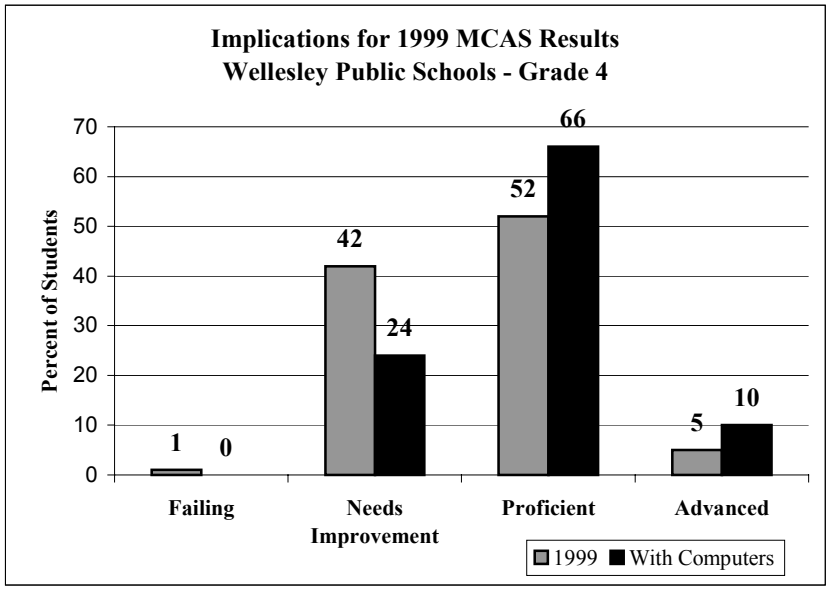


Figure 6-1 Mode of administration effect on grade 4 1999 MCAS results.
SOURCE: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College

As new technologies develop and become a regular component of classroom instruction, it is quite likely that they too will place similar pressures on testing. While it is difficult to see into the future, we are already starting to hear calls for the use of graphic calculators during math tests (Forster & Mueller, 2001). As voice recognition rapidly improves, it is likely that for many students it will become the preferred method of composing text.

Already some schools are using voice recognition software with students with learning disabilities. As students become comfortable and dependent on these and other emerging learning and production tools, it is likely that tests that prohibit the use of these tools will again underestimate the performance of these students. In turn, testing programs will be challenged to adapt their policies and procedures to accommodate these emerging technologies in order to maintain construct validity.

EFFICIENCY

Beyond their use at the classroom level, computer-based technologies can greatly increase the efficiency of testing. To a large extent, testing programs have already capitalized on the efficiencies afforded by technology. As one example, computer-adaptive testing combines a computer-based delivery system with algorithms that select items targeted at the test-taker's estimated ability, The algorithm refines this ability estimate as the test-taker succeeds or fails on each targeted item or sets of items; it then presents additional items targeted at the refined ability estimate. This iterative process occurs until the test-taker's ability estimate stabilizes. Computer-adaptive testing complicates the item selection and delivery process, but because it is usually able to obtain an ability estimate based on a smaller set of items, it is more efficient than traditional paper-based tests.

Computer-based technologies are also impacting the efficiency with which open-ended items are scored. Developments in computer-based scoring date back to the work of Ellis Page during the late 1960s. Since Page's pioneering efforts (1966, 1968), four approaches to computer-based scoring have evolved and have begun to be used to score student work, both in the classroom and on large-scale testing programs (see Rudner [2001] for an overview of these four methods). For all four approaches, studies have demonstrated that the scores produced by these computer algorithms are as reliable as scores produced by two independent human readers. Clearly, once in a digital format, the use of computer scoring systems can dramatically increase the speed with which open-ended responses are scored and reduce the costs required to compensate human scorers. The use of computer-based scoring systems also could allow examinees to obtain more immediate, or even preliminary, feedback, and this would increase the utility of open-ended tests to inform instruction in a timely manner.

Similarly, moves to administer tests via the internet have the potential to greatly increase the efficiency and utility of testing (Bennett, 2001). By eliminating the need to distribute, collect, and then scan paper-based tests, the internet can streamline distribution, administration, and scoring into a seamless and nearly instantaneous process. In turn, the rapid return of test results could provide valuable information to students and teachers in a timely manner.

As test developers continue to grow familiar with new and developing computer-based technologies, it is likely that they will discover other ways to improve the efficiency of testing. Already, some testing programs are experimenting with ways to generate large banks of test items via computer algorithms with the hope of saving the time and money currently required to produce test items manually (Bennett, 1999).

DISRUPTIVE APPLICATIONS OF COMPUTER-BASED TECHNOLOGIES

As Madaus has long emphasized, testing is its own technology with its own "body of special knowledge, skills, and procedures" (2001, p. 1). While the applications of computer-based technologies described above may increase the validity of inferences based on tests and may increase the efficiency of testing, these applications do not fundamentally impact the technology of testing itself. Even in the cases of adaptive testing and item generation, the psychometric principles and "rules" for test construction developed over the past 50 years are applied without significant alteration to determine which items are to be used in a given situation. In this way, applications to improve the validity or efficiency of testing are layered on top of the existing and long-established technology of testing.

Computer-based technologies, however, offer tremendous opportunities to dramatically alter the technology of testing. The ability of computers to present complex, multi-step problems that may incorporate several types of media, have several different paths to reach a solution, or have multiple solutions, coupled with the computer's ability to record the examinee's every action, creates opportunities to learn about students' knowledge, conceptual understanding, and cognitive development in ways that today's technology of testing cannot.

Although the principles and procedures of the current technology of testing are sound, several shortcomings arise. Despite efforts to incorporate open-ended items into some tests, most test items result in binary information about a student, namely, did he or she answer correctly or incorrectly? While scoring guides for some open-ended items focus on the procedures and cognitive process students use to solve problems, these items are dependent upon students' descriptions of their processes, which are often incomplete and inaccurate reflections of their actual processes. As a result, these items provide very indirect and crude insight into examinees' cognitive processes.

Similarly, while the educational community uses tests for a variety of purposes including diagnosing students' strengths and weaknesses; measuring achievement, aptitude, and ability; assessing the impact of instruction on student learning; and examining the quality of education students receive within a school, district, state or even country, test experts have long argued that the current technology of testing should be applied to meet a single purpose at a time. As Haney, Madaus, and Lyons argue, the fundamental problem with using a single test or assessment for multiple purposes is that "such tests require…fundamentally different characteristics" (1993, p. 264). Nonetheless, many current testing programs attempt to use a single test or set of closely related tests to fulfill multiple purposes. For example, in Massachusetts the MCAS uses results from 10th grade language arts and mathematics tests to: (1) make decisions about student competency and eligibility for graduation; (2) make decisions about the quality of education within individual schools; (3) identify exemplary educational programs; (4) assess the effectiveness of state and local interventions (such as tutoring); and (5) help teachers and schools diagnose student weaknesses. Despite including multiple-item formats and requiring several hours to complete, the tests contain roughly 50 items that are performed by all students across the state. While performance on the same set of items helps reassure the public that decisions about student competency and graduation eligibility are based on the same information, this limited set of items attempts to assess such a broad domain that only a handful of items are used

to measure the subdomains. As a result, there is very little information available to diagnose students' strengths and weaknesses. Moreover, the tests do not attempt to probe why students may have performed poorly within a given subdomain. By administering the same set of items to all students rather than spiraling item sets across sets of students, schools and districts are provided with very limited information about the strengths and weaknesses of their educational programs. In short, while tests like the MCAS ambitiously attempt to satisfy several purposes, they fail to meet these needs.

Beyond the MCAS, several state-developed and commercial tests attempt to help teachers diagnose student weaknesses. These tests, however, focus on specific content within a given domain and often use multiple-choice formats to measure student performance within the several subdomains. As a result, the diagnostic information provided to educators is typically limited to whether or not students tend to succeed or fail on items within a given subdomain. While this information helps educators identify those subdomains that may be in need of further instruction, these diagnostic tests tend to provide little or no information about why students may be struggling within a given subdomain. Rather than diagnosing the misconceptions and/or specific skills sets that interfere with students' mastery of the subdomain, most current diagnostic tests provide little more information than an achievement or mastery test.

Among other shortcomings of current testing practices is that most testing currently occurs outside of instruction. As a result, the amount of instructional time is decreased. Ironically, this problem is exacerbated in settings that administer diagnostic tests on a regular and frequent basis to help focus instruction, or that use a series of achievement tests to better measure the impact of instruction on student learning over time. Each test administration decreases instructional time, whether it is internal or external to the classroom and whether it is teacher-developed or developed external to the classroom. While some educators argue that embedded assessment (see Wilson & Sloane [2000] for an example of an embedded assessment system) will streamline the traditional instructional and assessment cycle, externally developed or mandated tests still diminish instructional time.

It is these shortcomings that disruptive applications of computer-based technology to the technology of testing could well address. Building on learning systems currently in use or under development can provide tremendous potential to capture information about students and their learning during the actual learning process. Presenting complex problems as part of the instructional process, examining the strategies students use to solve these problems, and then comparing these strategies to those of novice and experts in the field could expand the notion of mastery from the ability to consistently answer problems correctly to the ability to incorporate knowledge and skills in a way that resembles expertise. Information collected as students interact with the learning system could also be used to diagnose student learning styles, common errors or tendencies, and misconceptions. Once these elements are identified, the systems could not only help teachers intervene immediately, but also could help structure future instruction in a way that is compatible with the students' learning style. In addition, as students master subdomains, the systems could track student achievement. Because achievement is tracked throughout the year, attempts to assess the educational quality or effectiveness of schools, district, and states could be based on the full range of content and skills addressed during the

entire year. And because the information would be broader and deeper than that provided by current achievement tests, the need for external exams might be eliminated.

Below, I describe two recent collaborative efforts to apply computer-based technologies in a manner that substantially departs from current approaches to testing and assessment.

SURGICAL SIMULATION

The U.S. Army is often credited with sparking the growth of large-scale standardized testing. With the onset of World War I and the need to quickly assess recruits and assign them to various positions believed to require different levels of intelligence, the Army administered the Army Alpha and Beta intelligence tests to over 1.75 million recruits (Gould, 1996). Soon thereafter, school systems began using standardized achievement tests to evaluate program effectiveness (Madaus, Scriven, & Stufflebeam, 1983). Since then, testing has grown into a billion-dollar industry (Clarke, Madaus, Horn, & Ramos, 2001).

Given the initial merit and stimulus the U.S. military gave to the standardized testing industry, it is fitting that the military is now playing a major role in reshaping future assessment methodologies. In 1998 the General Accounting Office issued a report that underscored the need to provide military medical personnel with trauma care training that reflected the injuries encountered during wartime. In response to this report, the U.S. Army Medical Research and Material Command (USAMRMC) Telemedicine and Advanced Technology Research Center (TATRC) has launched several initiatives involving medical simulations. While the main purpose of these initiatives is to develop medical and surgical simulators to efficiently and more effectively train Army medics, these simulators are providing unique opportunities to assess kinesthetic abilities, content knowledge, and medical decision-making skills.

Working collaboratively, the Center for the Study of Testing, Evaluation, and Educational Policy (CSTEEP) at Boston College and the Center for the Integration of Medicine and Innovative Therapy (CIMIT) are applying computer-based technologies to assess several aspects of medic training and proficiency. For example, CIMIT has developed a chest tube and surgical airway simulator. The simulator is intended to train medics how to alleviate three conditions commonly caused by chest trauma, namely, tension pneumothorax (collapsing lung with trapped air under pressure), hemothorax (collapsed lung with blood in the chest cavity), and hemopneumothorax (blood and air in the chest cavity). All three conditions are life-threatening if not alleviated in a relatively short period of time. As part of the learning system, medic recruits first interact with a web-based tutorial that provides information on basic first aid, detailed descriptions of these three conditions, protocols and video demonstrations of the procedures required to alleviate the conditions, and detailed descriptions of common complications and appropriate counteractions. Opportunities for recruits to demonstrate the acquisition of the basic knowledge through traditional multiple-choice items are being incorporated into this component of the learning system. If recruits cannot demonstrate mastery of this information, they are presented with additional information to help them master the content. While this component of the learning system does not expand upon the current technology of testing, the actual simulator does.

Upon demonstrating mastery of the content knowledge, recruits are then introduced to the simulator. The simulator combines a sophisticated mannequin (Figure 6-2 below) that contains flesh-like tissue, bone-like ribs, and pockets of blood-like liquid with a computer that contains an exact model of the mannequin with the addition of internal organs. The surgical tools employed for these procedures are connected to tracking devices that record all movements made inside and outside of the mannequin's chest cavity. By combining the instrument tracking with the simulated model of the mannequin's internal organs, the simulator is able to record the amount of time it takes to perform each task required for a given procedure, while also monitoring the movements of the instruments in three-dimensional space. Using these recorded movements, calculations can be made of the factors that can impact the success of the procedure, such as the speed with which instruments enter the cavity, their angle of entry, and their depth of entry. In addition, it is possible to examine factors such as acceleration and deceleration and changes in the direction and angle of movement. The learning system is able to use the recorded movements to reproduce the procedure on screen and show the relationship between surgical tools, ribs, and key organs that could be harmed.

As a training tool, the simulator provides several benefits. Whereas medics typically practice these procedures on animals and do so only a couple of times, procedures can be performed repeatedly on the simulator (and without harm to animals). In addition, since the mannequin is a reproduction (both externally and internally) of a real human being and has tissue and bone properties very close to real flesh and bones, the training more accurately reflects procedures that will likely be performed in the field. Finally, the portability of the mannequin allows training to occur just about anywhere (even on a helicopter en route to a battlefield).

FIGURE 6-2 CIMIT Chest Tube Trauma Mannequin (also known as VIRGIL). Images taken from Metrics for Objective Assessment of Surgical Skills Workshop (2001, July).
SOURCE: Photo by author.

From an assessment perspective, the simulator enables unique approaches to diagnostics and mastery testing. As the simulator provides opportunities for the recruit to practice new procedures or introduces new complications, the system can identify tendencies such as inserting an instrument at a dangerously steep or flat angle or inserting instruments too deep or shallow. This information can then be shared with the recruit and the instructor.

The simulator also has potential to compare the skill of the recruits with that of masters. These comparisons can be made at the macro- or micro-level. At the macro-level, the enhanced-reality reproductions of the recruits can be layered on top of the reproduction of an expert; this allows the recruit to visually compare the "track" of his or her performance with that of the expert. Through this macro-comparison, important differences in technique may become

apparent. On subsequent attempts, the recruit can then adjust his or her technique until it reflects that of the expert. Figure 6-3 depicts the motion trajectories of novice, trainee, and experienced surgeons as they perform a simulated sinus surgery (note that this example is not from the chest tube simulator). As expertise increases, random motion decreases, and movements become more precise and focused on each of the four specific areas of work.

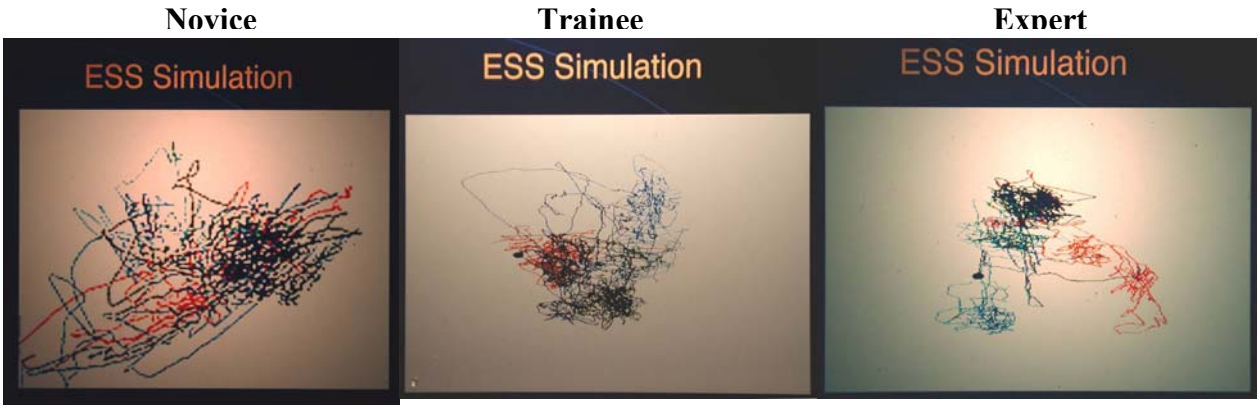


Figure 6-3 Motion trajectories of novice, trainee, and experienced surgeon.
SOURCE: Images taken from Metrics for Objective Assessment of Surgical Skills Workshop summary draft report, 2001.

At the micro-level, individual metrics, such as changes in velocity, angle of insertion, or depth of insertion, can be compared between the recruit and experts. Figure 6-4 compares both the amount of time required to complete a task and the velocity of movements made by advanced and beginning surgeons. In all three trials, the advanced surgeon completed the task in about half the time. In addition, the advanced surgeon executed the tasks with one initial burst of speed and then deliberately slowed down, whereas the beginner made several bursts of speed and inefficiently narrowed in on the target.

Whether focusing on the macro- or micro-images of performance, comparisons between the

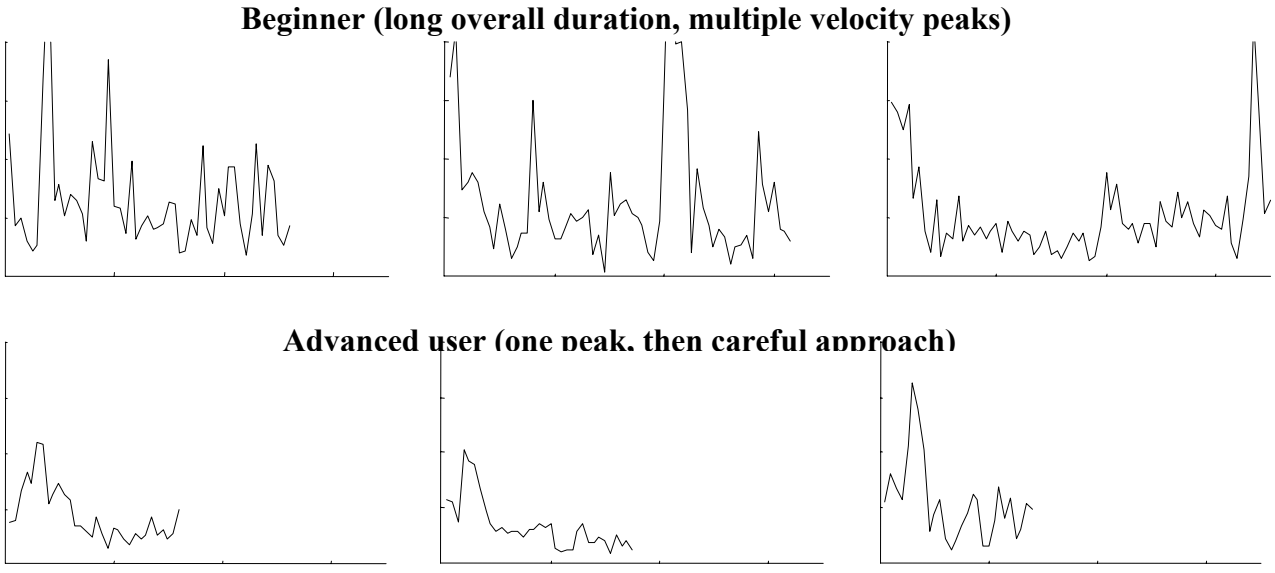


Figure 6-4 Results of time and velocity tracking of novice and expert surgeons.
SOURCE: Images taken from Metrics for Objective Assessment of Surgical Skills Workshop summary draft report, 2001.

performance of the recruits and that of experts may be a useful way to assess mastery. In addition, examining how many practice sessions are required before a recruit's performance reflects that of an expert helps identify recruits who seem to posses innate kinesthetic surgical skills and/or who are rapid learners—recruits who could be very valuable to the Army when training time is cut short by military engagement.

By presenting a recruit with scenarios that involve a range of complications and then examining how the recruit responds, the learning system provides opportuities to examine how well the candidate is able to integrate content and conceptual knowledge with kinesthetic skills. The realism of the scenario could be further enhanced by placing the chest tube simulator in a simulated war environment or by introducing the scenario after extended physical exercise or sleep deprivation. The recruit's ability to respond to complications and conduct the necessary physical movements can be examined in a real-life context. Finally, if the recruit is given access to reference materials that might be available in the field (either during initial training or during future training), his or her ability to rapidly access and apply information to resolve a problem could also be assessed.

K-12 LEARNING ENVIRONMENTS

At first brush, medical simulators may seem far removed from K-12 education. However, the approaches used to collect a diverse set of data about recruits and the challenge of figuring out how to make use of this set of data are directly applicable to learning systems currently in place or under development for K-12 schools.

Recently, CSTEEP and the Concord Consortium have begun to brainstorm ways in which assessments can be built into learning systems. To date, our discussions have been limited to BioLogica, a learning system developed by the Concord Consortium which focuses on genetics. The system is intended to help students learn about genetics through guided exploration. In its current form, BioLogica comprises 13 modules, each of which focuses on a different and increasingly more complex aspect of genetics. In most cases, the modules begin by asking students to explore a specific topic by manipulating genetic traits of a fictitious species of dragons. Figure 6-5 depicts the first exploration students encounter in the second module. In this exploration, students manipulate the dragon's chromosomes to determine how many different ways they can produce a dragon with horns. As each module progresses, new concepts are revealed through guided exploration. For example, the first set of explorations during lesson two culminates by asking students to describe how traits are produced in dragons (Figure 6-6). At times, the learning system presents textual or graphical information to explain concepts and provides students with access to various tools and pieces of information via menu selections. In addition, the system often asks students to demonstrate their understanding via written responses to specific questions, multiple-choice questions, and, most often, modifying different aspects of genetic codes to create dragons with specific traits or to determine how a trait suddenly appeared in a generation of dragons. All the students' interactions with the system are recorded.

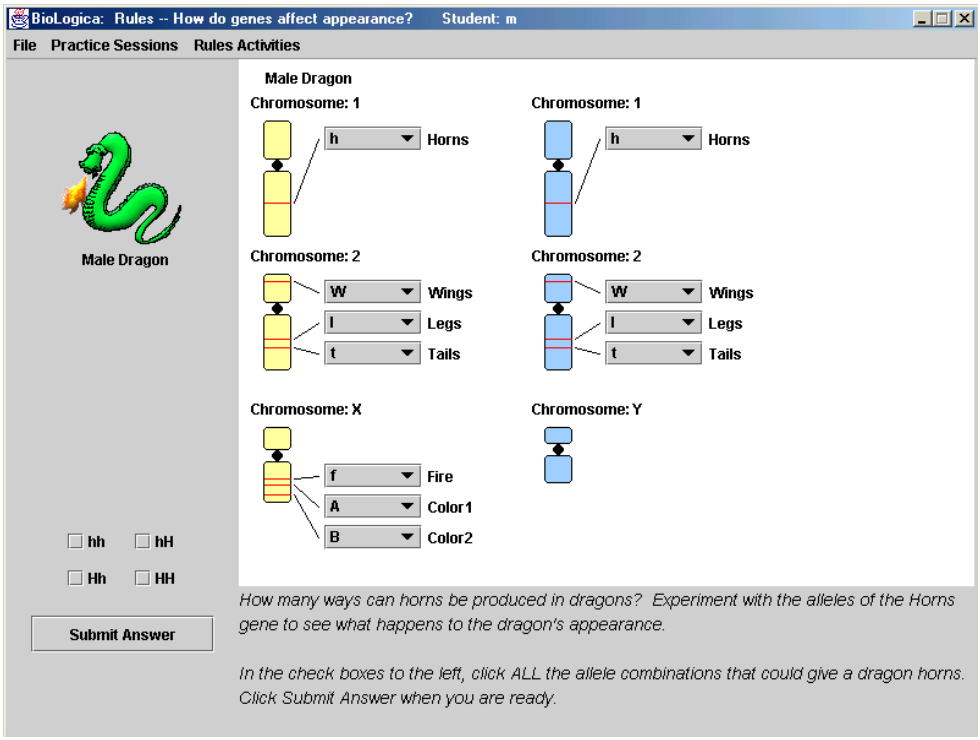


Figure 6-5 First exploration during second module of BioLogica.
SOURCE: BioLogica freeware.

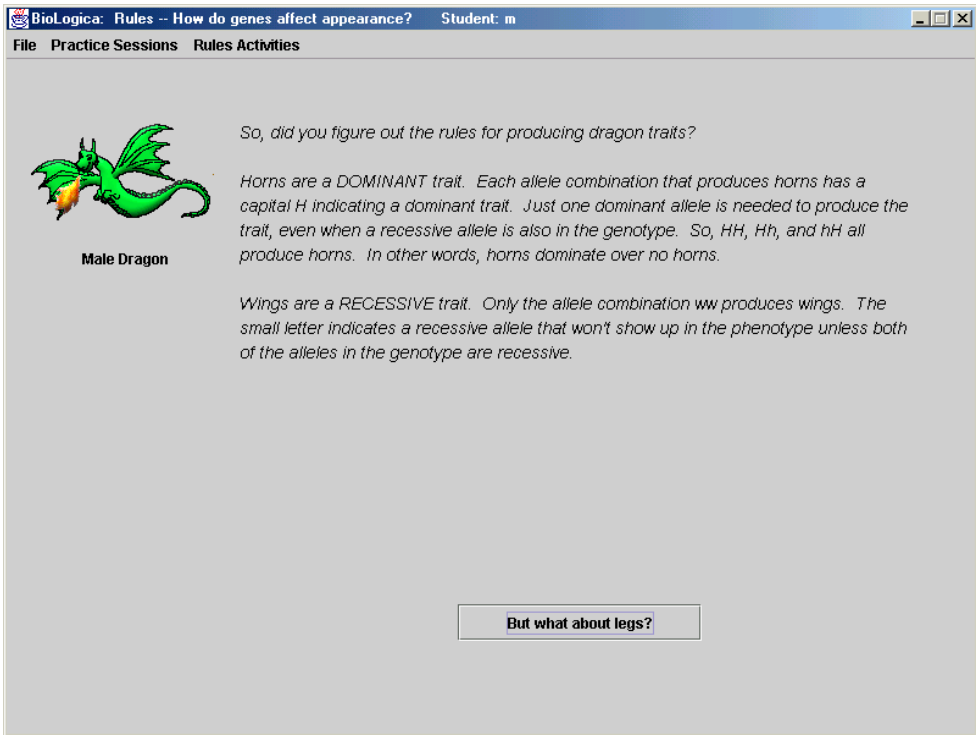


Figure 6-6 Generalizing from guided explorations to rules of genetics in BioLogica.
SOURCE: BioLogica freeware.

From an instructional perspective, BioLogica enables students to explore a complex topic via a variety of media, and it enables teachers to work individually or with small groups of students as questions arise. From an assessment perspective, the learning system provides a number of opportunities to assess student learning. Beyond examining students' understanding via their responses to the multiple-choice and open-ended questions (which could be analyzed by computer), the guided explorations and the problems posed to students present opportunities to:

1) examine students' conceptual understanding by examining the tools and information they opt to use, the amount of time required to solve problems, the type of strategies they employ (e.g., randomly changing chromosomes versus making initial changes on the appropriate chromosomes), as well as their success with the problem;
2) compare the students' pattern of interactions with those of "experts"; and
3) probe apparent misconceptions by presenting additional problems that focus on the specific misconception.

In addition, insight into students' learning styles might be gained by beginning modules and sub-modules in different ways. For example, a module might begin with a textual explanation of a concept, followed by an opportunity to demonstrate understanding. If the student does not demonstrate understanding, subsequent "instruction" might employ a guided exploration of the same concept. If the student still does not demonstrate understanding, a visual presentation of the concept might follow. Across multiple concepts, the order of presentation could be altered, and the efficiency with which the student mastered the concept recorded. After several iterations, the system might identify the preferred order of instructional strategy and utilize that order for that student during subsequent modules.

Finally, and perhaps most important, because the learning system provides multiple opportunities for students to demonstrate conceptual understanding, the need to administer a separate test on the material mastered could be eliminated. Moreover, because records could be sent electronically to any location, it would be possible to maintain a database that indicates which students have mastered each concept. This information could be used by the teacher to identify common misconceptions and inform instruction. In addition, this information could be used to assess achievement at the student level or at a higher level of aggregation. While there might not be much value in recording achievement data at an aggregate level for a single learning system, the value would increase rapidly as more learning systems are used within a school, district, and state. And, again, if this information proves to be redundant to information provided by on-demand, external tests, the external standardized tests might be eliminated.

MOVING FROM VIRTUAL POSSIBILITIES TO REALITY

While the possibilities are enticing, several challenges must first be overcome. These challenges fall into three broad categories: technical, political, and practical.

Technical Challenges

The first major technical challenge involves figuring out which information collected by these systems is most useful for a given purpose and then deciding how to combine this

information so that it is interpretable. By no means is this an easy challenge to overcome. Unlike a traditional multiple-choice test that may contain 50 to 100 pieces of binary information, these systems produce an amount of data spanning several pages and including everything—the amount of time between actions; number of changes made before a solution was found; materials and tools accessed; textual responses; and long lists of items clicked, alterations made, and items moved. While current psychometric models should not be abandoned altogether, new models will need to be created to make use of these multiple pieces of information.

Given the potential of computer-based technology to map actions, whether physical as in the case of surgical simulators or cognitive as in the case of K-12 learning systems, methods of analyzing graphical representations of processes should also be explored. Already, Vendlinski and Stevens (2000) have developed a set of Interactive Multimedia Exercise (IMMEX) programs that can capture a user's path, display the map graphically, and allow teachers and students to compare maps generated at different times. To help automate and standardize these comparisons, recent advances in biometrics may be applicable to assessment in education. As an example, advances in image recognition now make it possible to quickly identify people by comparing video images of their faces with digital photographs stored in large databases. Adapting this technology to compare the paths of learners and experts may prove a useful way to assess level of expertise.

In order to facilitate comparisons between learners and experts, a significant investment must be made in capturing the strategies and processes that experts employ. While this may be a relatively easy task in the case of physical skills (such as those employed during surgery), it is a significantly greater challenge for K-12 learning systems. This challenge is compounded at lower grade levels for which the definition of "expertise" may be radically different than for high school students. While settling on an appropriate definition of expertise may be more political than empirical, acceptable definitions will need to be reached before such comparisons will be broadly embraced.

Much work will also be needed to validate the decisions made as students work with these learning systems. This is particularly true for decisions about academic achievement. While these systems have the potential to greatly reduce or eliminate external testing, these radical changes will not occur unless it can be demonstrated that the information gleaned from these systems are redundant with the information provided by external tests. Moreover, in the current climate of high-stakes testing, it will also be necessary to develop methods of verifying the identity of the student working with the learning system.

Political Challenges

Currently, political and educational leaders strongly embrace large-scale and high-stakes testing, and educational accountability appears to be the top priority shaping our educational system. But political and education leaders appear deaf to the calls for the incorporation of multiple measures into these school accountability systems. One reason for the resistance to broadening the types of measures (be they grades, teachers' judgments, portfolios or work samples, or "performance-based" tests) may be the belief that standardized tests provide more objective, reliable, and accurate measures of student achievement. In part, the failure to expand

the measures used for accountability purposes results from the failure of critics to convince leaders of the utility and validity of these other measures. Although several years of research, development, validation, and disseminations are required before integrated learning and assessment systems could be widely available, efforts should begin now to familiarize political and educational leaders with these methods of assessment. To increase buy-in, roles in the development process should also be created for political and educational leaders.

Additionally, efforts are needed to help leaders see the potential role computer-based technology can play in expanding notions of accountability. As Haney and Raczek (1994) argue, current notions of accountability in education are narrowly defined as examining the performance of schools via changes in their test scores. Under this definition, the iterative process of reflecting on programs and strategies, providing accounts of the successes and shortcomings of those programs, and setting goals in response to those shortcomings is, at best, an informal and secondary component of school accountability. While computer-based learning and assessment systems have the potential to make information provided by current achievement tests redundant and thus eliminate the need for such external tests, computer-based technologies could also be applied today to disrupt current notions of school accountability by providing a forum for schools to account for their practices and to learn from those of other schools. Rather than simply transferring achievement testing from paper to a web-based delivery system (as is currently occurring in Virginia, Oregon, Georgia, and South Dakota), schools could use the internet to collect information about classroom performance (e.g., electronic portfolios or work samples), more closely scrutinize the reliability of scores given to such work, return data from multiple measures in more useful formats, share information and student work with a wider base of constituents, and provide a forum to account for school programs and strategies. Investing now in developing web-based accountability systems that broaden the definition of educational accountability will better set the stage for replacing external state-mandated achievement tests with assessments that are integrated with learning systems.

Practical Challenges

If these disruptive approaches to assessment are to become a regular practice within schools, learning systems like BioLogica will need to be developed in a wide range of topic areas. Anticipating the potential growth of these types of learning systems, the Concord Consortium has developed a scripting language that allows users to easily create new modules for current learning systems or to develop new learning systems. In a sense, this scripting language is analogous to HTML in that it has the potential to standardize learning systems and allow them to interact with one another. Not only will this scripting language be useful for those who want to develop new learning systems, it also provides an easy way to alter current systems so that assessment components can be added or modified.

The high initial cost required to develop a learning system, coupled with the need to have a learning system (or at least a prototype) in use with students before much of the technical work described above can be performed, poses a major obstacle. Not long ago, the National Board on Educational Testing and Public Policy worked with a coalition of schools, political and educational leaders, and internet-based database developers to develop a proposal to design a comprehensive web-based accountability system that builds on Massachusetts' current MCAS.

While the proposal dedicated substantial resources to piloting and validating the system, the high costs associated with developing database engines and interfaces resulted in a research and development budget that was too large to be attractive to funders. The same potential challenge exists for learning and assessment systems. One strategy is to focus first on those systems that are already in use or already have funding to support development. Collaborating with the developers of existing systems can substantially reduce the resources required to support the development and validation of new approaches to assessment; it can also mean access to sets of data from systems that are already in use in schools. For example, BioLogica is currently being used by some 10,000 students across the nation. Because BioLogica is delivered via the web, its modules can be easily updated and student data sent to a central database. Thus, rather than investing two to three years in developing a learning system, project planners who work with the highest quality systems that are currently in use (or will soon be in use) will have opportunities today to begin exploring some of the technical challenges outlined above.

A third practical challenge involves tapping expertise from a range of fields. As the NRC's *Knowing What Students Know* (2001) notes, collaboration among cognitive scientists, assessment experts, and content experts is needed to better inform the development of new approaches to assessment. But in addition, input from instructional leaders and developers of technology is also needed to better anticipate how these systems might be used within classrooms, and how emerging computer-based technologies might impact these systems. Finally, as noted above, political and educational leaders must be brought into the research and development process to better assure that these systems will be accepted as valid means of measuring student achievement.

Clearly, there is a tremendous amount of work that must be performed before these learning systems can adequately meet assessment needs. As the way students—whether they be children in the K-12 classroom or Army recruits—learn changes, there are important opportunities to acquire a more thorough and useful understanding of how and what students learn. Without question, as current and future computer-based technologies are used regularly in the classroom, they will continue to pressure changes in testing. While small-scale studies may be required initially to demonstrate the need to incorporate these technologies into testing, the primary responsibility for examining and implementing changes falls on the testing programs themselves. Given the financial rewards the testing industry will realize, it is likely that it will continue to take on the challenge of developing ways to apply computer-based technologies to increase the efficiency of testing. However, because of the potential of computer-based technologies to seriously disrupt the current technology of testing, it is unlikely that the testing industry itself will invest in researching and developing disruptive uses of computer-based technology. The potential positive impacts that integrated learning and assessment systems could have on teaching and learning, coupled with the vast amount of technical work that must be done to develop these new methodologies, make it imperative that the educational community follow the military's lead by investing now in developing disruptive applications of computer-based technology to the technology of testing and assessment.

REFERENCES

Bennett, R.E. (2001). How the internet will help large-scale assessment reinvent itself. *Educational Policy Analysis Archives*, *9*(5) [Available online: http://epaa.asu.edu/epaa/v9n5.html].

Bennett, R.E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practices, 18*(3), 5-12.

Clarke, M, Madaus, G., Horn, C., & Ramos, M. (2001). The marketplace for educational testing. *National Board on Educational Testing and Public Policy Statements, 2*(3).

Dion, G., Harvey, A., Jackson, C., Klag, P., Liu, J., & Wright, C. (2000). *SAT program calculator use survey*. Paper issued by the Educational Testing Service, Princeton, NJ.

Dunham, P.H., & Dick, T.P. (1994). Research on graphing calculators. *Mathematics Teacher*, *87*(6).

Forster, P.A., & Mueller, U. (2001). Outcomes and implications of students' use of graphics calculators in the Public Examination of Calculus. *International Journal of Mathematical Education in Science and Technology*, *32*(1), 37-52.

Gould, S.J. (1996). *The mismeasure of man*. New York: W.W. Norton.

Haney, W.M., Madaus, G.F., & Lyons, R. (1993). *The fractured marketplace for standardized testing*. Boston: Kluwer Academic.

Haney, W., & Raczek, A. (1994). *Surmounting outcomes accountability in education*. Paper prepared for the U.S. Congress Office of Technology Assessment.

Kenelly, J. (1990). Using calculators in the standardized testing of mathematics. *Mathematics Teacher*, *83*(9), 716-20.

Madaus, G.F. (2001). Educational testing as a technology. *National Board on Educational Testing and Public Policy Statements, 2*(1).

Madaus, G.F., Scriven, M S., & Stufflebeam, D.L., (1983). Program evaluation: A historical overview. In *Evaluation Models: Viewpoints on Educational and Human Services Evaluation*, G. Madaus, M. Scriven, & D. Stufflebeam (Eds.). Boston: Kluwer-Nijhoff.

Metrics for Objective Assessment of Surgical Skills Workshop: Developing Quantitative Measurements through Surgical Simulation. (2001, July). Summary (draft) report for conference held in Scottsdale, AZ.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). Board on Testing and Assessment. Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Page, E. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*, 238-243.

Page, E. (1968). The use of computers in analyzing student essays. *International Review of Education, 14*(2), 210-221.

Rudner, L. (2001). Bayesian Essay Test Scoring System – BETSY [Available online: http://ericae.net/betsy/].

Russell, M. (1999). Testing writing on computers: A follow-up study comparing performance on computer and on paper. *Educational Policy Analysis Archives, 7*(20).

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, *5*(3), http://olam.ed.asu.edu/epaa/v5n3.html.

Russell, M., & Plati, T. (2001). Mode of administration effects on MCAS composition performance for grades eight and ten. *Teachers College Record* [Available online: http://www.tcrecord.org/Content.asp?ContentID=10709].

Vendlinski, T., & Stevens, R. (2000). The use of artificial neural nets (ANN) to help evaluate student problem solving strategies. In B. Fishman & S. O'Connor-Divelbiss (Eds.), *Proceedings of the Fourth International Conference of the Learning Sciences* (pp. 108-114). Mahwah, NJ: Erlbaum.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181-208.

Chapter 7
Design of Automated Authoring Systems for Tests

Eva L. Baker
The Center for the Study of Evaluation
University of California, Los Angeles

INTRODUCTION AND BACKGROUND

This paper will address the goals and requirements of computer-based tools and systems to support the design of assessment tasks.[1] This is not a new idea, but one with a conceptual history (Baker, 1996; Bunderson, Inouye, & Olsen, 1989; O'Neil & Baker, 1997) and early work in item generation (Millman & Greene, 1993; Roid & Haladyna, 1982). The rationale for an increased investment in R&D in this area resides in the improved availability of software tools, modern understanding of assessment and validity, present practice, and unresolved difficulties in assessment design and use. There are five underlying claims that should affect any R&D on assessment design and development:

- Achievement test design needs improvement in order to meet the challenges of measuring complex learning, within cost and time constraints, and with adequate validity evidence.
- The theory of action underlying accountability-focused testing requires that single tests or assessments be employed for a set of multiple, interacting purposes: diagnosis, instructional improvement, certification, program evaluation, and accountability (Baker & Linn, in press).
- For the most part, tests developed for one purpose are applied on faith to meet other educational purposes. There is almost no validity evidence supporting these multiple purposes in widely used achievement tests. Such evidence is needed (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999)
- An integral part of improved learning is the idea that assessments that occur during instruction (whether computer- or teacher-delivered) need to provide relevant information about performance in the target domain of competence. Alignment of these tests is essential, and teachers are an essential target for test development assistance (Baker & Niemi, 2001).
- Performance assessments provide one source of practical knowledge for improvement, but sustained systematic strategies for their development, validation, and implementation have been neither clearly articulated nor widely accepted.

[1] "Assessment" and "test" are terms that will be used interchangeably.

A second set of assertions pertains to the present state of online assessment and authoring systems:

- Online assessments, using simulations, open-ended oral or verbal responses, other constructed responses, and automated approaches to development are relatively well in hand as proof of concept examples (Braun, 1994; Clauser, Margolis, Clyman, & Ross, 1997; Bennett, 2001).
- Authoring components to create integrated testing systems have been described by Frase and his colleagues (in press). Schema or template-based, multiple-choice development, and test management systems have made significant progress (Bejar, 1995; Bennett, in press; Chung, Baker, & Cheak, 2001; Chung, Klein, Herl, & Bewley, 2001; Gitomer, Steinberg, & Mislevy, 1995; Mislevy, Steinberg, & Almond, 1999).
- New assessment requirements, growing from federal statutes or from the expanded role of distance learning, will continue to propagate. Efficient means of online test design need to be built.

Much of the current effort has been devoted to improving computer-administered tests so that they provide more efficient administration, display, data entry, reporting, and accommodations. Ideally, computer administration will enhance measurement fidelity to desired tasks and the overall validity of inferences drawn from the results. A good summary of the promise of computerized tests has been prepared by Bennett (2001). Computerized scoring approaches for open-ended tasks have been developed. Present approaches to essay scoring depend, one way or another, on a set of human raters (Burstein, 2001; Burstein et al., 1998; Landauer, Foltz, & Laham, 1998; Landauer, Laham, Rehder, & Schreiner, 1997). Other approaches to scoring have used Bayesian statistical models (Koedinger & Anderson, 1995; Mislevy, Almond, Yan, & Steinberg, 2000) or expert models as the basis of performance scoring (Chung, Harmon, & Baker, in press; Lesgold, 1994). Let us assume that only propositional analyses of text remain to be done. These scoring approaches will apply ultimately to both written and oral responses.

DESIRABLE FEATURES OF AN AUTOMATED AUTHORING SYSTEM

If we argue that a significant R&D investment is needed to improve test design and, therefore, our confidence in test results, let us envision software tools that result in solving hard and persistent problems, as well as advancing our practice significantly beyond what present stage. What is on our wish list? The goals of one or more configurations of a system are identified below:

- improved achievement information for educational decision making;
- assessment tasks that measure challenging domains, present complex stimuli, and employ automated scoring and reporting options;
- assessment tasks that are useful for multiple assessment purposes;
- reduced development time and costs of high-quality tests;
- support for users with a range of assessment and content expertise, including teachers; and
- reduced timelines for assembling validity evidence.

Let us consider four key categories for organizing the development effort for assessment tasks: cognitive requirements, content, validity, and utility.

Cognitive Requirements

First, we would like to design assessments that require significant intellectual activity for the examinees. We need the assessments to focus primarily on open-ended responses, constructed at one sitting or over time, developed individually or by more than one examinee partner. We want the assessments to reflect explicit cognitive domains, described as families of cognitive demands, with clearly described attributes and requirements. At CRESST (Center for Research on Evaluation, Standards, and Student Testing), we have used a top-level formulation that has guided work by many of our team (Baker, 1997). These cognitive families involve performance tasks with requirements illustrated in Figure 7-1.

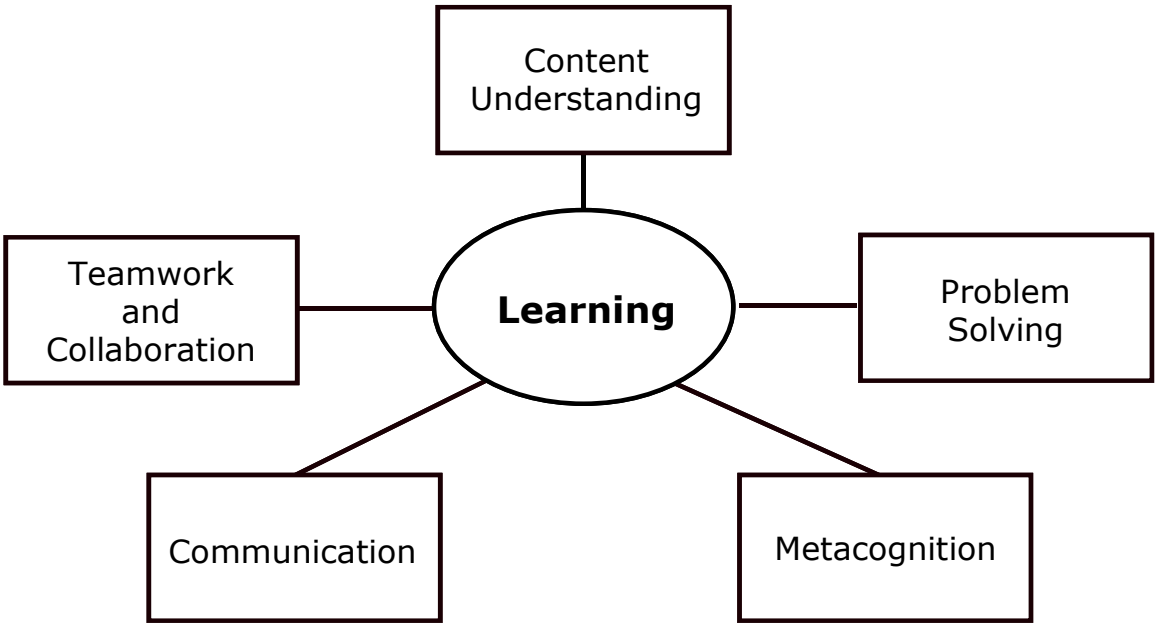


Figure 7-1 Families of cognitive demands as starting points for authoring tasks.

We have been conducting research on these components since 1987 (Baker, Linn, & Herman, 2000; Baxter & Glaser, 1998; Chung, O'Neil, & Herl, 1999; Glaser, Raghavan, & Baxter, 1992; Niemi, 1995, 1996, 1997; O'Neil, Chung, & Brown, 1997; O'Neil, Wang, Chung, & Herl, 2000; Ruiz-Primo, Schultz, Li, & Shavelson, 2001; Ruiz-Primo & Shavelson, 1996; Ruiz-Primo, Shavelson, Li, & Schultz, 2001). These cognitive requirements will call out specific features of tasks and responses, as well as criteria for judging responses. Initially implemented as templates, these cognitive demands should be available in componential form to enable the recombination of sub-elements. To computerize the design of such assessments, the key components or elements would need to be analyzed. For example, in problem solving, we would definitely need to have a component that dealt with problem identification. In a template form of an authoring system, screens would be sequenced that would step the author through the task of deciding how many cues to include and how embedded in text or graphics the presentation of the problem will be. In object form (where object is defined as a subroutine of computer code that

performs the same function), the author would be assisted in using appropriate language so that the problem would be well defined. It is key that the components of cognitive demands are used as the starting point, either in template or object form (Derry & Lesgold, 1996). These components, expressed as rules (or as operating software), are instantiated in subject matter by the author, much as linguistic rules for natural language understanding are instantiated in various content domains. Using cognitive demands as a point of departure, rather than subject matter analysis, will increase transfer of learning across topics and domains because similar frameworks or components will be used in different subject areas. Transfer occurs at the level of the learner, but an approach that starts with cognition may also have a higher payoff. It should enable more coherent instructional approaches for teachers in multiple-subject classrooms or interdisciplinary endeavors.

Content

In discipline-based achievement tests, it is *what* is learned that is of central importance. What is missing in most formulations of test authoring systems is computer-supported strategies to access content to be learned and measured. Some commercial authoring systems step people through the use of templates without providing any assistance on access and editing of content. The fact is that off-line test development has relatively simple approaches to content access. Visit a test development operation, and you may still find content examples and relevant questions stored on 3 x 5 index cards, ready to be sorted into the next tryout. The identification of relevant content, whether for problems, for text, or for examples, is clearly a major bottleneck in test design. Difficult conceptual work is required to identify the rules for inclusion of content in particular domains, a problem made harder and somewhat more arbitrary by the varying standards of clarity in top-level standards intended to be measured. Progress has been made in systems for organizing and searching content (Borgman, Hirsh, Walter, & Gallagher, 1995; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Leazer, 1997; Leazer & Furner, 1999; Lenart, 1995). One of the questions is whether search and organizational rules for document organization can be applied within documents to select candidate content for tests. Clearly, it is time for a merger of browser technology, digital library knowledge structures, and test design requirements. We propose an application of Latent Semantic Indexing (LSI) to the search and acquisition of content for automated design (Wolfe et al., 1998). Procedures to search and import candidate content for use in assessments, as explicit domains of content to be sampled, are needed immediately.

The difficulty of first creating credible, operable templates and then moving to objects (or computer subroutines) cannot be ignored. The problem is technical on two levels. By far the harder part is to identify and regularize the components of tasks, using one or another framework of cognitive demands as the point of departure. To accomplish these tasks, there would need to be agreement on components of key value, e.g., those in problem solving or content understanding. The next phase is to determine the order or orders in which such authoring would occur, including revision loops. Such functional specifications would need to be translated into supportive computer code and embedded in a system with user interfaces to accommodate the potential range of authors, from military trainers to K-12 test developers. Finally, there would need to be a set of activities that demonstrate that components resulted in comparable tasks, first within topics and disciplines and then between them.

The conditions required for the use of browser technology to search and acquire candidate software may be available, but this technology may also require a level of internal coding of content that so far has not been standard in the development of instructional materials. This internal coding would need a proof of concept implementation, so that the additional costs required could be underwritten. Neither of these tasks is easily accomplished, and both require intellectual and financial investment. They are provided as a part of the wish list that describes where we need to be if testing is to be a high-quality practice based on the best we know about human development and technological support. Start-up costs for each project should run around $5 million for about three years.

Validity

The best authoring system would allow users to generate assessments with high technical quality (AERA et al., 1999). The created assessments would provide an adequate degree of accuracy and validity arguments drawn from their subsequent empirical data to document their quality. Assessments intended to meet multiple purposes would require additional technical attributes and relevant evidence supporting their applicability for various uses: making individual, group, or program decisions or supporting prescriptions offered to ameliorate unsatisfactory results.

Acquisition of validity evidence is a second major bottleneck for high-quality tests, apparent because of the lack of evidence relevant for many current test uses. Because such magic is not available, can assessments be designed so that their *a priori* characteristics predict technical quality? There are at least three approaches to consider. One is to use automated review criteria to reduce likely validity problems. Consider an obvious example. There is a great deal of evidence that linguistic barriers (semantic, syntactic, and discourse levels) create construct-irrelevant variance in test performance (Butler & Stevens, 1997; Abedi, 2001). Parsers that identify and highlight such barriers could easily improve the probabilities of reducing this source of error. A second approach is to address characteristics that are known to support particular test purposes. For example, the diagnostic value of an assessment will depend upon the relationship of subtasks to criterion task performance, and the degree of diagnostic confidence is related to the number of items or breadth of contexts used in the assessment task. Another example of qualitative analysis relates to the idea of "objects" in design (Derry & Lesgold, 1996). We would want to assure that assessment tasks, intended to provide a reasonably equivalent level of difficulty in a particular domain, would be descriptively analyzed to be certain that critical features were shared by all tasks in the alleged domain. A third approach is to experiment carefully with features of examinations, and then generate comparable tasks and examine the extent to which they perform as intended, for example, whether they show sensitivity to different instructional interventions.

Perhaps the most challenging issue in the validity/technical quality area is finding ways to reduce the time it takes to assess the validity and accuracy of the test for its various purposes. Authoring systems that incorporate simulation and modeling, rather than relying on laboriously accumulated norming or tryout groups or year-long data collection efforts, are essential if the testing industry is to keep up with policy makers' desires. We believe such simulations are

possible if very small, carefully selected pilot data are used. Obviously, this claim would need to be verified.

Utility

For an authoring system to be useful, it will need to be adapted to the range of users who may be required to design (or interpret) tests. Thus, interfaces and technical expertise are required to make components or entire test design systems operate successfully. User groups with different levels of expertise will need systems to adapt to, and compensate for, limits in their expertise, interest, or time. These groups include teachers (who need to create assessments that map legitimately to standards and external tests), local school district and state assessment developers, the business community, and commercial developers. Not everyone intends to create full-service tests. For example, school district, state, and military personnel may use such an authoring system to design prototypes of tasks in order to communicate their intentions for assessment systems to potential contractors. A diverse audience will mean a range of expertise in the background knowledge required for the system. The range will include knowledge about testing, subject matter, and learning. Embedded tutorials, explaining default conditions and advising users on why decisions they make may be inappropriate, will need to be built and verified.

Although it may be obvious, it is still worth saying that exposure to such a system should result in positive payoff for instructional design and teaching. It is possible that analytical and creative thinking inspired by such authoring environments will spill over to teaching design as well.

Common standards for design, communication, and data reporting are also required of authoring systems. At the present time, the tension between proprietary test design and quality is clear, and far too often algorithms and procedures are cloaked by the shadow of commercial endeavor, a reality that makes choices among measures rely on preference for surface features of tests. A system like SCORM (Shareable Content Object Reference Model) would be ideal. SCORM, which is used by the U.S. Department of Defense for its training procurements, describes standards guiding the interoperability of components and content.

Design Phases

Competing complete systems should be designed and applied to high-priority areas. These will probably remain in the template mode in the short run. Generalizability of their utility for different content, tasks, cognitive demands, and examinees can be assessed. Simultaneous efforts should be made to create reusable components (objects) to improve aspects of the design process, including specifications and simulation authoring systems (see, for example, RIDES [Munro et al., 1997] and VIVIDS [Munro & Pizzini, 1998]). In addition, funding should be available for competing analyses of the objects or modules needed to develop fully object-oriented assessment design environments. Competing designs will differ on the level of granularity and on the degree to which they can be easily recombined to generate new assessment prototypes. Finally, we need a fundamental analysis of the components of performance, including task, content, and cognitive and linguistic demands. Using the metaphor

of the genome, we speak of the *Learnome* (Baker, 2000). Investment in the *Learnome* and its resulting primitives could greatly improve our understanding of the components of performance, assessment, and instruction.

RECOMMENDATIONS FOR RESEARCH PRIORITIES

- Fund competing, publicly available authoring components, requiring proof-of-concept to include validity evidence in at least three different task areas.
- Fund competing template-focused systems designed for users with different levels of expertise.
- Fund competing total object-oriented systems, requiring common interoperability standards, addressing different ages of learners and different task complexity.
- Specifically fund approaches that import candidate content for use in assessment design and development.
- Fund fundamental descriptive domain and performance analyses intended to result in primitives for use in future object-oriented systems.
- Fund research intended to model and speed up validity evidence for the development of new measures.

REFERENCES

Abedi, J. (2001). *Standardized achievement tests and English language learners: Psychometrics and linguistics issues* (Technical Report). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Baker, E.L. (1997). Model-based performance assessment. *Theory Into Practice, 36*(4), 247-254.

Baker, E.L. (1996). Readying NAEP to meet the future. In G. Bohrnstedt (Ed.), *Evaluation report on the 1994 NAEP trial state assessment*. Palo Alto, CA: National Academy of Education.

Baker, E.L. (2000, November). *Understanding educational quality: Where validity meets technology*. William H. Angoff Memorial Lecture Series. Princeton, NJ: Educational Testing Service.

Baker, E.L., & Linn, R.L. (in press). Validity issues for accountability systems. In R. Elmore & S. Fuhrman (Eds.), *Redesigning accountability*.

Baker, E.L., Linn, R.L., & Herman, J.L. (2000). *Continuation proposal* (submitted to the Office of Educational Research and Improvement, U.S. Department of Education). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E.L., & Niemi, D. (2001). *Assessments to support the transition to complex learning in science* (Proposal to the National Science Foundation). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baxter, G.P., & Glaser, R. (1998). The cognitive complexity of science performance assessments. *Educational Measurement: Issues and Practice*, *17*(3), 37-45.

Bejar, I.I. (1995). From adaptive testing to automated scoring of architectural simulations. In E.L. Mancall & P.G. Bashook (Eds.), *Assessing clinical reasoning: The oral examination and alternative methods* (pp. 115-130). Evanston, IL: The American Board of Medical Specialties.

Bennett, R.E. (in press). An electronic infrastructure for a future generation of tests. In H.F. O'Neil, Jr. & R. Perez (Eds.), *Technology applications in education: A learning view*. Mahwah, NJ: Erlbaum.

Bennett, R.E. (2001). How the internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives*, *9*(5), 1-26.

Borgman, C.L., Hirsh, S.G., Walter, V.A., & Gallagher, A.L. (1995). Children's searching behavior on browsing and keyword online catalogues: The science library catalogue project. *Journal of the American Society for Information Science*, *46*, 663-684.

Braun, H. (1994). Assessing technology in assessment. In E.L. Baker & H.F. O'Neil, Jr. (Eds.), *Technology assessment in education and training* (pp. 231-246). Hillsdale, NJ: Erlbaum.

Bunderson, C.V., Inouye, D.K., & Olsen, J.B. (1989). The four generations of computerized educational measurement. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 367-408). New York: Macmillan.

Burstein, J. (2001, April). *Automated essay evaluation with natural language processing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Burstein, J., Kukich, K., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Lu, C., Nolan, J., Rock, D., & Wolff, S. (1998). *Computer analysis of essay content for automatic score prediction: A prototype automated scoring system for GMAT analytical writing assessment* (ETS Rep. RR-98-15). Princeton, NJ: Educational Testing Service.

Butler, F.A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations* (CSE Tech. Rep. No. 448). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Chung, G.K.W.K., Baker, E.L., & Cheak, A.M. (2001). *Knowledge mapper authoring system prototype*. (Final deliverable to OERI). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Chung, G.K.W.K., Harmon, T.C., & Baker, E.L. (in press). The impact of a simulation-based learning design project on student learning. *IEEE Transactions on Education*.

Chung, G.K.W.K., Klein, D.C.D., Herl, H.E., & Bewley, W. (2001). *Requirements Specification for a knowledge mapping authoring system*. (Final deliverable to OERI). Los

Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Chung, G.K.W.K., O'Neil, H.F., Jr., & Herl, H.E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior*, *15*, 463-494.

Clauser, B.E., Margolis, M.J., Clyman, S.G., & Ross, L.P. (1997). Development of automated scoring algorithms for complex performance assessments: A comparison of two approaches. *Journal of Educational Measurement*, *34*(2), 141-161.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391-407.

Derry, S., & Lesgold, A. (1996). Toward a situated social practice model for instructional design. In D.C. Berliner & R.C. Calfee (Eds.), *Handbook of educational psychology* (pp. 787-806). New York: Macmillan.

Frase, L.T., Almond, R.G., Burstein, J., Kukich, K., Mislevy, R J., Sheehan, K. M., Steinberg, L S., Singley, K., & Chodorow, M. (in press). Technology and assessment. In H.F. O'Neil, Jr. & R. Perez (Eds.), *Technology applications in education: A learning view*. Mahwah, NJ: Erlbaum.

Gitomer, D.H., Steinberg, L.S., & Mislevy, R.J. (1995). Diagnostic assessment of trouble-shooting skill in an intelligent tutoring system. In P.D. Nichols, S.F. Chipman, & R.L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 73-101). Hillsdale, NJ: Erlbaum.

Glaser, R., Raghavan, K., & Baxter, G.P. (1992). *Cognitive theory as the basis for design of innovative assessment: Design characteristics of science assessments* (CSE Tech. Rep. No. 349). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Koedinger, K.R., & Anderson, J.R. (1995). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8,* 30-43.

Landauer, T.K., Foltz, P.W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259-284.

Landauer, T.K., Laham, D., Rehder, B., & Schreiner, M.E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In M.F. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412-417). Mahwah, NJ: Erlbaum.

Leazer, G.H. (1997). *Examining textual associations using network analysis*. Paper presented at the International Association for Social Network Analysis (Sunbelt), San Diego, CA.

Leazer, G.H., & Furner, J. (1999). Topological indices of textual identity networks. *Knowledge: Creation, Organization and Use: Proceedings of the 62nd ASIS Annual Meeting*, *36*, 345-358.

Lenart, D.B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, *38*(11), 32-38.

Lesgold, A. (1994). Assessment of intelligent training technology. In E.L. Baker & H.F. O'Neil, Jr. (Eds.), *Technology assessment in education and training* (pp. 97-116). Hillsdale, NJ: Erlbaum.

Millman, J., & Greene, J. (1993). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335-366). New York: Macmillan.

Mislevy, R.J., Almond, R.G., Yan, D., & Steinberg, L.S. (2000). *Bayes nets in educational assessment: Where do the numbers come from?* (CSE Tech. Rep. No. 518). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (1999). *Evidence-centered assessment design*. Princeton, NJ: Educational Testing Service.

Munro, A., Johnson, M., Pizzini, Q.A., Surmon, D., Towne, D., & Wogulis, J. (1997). Authoring simulation-centered tutors with RIDES. *International Journal of Artificial Intelligence in Education*, *8*, 284-316.

Munro, A., & Pizzini, Q.A. (1998). *VIVIDS reference manual*. Los Angeles: University of Southern California, Behavioral Technology Laboratories.

Niemi, D. (1996). Assessing conceptual understanding in mathematics: Representation, problem solutions, justifications, and explanations. *Journal of Educational Research*, *89*, 351-363.

Niemi, D. (1997). Cognitive science, expert-novice research, and performance assessment. *Theory into Practice*, *36*(4), 239-246.

Niemi, D. (1995). *Instructional influences on content area explanations and representational knowledge: Evidence for the construct validity of measures of principled understanding—mathematics* (CSE Tech., Rep. No. 403). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

O'Neil, H.F., Jr., & Baker, E.L. (1997). A technology-based authoring system for assessment. In S. Dijkstra, N.M. Seel, F. Schott, & R.D. Tennyson (Eds.), *Instructional design: International perspectives. Vol. II: Solving instructional design problems* (pp. 113-133). Mahwah, NJ: Erlbaum.

O'Neil, H.F., Jr., Chung, G.K.W.K., & Brown, R. (1997). Use of networked simulations as a context to measure team competencies. In H.F. O'Neil, Jr. (Ed.), *Workforce readiness: Competencies and assessment* (pp. 411-452). Mahwah, NJ: Erlbaum.

O'Neil, H.F., Jr., Wang, S-L., Chung, G.K.W.K., & Herl, H.E. (2000). Assessment of teamwork skills using computer-based teamwork simulations. In H.F. O'Neil, Jr. & D.H. Andrews (Eds.), *Aircrew training and assessment* (pp. 245-276). Mahwah, NJ: Erlbaum.

Roid, G.H., & Haladyna, T.M. (1982). *A technology for test-item writing*. New York: Academic Press.

Ruiz-Primo, M.A., Schultz, S.E., Li, M., & Shavelson, R.J. (2001). Comparison of the reliability and validity of scores from two concept-mapping techniques. *Journal of Research in Science Teaching*, *38*(2), 260-278.

Ruiz-Primo, M.A., & Shavelson, R.J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, *33*(6), 569-600.

Ruiz-Primo, M.A., Shavelson, R.J., Li, M., & Schultz, S.E. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment*, *7*(2), 99-141.

Wolfe, M.B.W., Schreiner, M.E., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W., & Landauer, T.K. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes*, *25*(2-3), 309-336.

Appendix A
Workshop Agenda

The National Academies
BOARD ON TESTING AND ASSESSMENT
Workshop on Technology and Assessment: Thinking Ahead
Green Building Room 130, 2001 Wisconsin Avenue, NW
Wednesday, November 14, 2001

8:00 a.m. Breakfast

8:30 Welcome and introductions

- Marshall (Mike) Smith, The William and Flora Hewlett Foundation
- Michael Feuer, Director, NRC Center for Education
- Pat DeVito, Director, Board on Testing and Assessment

9:00 Advances in cognition, measurement, and technology highlighted in the National Research Council report *Knowing What Students Know*

- *Rethinking the foundations of assessment*
  Jim Pellegrino, University of Illinois, Chicago
- *Advances in the sciences of thinking and learning*
  Rich Lehrer, University of Wisconsin
- *Advances in measurement and statistical modeling*
  Mark Wilson, University of California, Berkeley
- *Assessment design and use and the role of technology*
  Jim Pellegrino
- *Reemphasizing an important message: The need for informative assessments*
  Robert Glaser, Learning Research and Development Center, University of Pittsburgh
- *A perspective from the sponsor*
  Larry Suter, National Science Foundation

10:30 Break

10:45 Discussants

- Lorrie Shepard, University of Colorado
- Jose Mestre, University of Massachusetts

Followed by audience questions for panel

11:45 An example of the principles set forth in *Knowing What Students Know*: The Algebra I Cognitive Tutor

- Albert Corbett, Carnegie Mellon University

12:00 noon Lunch

Demonstration: The Algebra I Cognitive Tutor

- Albert Corbett

1:00 p.m. Information technologies: Opportunities for advancing educational assessment

- *Session moderator*
  Jim Pellegrino
- *Technology and the unmasking of constructs*
  Drew Gitomer, Educational Testing Service
- *Surgical simulations and other learning systems that offer potentially rich assessment information*
  Mike Russell, Boston College
- *Computerized speech recognition and the assessment of reading*
  Susan Williams, University of Texas, Austin
- *Technology supports for developing assessments of science inquiry*
  Barbara Means and Geneva Haertel, SRI International
- *Is it worth it? Cost benefits from technology-based assessment in the military*
  Dexter Fletcher, Institute for Defense Analyses

3:00 Break

3:15 Discussants

- Lauren Resnick, Learning Research and Development Center, University of Pittsburgh
- Paul Holland, Educational Testing Service

Followed by audience questions for panel

4:00 Group discussion of research and development priorities

Discussion leader

- Mike Smith

Synthesizer

- *Michael Feuer*

5:00 Adjourn

Appendix B
Board on Testing and Assessment Membership

EVA L. BAKER *(Chair),* The Center for the Study of Evaluation, University of California, Los Angeles
LORRAINE McDONNELL *(Vice Chair)*, Departments of Political Science and Education, University of California, Santa Barbara
LAURESS L. WISE *(Vice Chair)*, Human Resources Research Organization, Alexandria, Virginia
CHRISTOPHER F. EDLEY, JR., Harvard Law School
EMERSON J. ELLIOTT, Independent Consultant, Arlington, Virginia
MILTON D. HAKEL, Department of Psychology, Bowling Green State University
ROBERT M. HAUSER, Institute for Research on Poverty, Center for Demography, University of Wisconsin, Madison
PAUL W. HOLLAND, Educational Testing Service, Princeton, New Jersey
DANIEL M. KORETZ, Graduate School of Education, Harvard University
EDWARD P. LAZEAR, Graduate School of Business, Stanford University
RICHARD J. LIGHT, Graduate School of Education and John F. Kennedy School of Government, Harvard University
ROBERT J. MISLEVY, Department of Measurement and Statistics, University of Maryland
JAMES W. PELLEGRINO, Department of Psychology, University of Illinois, Chicago
LORETTA A. SHEPARD, School of Education, University of Colorado, Boulder
CATHERINE E. SNOW, Graduate School of Education, Harvard University
WILLIAM T. TRENT, Department of Educational Policy Studies, University of Illinois, Urbana-Champaign
GUADALUPE M. VALDES, School of Education, Stanford University
KENNETH I. WOLPIN, Department of Economics, University of Pennsylvania

PASQUALE J. DEVITO, *Director*
LISA D. ALSTON, *Administrative Associate*