WILLIAM H. ANGOFF MEMORIAL LECTURE Series

ETS®
Educational Testing Service
Policy Information Center

UNDERSTANDING EDUCATIONAL QUALITY: WHERE IT MEETS VALIDITY WHERE MEETS TECHNOLOGY

BY
EVA L. BAKER

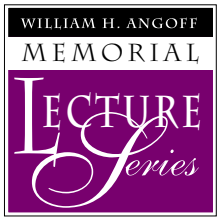
William H. Angoff
1919 - 1993

William H. Angoff was a distinguished research scientist at ETS for more than forty years. During that time, he made many major contributions to educational measurement and authored some of the classic publications on psychometrics, including the definitive text "Scales, Norms, and Equivalent Scores," which appeared in Robert L. Thorndike's Educational Measurement. *Dr. Angoff was noted not only for his commitment to the highest technical standards but also for his rare ability to make complex issues widely accessible.*

The Memorial Lecture Series established in his name in 1994 honors Dr. Angoff's legacy by encouraging and supporting the discussion of public interest issues related to educational measurement. The annual lectures are jointly sponsored by ETS and an endowment fund that was established in Dr. Angoff's memory.

The William H. Angoff Lecture Series reports are published by the Policy Information Center, which was established by the ETS Board of Trustees in 1987 and charged with serving as an influential and balanced voice in American education.

Understanding Educational Quality:
Where Validity Meets Technology



The fifth annual William H. Angoff Memorial Lecture was presented at Educational Testing Service, Princeton, New Jersey, on November 8, 1998.

Eva L. Baker
University of California, Los Angeles
National Center for Research on Evaluation, Standards, and Student Testing

Educational Testing Service
Policy Information Center
Princeton, NJ 08541-0001

Preface

The ETS® Policy Information Center is pleased to publish the fifth annual William H. Angoff Memorial Lecture, given at ETS on November 8, 1998 by Dr. Eva L. Baker of the University of California, Los Angeles and the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

The William H. Angoff Memorial Lecture Series was established in 1994 to honor the life and work of Bill Angoff, who died in January 1993. For more than 50 years, Bill made major contributions to educational and psychological measurement and was deservedly recognized by the major societies in the field. In line with Bill's interests, this lecture series is devoted to relatively non-technical discussions of important public interest issues related to educational measurement.

Dr. Baker's lecture focuses on testing and technology and the connections between the two. She suggests that technology provides a venue and some tools to attack persistent problems in educational assessment, particularly in the K-12 system. Noting that assessment is becoming potentially more powerful as demands for accountability are increased, Dr. Baker calls for an understanding of the conceptual and scientific basis of student learning. To begin this extensive task, she proposes that we adopt the metaphor of the human genome mapping project.

At CRESST, special thanks go to Larry Casey for graphic design and David Westhoff and Katherine Fry for their tireless support. At ETS, Drew Gitomer and Madeline Moritz provided support for the lecture, Amanda McBride was the editor, and Carla Cooper provided desktop publishing services.

Richard J. Coley
ETS Policy Information Center

Preamble

First of all, I'd like to thank the leadership of ETS, and Henry Braun, in particular, for inviting me to speak today. It is a treat to be invited to discuss my ideas with so many of those who have made significant contributions to educational research. I am particularly honored to lecture in commemoration of William Angoff, whose work resulted in major technical and practical insights for the benefit of educational measurement. As advertised, I will discuss today testing and technology, considering both the expected and the less obvious connections between the general areas of educational information and emerging technologies of the future.

My presentation will be in four parts. First, I will start with a discussion about technology itself, so we have some common premises. Second, I will describe some work illustrating how technology might be useful in meeting present purposes for testing, including tests used for communicating with the public, improving teaching and learning, and supporting the increasing demands for accountability. These examples will span the functions of test design, administration, scoring, and reporting. This section, then, deals with present and short-term possibilities. In the third part of my presentation, I will examine some of the persisting anomalies in the design and use of tests in K–12 settings. I will argue that there are many inconsistencies in the present system that demand important and fundamental change. Unless such changes are made, the utility, credibility, and ultimate validity of the K–12 testing system will continue to erode. Fourth, and most important, I will suggest that technology provides a venue and some tools for us to attack these persistent problems in a systematic way. Because the general public still sees technology as a new and separate phenomenon, I will suggest that the use of technology may be especially timely. Technology continues to surprise and astound us, and for this reason, technology now offers us both a context and a shield to investigate paradoxes that cannot be otherwise approached in the world of testing.

Technology Is

To begin, let us reprise what we mean by technology. Most people think about technology in terms of hardware, that is, the platform, the titanium box, displays, buttons, and cables, or more colloquially, the bells and whistles. A more general definition of technology is that it consists of replicable procedures designed to attain specific goals. These procedures might be designed to make existing services and functions more regular and more reliable (such as improved telephone service). Such procedures also can be applied to serve new goals, for instance, the anytime-anywhere use of cellular phones. For either of these purposes, new procedures may be—but don't necessarily have to be—instantiated in supporting software and hardware.

All technology is on a continuum, of course, from slight modifications of the very familiar to the design of radical new systems. All technology also achieves its goals at the expense of others. To many of us, technology also conjures up an enterprise that is recent, changing, and innovative, and that forces us on occasion to leave behind more familiar and sometimes more comfortable options. The "newness" in technology is a matter of vantage point. Consider the topic of windows—glass ones, not the kind that Microsoft® produces. The window was a new technology in the Middle Ages. In churches, its purpose was to give light; to create a particular, prayerful ambiance; and, more practically, to tell religious stories with the result that heretofore private knowledge became distributed. Sort of like the Internet. Windows evolved so that they also could help control temperature in homes, workplaces, and vehicles. If our starting point for investigating the technology of windows were 1930, "new" technology might be the kinds of switches that control the rates by which windows are raised and lowered in our cars, or the advent of improved weather resistance. The rule this example illustrates is that after the technology has been used in one form, it continues to evolve or becomes irrelevant.

We also have learned that as it evolves, technology pushes us to new knowledge and to new applications that we might not have conceived of without its invention. The telescope and Scotch™ tape are examples of inventions that resulted in continual interactive change. The telescope led us to understand more about the cosmos and created the need for more sophisticated optical and radio approaches to exploring the universe. On a more mundane level, with Scotch tape, we moved from its use in masking parts to be left unpainted on cars to, in the '70s, its use in fastening curls or allowing us to decorate kitchens with coral- or avocado-colored decorative adhesive stickers, to the current omnipresence of Post-it® Notes . . . now evolved into technological form on my computer. Like all phenomena of worth, technology has both surface and deeper features: Surface features are where we start, and deeper features support transfer.

Testing and Technology

Turning specifically to the field of testing, technology has already played significant roles in the form of scoring technology, analytical practices, and development strategies. Large-scale testing has depended in great measure on the refinement of machine-scorable approaches to processing student papers. Without such technology, our ability to address and confront uses of testing on a mass level of the sort used in college admissions would not have occurred.

Two common functions of technology and testing are currently at work. First, there is the use of technology to meet more efficiently existing goals: Computer-adaptive tests that tailor the sequence of questions to students' answers is one obvious example. Experiments in automated scoring of student essay performance undertaken by Burstein, Kukich, Landauer and others represent another example of technology that is focused primarily on improving the efficiency of existing practices.

Technology for New Tests

The second role of technology and testing, which I intend to spend more time on, is its use to expand the domains of testing—so that we can measure domains of performance in areas heretofore inaccessible on a broad scale. One example is the use of simulation-based assessment tasks, of the sort ETS has developed in its architecture certification assessment or embedded in intelligent tutoring systems such as HyDrive. In these examples, technology is used to create task fidelity or verisimilitude, as well as to incorporate constraints and goal structures in the testing setting. Selective fidelity or only partially representing tasks and constraints is done in some virtual reality systems, such as in the SIMNET approaches to assessing military performance. Other simulations go for full, multicolor realism, as depicted in the use of patient surrogates in medical assessments. In either case, the role of technology in testing is to expand the bandwith of what can be measured, and it affects the validity of the tasks relative to the domain of performance.

At CRESST (Center for Research on Evaluation, Standards, and Student Testing), we have undertaken cameo projects that may turn out to have significant impact in this area. In one set of work, we have attempted to measure team performance using tasks where groups engage in networked environments (Chung, O'Neil, & Herl, 1999; O'Neil, Chung, & Brown, 1997; O'Neil, Wang, Chung, & Herl, 2000). Using a theory of team performance created by Salas and his colleagues (1992), we have kept track not only of the team's success or level of attainment in meeting goals, but also of the process they used to reach the goal. Essentially, we observed how the team, as a group and as individuals, dealt with five dimensions: adaptability, coordination, decision making, interpersonal support, and leadership. We were able to generate scores in this teamwork domain in real time. The academic tasks, serving as the learning context, were conceptual (showing how much a team knew about a particular science topic) and focused on problem solving (how the team should negotiate on particular workforce issues).

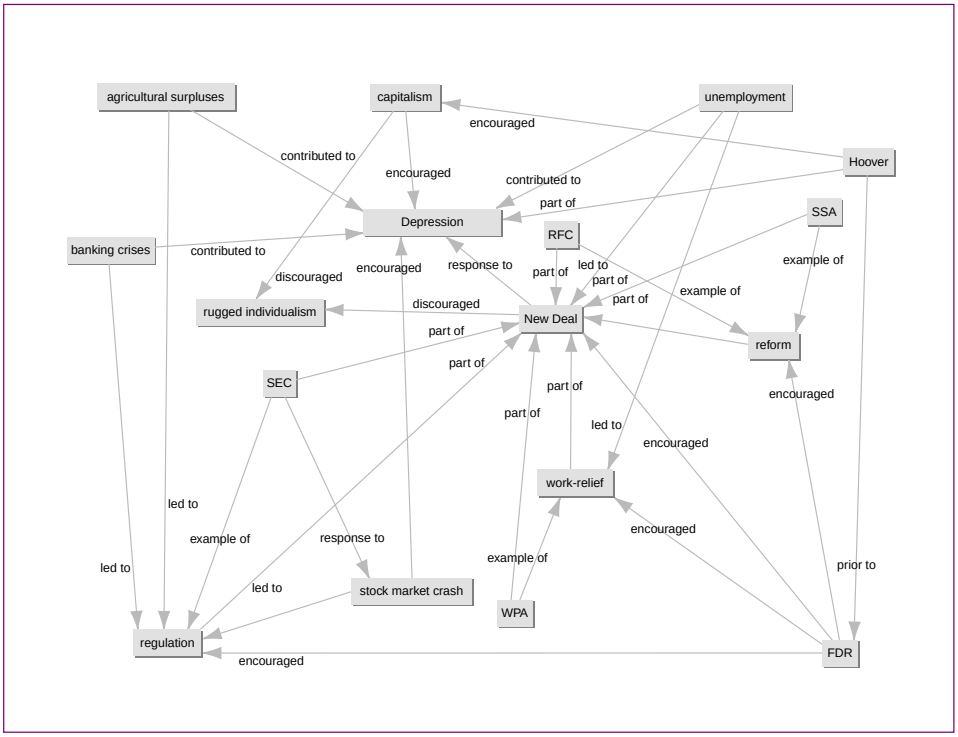
These studies have taught us that even a relatively modest technology goal, such as to expand the domain of what can be measured, creates new challenges. O'Neil (Chung, O'Neil, & Herl, 1999; O'Neil,

Chung, & Brown, 1997; O'Neil, Wang, Chung, & Herl, 2000) and his CRESST colleagues have been struggling with the problem of creating a team or group student model in order to learn how differences in background knowledge or task expectations influence performance, and how to value the performance obtained.

In a related set of studies, we have captured student maps of particular topics to measure the declarative and procedural knowledge of students (Aguirre-Muñoz, 2000; Baker, Niemi, Novak, & Herl, 1992; Herl, O'Neil, Chung, & Schacter, 1999; Herl, O'Neil, Schacter, & Chung, 1998; O'Neil, Herl, Chung, Bianchi, Wang, Mayer, Lee, Choi, Suen, & Tu, 1998; Schacter, Herl, Chung, Dennis, & O'Neil, 1999). To determine the feasibility of looking at mapping performance as an outcome measure, students were first provided access to information (for instance, in history, using primary texts), and then asked to show their understanding of the field using a computer to demonstrate their knowledge (see Figure 1).

With simple interfaces involving pull-down menus students created maps using specified directional links (for instance, that concept x *preceded* concept y) and knowledge elements consisting of facts, concepts, and processes. Performance was scored in real time in comparison to expert maps in the same domains. We have experimented using maps for students with language facility in Korean, Chinese, and Spanish and found these techniques to work with

Figure 1. Great Depression Knowledge Map

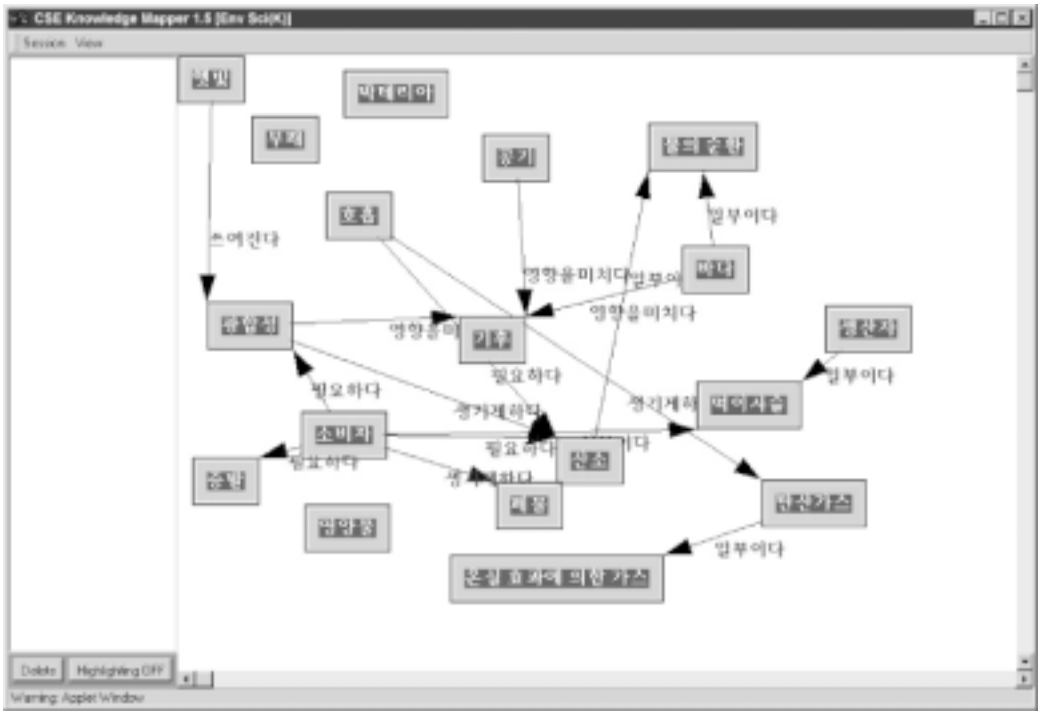


The Knowledge Mapper was developed with funding from the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The Knowledge Mapper does not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

different languages and to be particularly useful for students who do not have full command of English discourse but who do know content (see Figure 2). For English speakers, we have found strong relationships with writing tasks stimulated by comparable texts (Aguirre-Muñoz, 2000; Baker, Niemi, Novak, & Herl, 1992).

At CRESST, we are expanding our R & D to attempt to measure the cognitive processes of students as well as their level of attainment of particular academic goals (Chung & Baker, in process). Analogous to student models used in artificial intelligence systems, we want to obtain parsimonious student models of what students know and how they acquire knowledge. To this end, in one set of studies, we have given students a multi-step task. First, students were asked to complete a knowledge map to show their level of understanding for a science domain in environmental science. Students were able to use many different approaches to create their maps, and maps varied in their superficial characteristics. Nonetheless, students received credit for very different-looking maps. Following the completion of their maps, students were immediately given feedback on the quality of their maps. Then students were given an Internet simulation, consisting of more than 300 pre-selected Web pages that they could search to find knowledge needed to improve their maps. Some of these destinations were more or less relevant and more or less helpful to the knowledge-acquisition task. By keeping track of the sites students visited,

Figure 2. Environmental Science Computer Knowledge Map in Korean



The Knowledge Mapper was developed with funding from the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The Knowledge Mapper does not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

what they abstracted, and how that information influenced their final representations, we were able to make inferences about the role of prior knowledge, content salience, and even the influence of initial representation on final performance.

Taken together, this set of tasks allowed us to make inferences about the learning that children can obtain from Internet exposure and whether Internet activity improves their performance. On the basis of preliminary data obtained in a U.S. Department of Defense study, our tentative answer is that Internet exposure affects performance related to students' improved content knowledge (Herl, O'Neil, Schacter, & Chung, 1998).

We are also applying this approach to problem solving, where we are trying to assess procedural knowledge involving core laws of physics. In addition, we are exploring its use in narrative comprehension, for application in early literacy assessment.

These examples demonstrate the use of technology to broaden the domains that can be tested validly and credibly. These examples illustrate uses serving important, but relatively near-term, goals. Technology applied to testing in the long run will help build acceptance in user communities, as well as provide significant test beds for all to work out more general principles.

Technology to Improve Understanding of Educational Findings

An area of high interest, rapid growth, and great need involves the reporting of test results, along with other indicators of quality, to various publics. We are currently considering data at four levels: state, district, school, and individual. Our goal is to provide ways to manipulate and draw inferences from data.

Our strategy started with the design of a school-level information system, the Quality School Portfolio (QSP). The initial goals of this software were to help school staff make judgments about their students' progress and focus on the continuous improvement of school performance. The QSP software contains three components: the Data Manager, the Resource Kit, and the Reporting Module. The Data Manager allows school personnel to take multiple data sets from external sources and reconfigure them into a single, longitudinal, individual student database. The database is interactive, allowing eight linked levels of query. School staff can query the system in order to disaggregate results to determine how subsets of students have been influenced by particular sets of variables. For instance, we can find out the level of math performance for fourth-grade African American boys who have been in the school for two years, whose reading scores are below the median, but whose attendance is above average. Asking questions about imported data is part of the top-down information and accountability system, since decision makers beyond each school decide on what tests and information will be commonly used.

The second component of the system, the Resource Kit, focuses on local concerns at the school site and provides instruments and strategies to encourage schools to answer local questions of concern. For example, the Resource Kit includes topics such as local curriculum, safety and security, parent involvement, and extended-day programs. Resource instruments and procedures are available so that the school can undertake, albeit in a limited and targeted way, to

answer its own questions and to engage in systematic self-assessment.

The last structural component of QSP is its analytic and reporting capability. Common report formats (for instance, meeting Title I or state requirements) are configured and offer a range of reporting displays, from simple graphs to icons. The reports use a set of core indicators, including many developed by the cross-site team for the Annenberg Challenge. One extension of this project became obvious—program evaluation. We believed that QSP could improve the usual way program evaluation is conducted. By using QSP as a system to collect and export information from schools to program managers or evaluators, the system could, at the same time, give feedback to schools on their current status. QSP could simplify the evaluation burden at the school site, engage participants in the evaluation and improvement process, and meet program managers' needs by summarizing findings across a variety of schools.

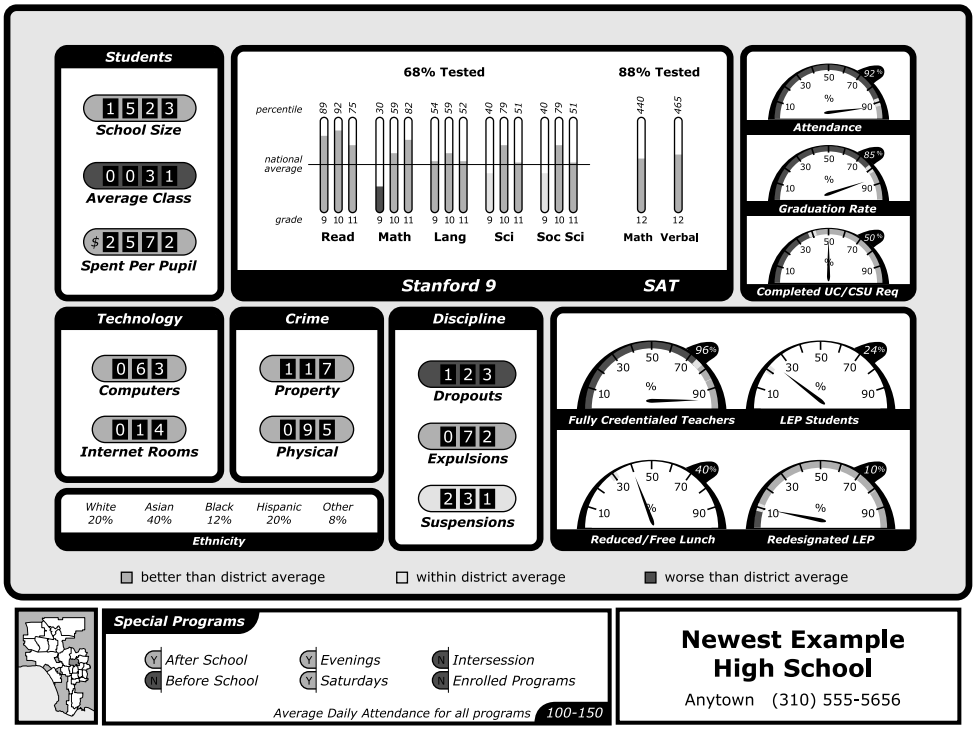
A third modification of QSP is underway. By adding a different interface, small districts, where data analysis and information support may be weak, could use QSP to summarize school performance. Finally, we are working collaboratively with national organizations (the American Association of School Administrators and the National School Boards Association) to help decision makers use the system to report and simulate the impact of potential interventions given the state of findings.

The utility of the QSP system is limited by what is in it (specifically, by the quality and frequency of external information) and by the energy of local users to add data. If the data result from a once-a-year standardized test, then only a small set of principal inferences and interventions is likely to be useful. A second limit of the system is the level of sophistication of its uses: What particular questions are posed to the system by the school community will affect how much performance is improved. We are in the process of working with teams of administrators and teachers to learn how to focus attention on questions that are likely to lead to improvements in school processes and outcomes. The quality of the information in the system and the quality of the questions are linked. Good information is needed for sophisticated questions. Penetrating questions are undermined by inappropriate data.

We also have focused, in another project, on reporting information about schools to the public. Building on work presented in the *Harvard Business Review* in the area of human factors, on efforts of ETS experts like Howard Wainer, and on previous CRESST research, we have approached the problem of increasing the meaning of reports of data on educational quality. One tactic has resulted in a simple redesign of the selection and representation of available indicators of school performance. Here is an example from the *CRESST School Report Card* (see Figure 3), in which the salient issues (what schools can actually change) are highlighted, in comparison to the usual emphasis given to input variables. The metaphor, a dashboard, is peculiarly appropriate to Los Angeles—where everyone drives everywhere. The task is clandestinely instructional.

Through the array of icons and use of color we are trying to build recognition of a new set of conventions in the eye of the reader. Standardized meanings are used for icons: Dials display only percentages, odometers show counts, and gauges depict transformed

Figure 3. CRESST School Report Card

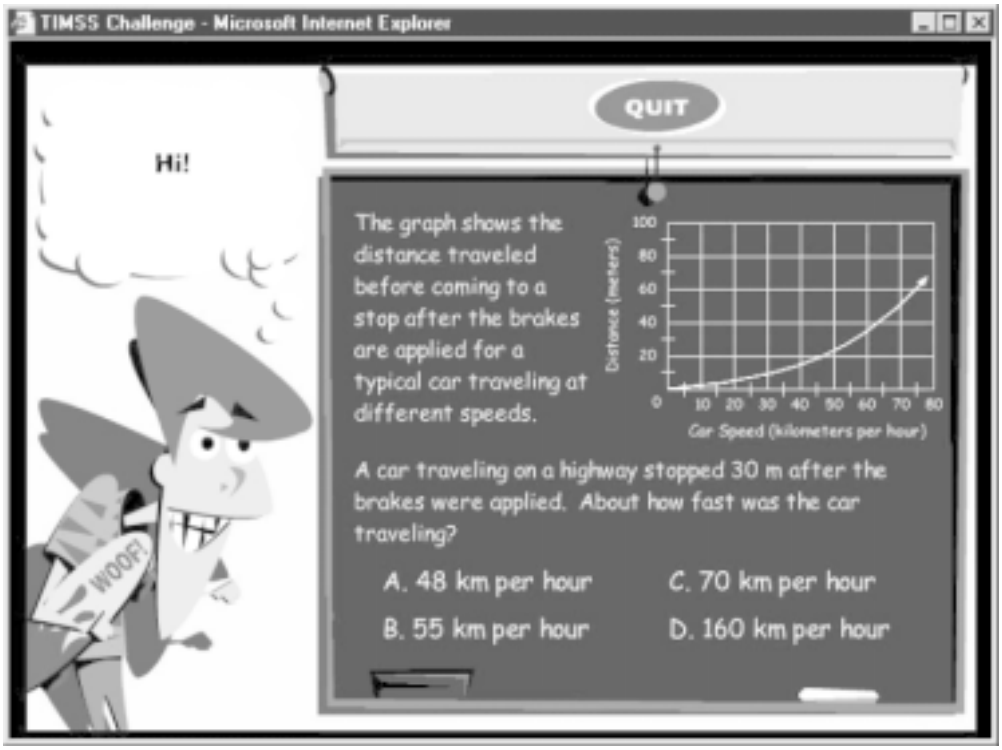


The School Report Card was developed with funding to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) from the Office of the Mayor, City of Los Angeles, Rx for Reading Foundation, contract # 18221. The School Report Card is the work of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) and does not necessarily reflect the positions or policies of the Office of the Mayor, the City of Los Angeles, or the Rx for Reading Foundation.

or otherwise scaled scores. Supplementing the pictures is brief verbal backup explaining the indicators and source of information. Displays are also available in Web form and, we hope, printed in a consumable form, like a telephone book. We are exploring other metaphors to guide representations so as to increase parent understanding of findings as well, including the way a normative or standard benchmark is depicted and understood. Our interest in reporting is focused on improving access to data, whatever it is, and displaying findings in ways that can lead to appropriate inferences given the preexisting mental models of the users.

Our last technology approach—providing tests on the Internet—is intended to make public the types of expectations embedded in regularly given or high-profile tests. In this instance, we created a destination Web site to give people (parents and kids) real experience with the Third International Mathematics and Science Survey (TIMSS) measures (see Figure 4). This example illustrates how technology can open up and expand access to measures of performance, but the technology application is again limited by the nature of the test used—in this case, a multiple-choice measure.

Figure 4. TIMSS Online Challenge



The TIMSS Online Challenge in the CRESST Web site was supported by a U.S. Department of Education grant (National Center for Research on Evaluation, Standards, and Student Testing [CRESST]), having reference number Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The TIMSS Online Challenge portion of the CRESST Web site was designed and produced by Imagistic Media Studios, Inc. The TIMSS Online Challenge is the work of CRESST and does not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education. The Web site of the Third International Mathematics and Science Survey (TIMSS) can be found at http://ustimss.msu.edu.

Practical Limits of Technology

The previous discussion has focused on the use of technology to make substantial improvements in our ability to assess and understand educational quality, principally measured by student achievement. But there is plenty of resistance to even this level of technology application. There is a widespread concern about computer uses in testing because of equity concerns. It is true that the present distribution of computers does not favor poor children, although the ratio of children to classroom computers is rapidly falling. There is also substantial growth among poorer and less educated people in their access to home computing as the cost of a high-end system is now well below $2,000. I believe, however, that along with dramatically reduced hardware costs, the creation and use of tests that require computers will provide a spur to remedy this distribution problem.

A second, commonly heard concern is the lack of teacher preparation in the use of computers in general. Again, I see this as largely a symbolic problem, and if the numbers related to turnover in the system are true, we are about five years away from adding significant numbers of comfortable computer users to teaching

jobs in schools, assuming we can fill the available positions with qualified professionals. For the moment, professional development is needed, but it should be focused on what the curriculum is about and how computers can support those goals, rather than on the area of computer literacy alone.

Educational Testing in America—An Apparently Sensible But Sadly Irrational System

To this point, we have reviewed some approaches that can be used with existing technology to improve the quality of what we test, the efficiency of the tests, and the usefulness of the information to different audiences. I next will sketch out for you some basic inconsistencies in K–12 testing as it now occurs in this country and suggest how we might embark on a serious research and development program to move ahead. Naturally, computer technology fits in here.

What's Wrong With the Testing System?

My thesis, that there is something wrong with our system of K–12 testing, does not flow from the same impulse as many such analyses. It is not developed as a critique of the factory model of education, the one that sees children as outputs and that is a vestige of the industrial era. It does not attack tests and their results as reductionist oddities. It does not compete with the findings of tests developed by the capital letters TESTING INDUSTRY against a sometimes more romantic view of the wisdom and accuracy of classroom teachers' judgments of their students' performance. Last, it will not deny that policymakers have the right and responsibility to demand testing programs that shed light on school progress and real policy options, and that such programs be developed on a schedule shorter than the Pleistocene era.

The usual underbrush now cleared, we can get down to business. My premise is simply this: The testing system as it is conceived and operated in K–12 education appears to serve, but really does not support, the goals that we have for it. Most certainly, it will not meet the expectations of guiding practice and improving learning, and, in fact, it may mislead us. With its current premises and traditions, I'm not sure it can be made to work well. I am not describing the value of any particular tests. I am discussing testing as it has come to be used in public policy.

Specific Problems

What are the problems with the system? They are as much conceptual as practical. First, as we all know, measurement experts have been able to identify discrete uses for tests (see Figure 5). These purposes include tests that are used to certify the performance of the test taker, test information that is used to monitor the education system, test results intended to help select individuals for educational or career opportunities, and test results that provide diagnostic information to encourage better matches between the learner and instructional options. Tests are also used as ways to provide quantitative information, in part to determine the effectiveness of programs and policies.

Over the years, measurement experts have told us that, for the most part, separate tests are needed to optimize each of these purposes. Every set of technical standards for tests developed by the Joint Committee of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (see 1999) has in one way or another cautioned users about blindly appropriating tests intended for one purpose and using them for an entirely different purpose. At

minimum, users are enjoined to develop an evidentiary argument, of the sort that Sam Messick, Bob Mislevy, Linda Steinberg, and Russell Almond have eloquently described, as a way of justifying the application to a different purpose of a test developed for a particular use. For the most part, many of the arguments about keeping tests in the silos of particular purposes (see Figure 5) have been made on technical grounds. In an example we all understand, the degree of certainty with which we need to recommend that a student be admitted to a particular university has to be higher than we would require if the student's performance, combined with scores of other students, is used to report on the general progress of a statewide program.

But this purpose-by-test link creates enormous problems for the K–12 educational world. First, unlike measurement people, the policymaker (think "a presidential candidate") or the public (think "the reporter from your favorite newspaper") is not concerned with nor even much interested in the boundaries among test purposes. In contrast, the technical community is talking about indicator or testing systems rather than considering one test at a time. In a system, multiple measures of student performance and multiple indicators of system performance address different kinds of performance and different inferences. The assumption is that taken together, the system information leads us to better understanding and ultimately to higher levels of attainment or rates of improvement.

This shift requires experts to think about system validity, not just the validity of a single measure. I must emphasize that this is a different order of problem for both measurement and policy experts, and so far, neither group has been up to the job. What has brought the system focus into the spotlight is that tests have become the major part of the reform agenda. Claims are regularly made (and believed by the public) that tests can do many important things in education. Test content, incentives, and sanctions on results are claimed to operationalize student content and performance standards, provide instructional guidance, motivate students, identify good programs, point out institutions that need to be fixed, and so on. Furthermore, the policy community has decided that we don't need six different tests to perform these miracles. We

Figure 5. Purposes of Tests

- System monitoring
- Accountability of schools
- Program evaluation
- Instructional improvement
- Communication
- Certification, promotion, retention
- Admission/selection
- Assignment
- Teacher and principal accountability
- Diagnosis

might only need one or two—one test that measures content standards, or what passes for curriculum, and the other to give us a comparison (their compromise on the conflicting desire for both local control over education and uniform comparisons).

Certainly, a multipurpose test is a heretical idea to many measurement experts. What if we could create a design theory that would allow conceptually linked tests to provide information to serve multiple purposes? We know we can do some of it. An easy example is looking at purposes connected to system monitoring and instructional improvement. We have designed tests in Los Angeles using the same core design models and content assumptions that are reported differently for teachers and for the public. This by itself is nothing new. What is difficult, and where the multiple purpose idea falters, is the assumption that teachers and others can use the assessment results formatively, that they can learn from the results and create useful alternative approaches, enabling them to show subsequent impact on the system without serious guidance or intervention. For instrumental improvement to occur over the long term, three assumptions must be met: (1) that all tests in a system attempt to measure the same domains, (2) that the domain definitions and characteristics are made public in sufficient detail to be actionable, and (3) that improvement on a test for one purpose will transfer to gains on assessments intended for other purposes.

In the emerging system, supported by most states with incentives from Title I, a rhetorical premium has been placed on multiple measures of performance. The intention was twofold: first, to provide different ways to assess performance to insure some breadth; second, to provide a commercial space for existing test publishers to perform. Yet the measures have been divided by widely varying domain definitions, inarticulate domain descriptions, and no clear path for action.

As an example, let's consider the common case where a state has its own performance-based tests, uses a standardized test, and participates in the National Assessment of Educational Progress (NAEP). The assumption about these multiple instruments all measuring the same domain is incredibly weak. It is only recently that the domains have been identified through the use of content standards, and for almost every case, the domains specified by such standards are too amorphous by themselves to guide careful design of one measure, let alone provide a crosswalk linking one measure to another. In fact, many extant tests have been adopted by a process that retrofits an existing test to newly developed standards (see the California accountability system, for example). Take an example: "Apply estimation procedures to real problems." When there is only a general-level description to guide test developers and teachers, it is no surprise that inferences and products will vary dramatically. How likely is it that test developers and classroom teachers would infer the same class of examples? They can't and haven't.

Perhaps an analogy or two will help, and it is for the general public rather than for measurement experts. When we talk about tests, many parents believe they know what we mean. We've all experienced them, perhaps even recently under the auspices of the Motor Vehicle Department. But then, our

common understanding quickly falls apart. Think for a moment about a currency system. We all know that the U.S. system uses coins and dollars in different denominations—1, 5, 10, 20, 25, 50. This is also true of the Italian lira. Most people think performance on different tests about the same topic should be exchangeable. We know this idea would get us in trouble if we thought a note with a 20 on it could buy us the same commodity in Italy as in the U.S. But the exchange rate for tests nominally in the same domain but used for different purposes is not shared; it is not even known. Yet, the public makes inferences as if one test score is about the same as another.

Second, there is some evidence to show that growth is predictable among measures, but little data to show that growth on important measures can be attributed to changes in instructional practices or initiatives under a school's control. Test results among different measures will naturally co-vary at the outset. Smarter students do well; less prepared, less motivated students do poorly. Some improvement occurs, because initial performance is very low across the board and because of motivational effects. But, can a system based on tests derived from different interpretations of a standard sustain systematic growth? I'd guess not. Even now, differences among measures lead to interesting but potentially flawed interpretations. For example, proportions of students scoring in a top category on NAEP and on a state assessment may differ. Is that a problem? Which should be the criterion? Score differences might be attributable to greater sensitivity between the state test and classroom instructional practices compared to NAEP. For some state tests, just the opposite might be true. Higher scores do not necessarily mean the test is less rigorous; the test might be reinterpreted as a more instructionally sensitive test in a state making great strides. High stakes on one test affects the entire system. The system becomes deformed for school personnel and for students. And high stakes is clearly the trend, whether it is at school exit, or transition, or just part of the reaction against social promotion.

When high stakes kick in, the lack of publicness and of explicitness of test attributes lead teachers, school personnel, parents, and students to focus on just one thing: raising the test score by any means necessary. The domains of learning and knowledge are often left behind in favor of meeting a target by pushing up scores. Because people are smart and understand sanctions, they will adopt whatever options are available: giving students practice on like items and similar formats, giving practice on the items themselves, providing illegitimate "help" for students, selectively excluding low-performing students (through a variety of increasingly subtle techniques), and even recruiting high-performing students. Perhaps there will be a student draft similar to the one used by the NFL. Rather than conceiving of the tests as instruments to detect changes in real learning of a curriculum, the tests become the singular object of focus. There is really no way that current tests can simultaneously be a legitimate indicator of learning and an object of concerted attention. The answer to this paradox from the measurement community is to keep purposes separate, if we can. But what if we can't? Can you devise a test that can both provide information relevant to a variety of purposes, including accountability, and at the same time provide real guidance toward improved instructional practice and that will be sensitive to growth? More likely, the measures with high stakes

will have salience. One option is to use course-based examinations. But these may operate against a desire for comparisons and lead to the arbitrary restriction of content.

Technology on a White Horse

It is true that I grew up in Hollywood, in the shade of the palm trees and the old movie studios, going to Saturday and Sunday double-feature matinees. At my favorite theater, the Hitching Post, we dressed appropriately (In those unenlightened times, I could dress only as Dale Evans.) and checked our cap pistols (again unenlightened) before we entered to see Roy Rogers and Hopalong Cassidy confront all problems. Solutions invariably rode up on a white horse, materializing out of the dust. And so all my life I have waited, continuing to squint at the horizon, scanning for the good guys. Waiting doesn't work for me any longer. Assessment is becoming potentially more powerful as demands for accountability are unleashed without real understanding of what is desired and what should count as appropriate measures. The lack of precision with which we define what we want to occur, how we teach, and how we measure, even for different purposes, may make it difficult for schools to respond appropriately. In common educational talk, this problem is rooted in the underspecification of content and cognitive demands of learning, results in misalignment among components of the educational system, and inhibits sensible design and interpretation of measures of learning. The simple fact is that we have quantitative data in black hats masquerading as the good guys. Where is that white horse? We need a good way to hold schools and students accountable for their accomplishments.

Knowing What Is Wanted, What Should Be Taught, and What Should Be Measured

An important step toward a solution requires an understanding of the conceptual and scientific basis of student learning. Rather than continue to patch and accrete more and more incompatible solutions, we need a clear, fully articulated, descriptive system. A major scientific effort is needed to specify the goals, instructional requirements, and potential measured outcomes of learning. A linked, exhaustive database of detailed content and well-specified cognitive demands must be created. Ultimately, we should attempt to develop a database that gives precise, coded descriptions of the full range of student learning options. Yet, once even a preliminary version of a domain is completed, we can test the benefits directly.

Creating a fundamental, comprehensive (and, of course, revisable) set of descriptions would enable us to use a rational basis for the selection of goals and learning options, including the design and use of software applications and classroom instruction, and the design and interpretation of tests. Gaps would be evident, and linkages among tests for different purposes could be easily documented.

In Summary

1. What if we took it as a goal to map, as explicitly as possible, the domains of school performance? The map would reflect a number of key features of learning. Suppose we specified in detail the range of cognitive demands for school learning, giving examples at different grade levels. This description of possible learning tasks would include general prior knowledge requirements as well as task requirements.

2. What if we detailed academic and skill content at a very specific level? We would explicate declarative, procedural, and system knowledge to be the focus of academic learning. Prior knowledge requirements, including the particular facts, procedures, skills, and content, would be included.

3. What if we described linguistic requirements and special task demands of assessment settings in a developmental framework?

To begin this extensive task, I propose that we adopt the metaphor and development practices of the genome mapping project. Although the difference between biological domains based on DNA sequencing and an artificial one such as learning may seem enormous, there are actually a number of points of comparison.

First, the goals of both projects are comparable, although the learning map is more difficult and requires some arbitrary definitions. In the learning map, we seek the full articulation (or set of instructions) of elements and relationships; we wish to describe learning domains completely. Second, we have to develop a useful system to conceptualize elements and model their results, so that the findings can be verified and so that we will develop applications to improve our ability to solve human problems. In the case of testing, we will be able to create and interpret measures by understanding in great detail what aspects of the domain they represented. Tests for different purposes could be developed from sampling profiles from the mapped domains. Growth on one test might or might not predict growth on

another test of known characteristics. Instead of only global correlations used to link tests, all tests would be connected explicitly to the content and cognitive domains they measure and would vary technically as demanded for the decisions flowing from them. An additional similarity to the genome project would be the use of differentiated expert teams, for example, content, cognitive, linguistic, and neuroscience. Another relevant comparison is that there would undoubtedly be competition in approach as well as positive commercial consequences for the mapping project groups. The project would also be costly, but could be scoped to achieve particular benchmarks. I assume funding would come from both private and governmental sectors.

A mapping R & D program is possible, and we have made some progress in thinking about strategies. Two obvious approaches should be undertaken simultaneously. One strategy would be working from the existing stuff that now comprises education, its standards, its conceptions of learning, its test content and format, and its instructional materials. This bottom-up approach would require that we compile domains. A compelling, parallel strategy is a top-down, theory-driven approach, where we would be trying to design domains more systematically. On a much smaller scale, specifically in an analysis of expert content domains to design expert systems, Wenger (1987) identified some useful differences between the approaches. Wenger noted that tradeoffs would need to be made between the two strategies. If we use a bottom-up approach, we would compile knowledge from existing sources (imagine the analysis of existing sets of items, such as those on eighth-grade NAEP and TIMSS mathematics tests). The result would be a domain that is more redundant, arbitrary, and isolated. It would be more efficient in the short run, but would do a poor job of supporting adaptability. It would be more difficult to modify and would not be readily useful for transfer. In contrast, a top-down approach based on designed content domains would be more independent. Domains would permit the expression of explicit relationships. They would be more generally useful in the long run, because they would support transfer, be modifiable, and be adaptable to different contexts and learners. These contrasting development strategies should not be mutually exclusive. Most important is our resolve to find a strategy to change the current system so that it can meet its intentions and lead to real improvement of children's learning.

Let me also point out that there are and have been initial efforts in this regard, but not explicitly to create a rational system of tests serving different purposes. Minstrell (2000) in science and Wiley and Haertel (1989) in mathematics have made inroads in the content part of this area, and databases of content in particular subject matters may suggest particular topics for initial consideration.

What has been done? A group of experts has agreed to undertake the task of building the "knowledge sphere." The group has participation by content experts, cognitive and other learning psychologists, instructional specialists (teachers), and psychometricians. A preliminary international project is underway to map a limited set of content, cognitive, and linguistic domains. I am happy to report that CRESST and ETS are collaborating in this work.

The ultimate goal would be to be exhaustive and inclusive in content and potential cognitive demands, rather than to reach politically driven consensus. Considerable practical issues have already arisen using a bottom-up, development strategy (classifying extant items). These concerns include proprietary status, valuing effort, publication, and extensibility to other domains, among others.

Where does technology fit in? It is, of course, the white horse. First, the task can be greatly simplified by using strategies from computer information science to classify and do first-level depictions of items (using software strategies such as search engines and digital classifiers). Second, relationships among the primitives (the lowest unit of cognitive demand or content knowledge) can only be modeled using technology. Third, the retrieval and analysis system can be made available, through various mechanisms, to the public-at-large, in non-technical terms. Fourth, the automation of test design and development can be facilitated.

New Utility

The utility of this project would be twofold. First it would be of important scientific value and lead to new theories of knowledge classification, test design and interpretation, and teaching. Second, its practical applications would seem to be myriad and particularly relevant to my topic today. We would be able to model clearly the types of content and cognitive demands included on tests for different purposes at different age levels. We could obtain a quantifiable measure of the conceptual, cognitive, and content relationships among measures. We would have a basis for the systematic design of instruction, software, and professional development, as well as a classification system that would be useful to parents, test designers, and users of out-of-school learning materials and systems. We would provide a warrant to those charged with instruction and school improvement to address their problems without the conceptual handcuffs they now wear. Because the resulting system would be open, both in access and in its extensibility, the quality of school outcomes and public understanding of education would be strengthened. Technology applied in the service of understanding the learning we want will help us fix the presently unfixable—the deep validity problem at the heart of our testing system.

References

Aguirre-Muñoz, Z. (2000). *The impact of language proficiency on complex performance assessments: Examining linguistic accommodation strategies for English language learners.* Unpublished doctoral dissertation, University of California, Los Angeles.

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., Niemi, D., Novak, J., & Herl, H. E. (1992). Hypertext as a strategy for teaching and assessing knowledge representation. In S. Dijkstra, H. P. M. Drammer, & J. J. G. van Merriënboer (Eds.), *Instructional models in computer-based learning environments* (pp. 365-384). Heidelberg: Springer-Verlag.

Chung, G. K. W. K., & Baker, E. L. (in process). *Assessing problem solving skills with model-based simulations.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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.

Haertel, E. H., & Wiley, D. (1989). *Poset and lattice representations of ability structures: Implications for test theory.* Presentation to the National Academy of Education, New York.

Herl, H. E., O'Neil, H. F., Jr., Chung, G. K. W. K., & Schacter, J. (1999). Reliability and validity of a computer-based knowledge mapping system to measure content understanding. *Computers in Human Behavior, 15,* 315-334.

Herl, H. E., O'Neil, H. F., Jr., Schacter, J., & Chung, G. K. W. K. (1998). *Assessment of CAETI STS1 technologies* (CAETI Deliverable to ISX). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Minstrell, J. (2000). Student thinking and related assessment: Creating a facet-based learning environment. In N. S. Raju, J. W. Pellegrino, M. W. Bertenthal, K. J. Mitchell, & L. R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (Report of the Committee on the Evaluation of National and State Assessments of Educational Progress, Commission on Behavioral and Social Sciences and Education, National Research Council; pp. 44-73). Washington, DC: National Academy Press.

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: Lawrence Erlbaum Associates.

O'Neil, H. F., Jr., Herl, H. E., Chung, G. K. W. K., Bianchi, C., Wang, S. L., Mayer, R. E., Lee, C. Y., Choi, A., Suen, T., & Tu, A. (1998). *Final report for validation of problem-solving measures* (Final

deliverable to Statistics Canada). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

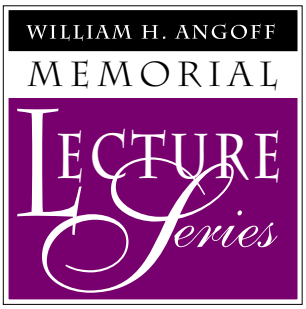
O'Neil, H. F., Jr., Wang, S. L., Chung, G. K. W. K., & Herl, H. E. (2000). Assessment of teamwork skills using computer-based teamwork simulations. In H. F. O'Neil, Jr., & D. Andrews (Eds.), *Aircrew training and assessment* (pp. 245-276). Mahwah, NJ: Lawrence Erlbaum Associates.

O'Neil, H. F., Jr., Wang, S. L., Chung, G. K. W. K., & Herl, H. E. (1998). *Final report for validation of teamwork skills questionnaire using computer-based teamwork simulations* (Final deliverable to Statistics Canada). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Salas, E., Dickinson, T. L., Converse, S. A., & Tannenbaum, S. I. (1992). Toward an understanding of team performance and training. In R. W. Swezey & E. Salas (Eds.), *Teams: Their training and performance* (pp. 3-30). Norwood, NJ: Ablex.

Schacter, J., Herl, H. E., & Chung, G. K. W. K., Dennis, R. A., & O'Neil, H. F., Jr. (1999). Computer-based performance assessments: A solution to the narrow measurement and reporting of problem-solving. *Computers in Human Behavior, 15*, 403-418.

Wenger, E. (1987). *Artificial intelligence and tutoring systems: Computational and cognitive approaches to the communication of knowledge*. Los Altos, CA: Morgan Kaufmann.



POLICY INFORMATION CENTER
Educational Testing Service
Princeton, NJ 08541-0001

88502-006067 • S110M4 • Printed in U.S.A.
I.N. 990040