



FROM THE PRESIDENT: NCHE AND NCME ANNOUNCE 4TH EDITION OF *EDUCATIONAL MEASUREMENT* UNDERWAY WITH ROBERT L. BRENNAN, EDITOR

Most of an NCME president's duties involve overseeing important activities that occur routinely in each president's term (e.g., annual meeting and appointments of committees). In addition, however, the one-year presidential term allows just enough time to implement one or two additional goals that can advance the organization and the measurement field. My goal in 2002-2003 has been to re-establish the partnership between NCME and the American Council on Education's National Council on Higher Education (NCHE) and to launch the production of the 4th Edition of *Educational Measurement* under the co-sponsorship of the two organizations. Our past-president, H.D. Hoover, had initiated a contact with NCHE, and I am pleased to report progress on this project.

A brief review of the history of this publication seems in order. In 1951 the American Council on Education (ACE) first published *Educational Measurement*, edited by E.F. Lindquist, to reflect the organization's interest in appraising the quality and usefulness of instruments used in postsecondary education. An advisory board identified critical topics and invited experts in various areas of measurement to author different chapters. The result was a much acclaimed and widely used reference. As the contents became dated, ACE received numerous requests to update this landmark work. Consequently, ACE invited the American Educational Research Association to assist with appointment of an advisory board for another edition. Robert L. Thorndike was editor for the 2nd edition, published in 1971. To illustrate the richness of the content of this volume, consider that the now-classic "Angoff" method for standard setting was proposed here in a mere footnote to a chapter on scales and norms. By the 1980s, there was again demand for updating this publication. The 3rd (and most recent) edition appeared in 1989, under the editorship of Robert L. Linn. Among its wealth of ideas, this volume contained Samuel Messick's classic and provocative chapter on validity. NCME was a collaborating sponsor with NCHE for the 3rd edition. Now, for over half a century, graduate students in measurement have immersed themselves in one or more of these editions as they prepared for qualifying examinations, and generations of measurement scholars and practitioners have turned to this source repeatedly to confirm their decisions or seek new solutions. These volumes are perhaps the most revered works in our field.

Against this historic backdrop, it gives me great pleasure to announce that NCME will serve as co-sponsor with the NCHE for an upcoming 4th edition of *Educational Measurement* with Robert L. Brennan as editor. Dr. Brennan is E.F. Lindquist Professor at the University of Iowa and a past-president of NCME. The intended publisher of the volume will be Greenwood Press. Given the enthusiasm of the publisher and NCHE for this project and their commitment to quality reference publications, I am confident that the tradition of excellence for *Educational Measurement* will continue into the 4th edition and delighted that NCME can play a leadership role.

Linda Crocker

LEGAL CORNER: *Alternate Assessment*

S.E. Phillips

The new NCLB Act Implementing Regulations for Standards & Assessments require states to provide alternate assessments for students with disabilities whose IEP teams have determined cannot participate in the state's regular assessments with *appropriate accommodations*.¹ Alternate assessments must initially be provided in the subject areas of language arts and math; science must be added beginning in the 2007-08 school year.²

Appropriate accommodations remove the effects of extraneous factors associated with the disability that are not intended to be assessed while preserving the knowledge and skills intended to be measured. For example, a visually impaired student might be tested in math with a large print test booklet because reading small print is not part of the math skills intended to be tested. However, if the math test included estimation and computation questions, providing a calculator to a learning disabled student would substitute calculator literacy for the intended skills of approximating and application of algorithms that were intended to be measured. Such an alteration in testing conditions would more appropriately be termed a *modification*, which may be considered one type of alternate assessment.

Besides modification of a regular assessment, other types of alternate assessment include: (1) a test in the same academic subject designed for a lower grade level where the tested skills match the instruction specified in the student's IEP (also called out-of-level testing); (2) an individually administered standardized achievement test covering the same academic subject; (3) a test constructed to measure the specific academic curriculum prescribed by an individual student's IEP; (4) a portfolio of student academic work in the subject area; (5) a checklist or rating scale of prerequisite behaviors that must be learned prior to beginning instruction on the tested subject matter; and (6) an evaluation of progress in achieving and maintaining nonacademic skills specified in the student's IEP. In providing for such alternate assessments, the intent of the NCLB Act, consistent with the IDEA and the ADA, is to include students with disabilities in regular, on-grade-level instruction and assessment to the maximum extent possible and, when students with disabilities cannot do so, to ensure that schools are held accountable for systematic evaluation of their progress in meeting IEP goals related to state standards in the tested subject.

The NCLB and its regulations state conflicting goals which have not yet been totally reconciled. On the one hand, all students are to be assessed at grade level based on the same state content standards. On the other hand, alternate assessments are to be provided when students with disabilities are unable to participate in regular state assessments with appropriate accommodations. By definition, students taking alternate assessments are being tested on different content and skills than students taking the regular assessment because the

content has been modified, is at a lower grade level, or involves nonacademic behaviors. The provision of alternate assessments meets the policy goal of including all students in the assessment system but does not meet the policy goal of holding all students to the same grade-level content standards. However, it would clearly be unfair to require students with disabilities to be tested on content and skills different from those prescribed by their IEPs.



The competing policy goals of *holding all students to the same standards* and *assessing all students* are partially reconciled by: (1) expanding state standards to include prerequisite and developmental skills necessary for achieving each grade level content standard; and (2) using some portion of the 5% of students the law allows not to be tested (which also includes absentees) to exempt some students with disabilities from academic tests (substituting appropriate nonacademic behavioral assessments). However, although such actions allow students with disabilities to be assessed with the most appropriate tests consistent with their IEPs, the problem of reporting remains. When students take different tests, including regular statewide assessments administered with modifications, the resulting test scores are not comparable to scores from the regular statewide assessment. Thus, it is not valid to aggregate the scores from these disparate measures into a single statistic purporting to indicate the percent of students who have met state standards.

continued

NEWSLETTER ADVISORY BOARD

BETTY BERGSTROM, Promissor
GREGORY CIZEK, University of North Carolina, Chapel Hill
JOAN HERMANN, CRESST/UCLA
SHARON LEWIS, Council of the Great City Schools
DUNCAN MACQUARRIE, Tacoma Public Schools
KAREN MITCHELL, SRI International
LORA MONFILS, Graduate Student – Rutgers University
S.E. PHILLIPS, Consultant
NAMBURY RAJU, Illinois Institute of Technology
LAWRENCE RUDNER, ERIC/University of Maryland
STEVE SIRECI, University of Massachusetts, Amherst
LISA F. SMITH, Kean University
JON TWING, NCS Pearson
DENNY WAY, Educational Testing Service
DOUGLAS F. BECKER, EDITOR, Riverside Publishing

Send articles or information for this newsletter to:

Douglas F. Becker Phone: (800) 767-8420, ext. 7006
Riverside Publishing Fax: (630) 467-7126
425 Spring Lake Drive e-mail: douglas_becker@hmco.com
Itasca, IL 60143

The *NCME Newsletter* is published quarterly. The *Newsletter* is not copyrighted; readers are invited to copy any articles that have not been previously copyrighted. Credit should be given in accordance with accepted publishing standards.

¹ 34 C.F.R. Part 200, §200.6(a)(2)(i); Authority: NCLB, 20 U.S.C. 6311(b)(3).

² *Id.* at §200.6(a)(2)(ii).

Alternate Assessment, continued

Similarly, when students with disabilities have taken different tests under different conditions (e.g., regular math computation test with a calculator, math computation test designed for students two grades below the student's grade placement, oral test of counting skills, etc.), those test results also cannot validly be aggregated to produce a meaningful result. The most that could be said for such aggregations is that they would indicate the percent of students who had met whatever standard (at or below grade level, academic or nonacademic) that had been set for them in their regular education, Title I, or IEP program. Moreover, this interpretation does not answer the policymakers' questions about how many students have achieved state standards in each academic subject at each grade level.

A compromise position may be to separately report: (1) the percent of students subject to regular assessments who met state standards (with or without valid accommodations); (2) the percent of students with disabilities taking alternate academic tests who met the standards specified for them in their IEPs; and (3) the percent of students with disabilities who met the nonacademic goals specified for them in their IEPs. Each of these statistics could be included in the state accountability system. In addition, it may be helpful for states to provide incentives for instructing and assessing students with disabilities at the highest academic levels possible.

The alternate assessment requirement is more than the development of a single test. The alternate assessment administered to a student with a disability must be matched to the specific instruction that student is receiving. Different students will require different alternatives depending on what their IEPs prescribe. As a state begins planning for alternate assessments, consideration must be given to the various types of alternate assessments that will be needed. It may be relatively straightforward to modify regular assessments. Assessments may already be available for students who need to test one or two grade levels below their grade placement. Rated portfolios or state-developed observations/ checklists may be appropriate for the most severely disabled students who are focusing on enabling skills or nonacademic behaviors. English language proficiency tests (reading, writing, and speaking) may be needed for ELLs who are not yet proficient in English. In short, alternate assessment is really a system of assessments appropriate for different types and degrees of disability and consistent with IEP requirements.

Once a state has decided which tests or assessment procedures will be administered as alternatives to which types of students, the difficult task remaining will be to determine how to incorporate that information into the state's accountability system. As long as each student with a disability is assessed, and nearly all are assessed on academic skills in reading, math and later science, the states should have some flexibility in determining how those results will be reported to provide valid and meaningful information about attainment of standards by students within the state. The state must strive to provide information that is accurate and not misleading while providing all students with disabilities the opportunity to

achieve grade-level academic standards when the IEP team concurs. The challenges for the state are: (1) to raise expectations for students with disabilities while retaining enough flexibility for IEP teams to prescribe appropriate goals for each disabled student, and (2) to ensure that schools are held accountable for the progress of their disabled students.

THE NO CHILD LEFT BEHIND ACT: REGULATORY GUIDANCE

Duncan MacQuarrie, Tacoma (WA) Public Schools

The Secretary of Education, on July 5th, issued "final regulatory guidance" for implementing the standards and assessment requirements of Title I, subpart A, of the Elementary and Secondary Education Act (ESEA) as amended by the No Child Left Behind (NCLB) Act of 2001. In addition, on August 6th, the Secretary issued proposed rules that address, among other topics, those sections of subpart A dealing with state and district participation in NAEP, state accountability systems, and the requirements for achieving "adequate yearly progress" on state assessments.

Some commentators on the rules have championed them as positive examples of regulatory flexibility, while others see the Department of Education as providing too little guidance in critical areas. The final and proposed rules do make helpful clarifications of the language contained in NCLB, sometimes by merely reorganizing the statutory provisions into a more coherent statement of the requirements. However, there are two particular troubling topics in the regulations regarding testing. In one case the vagueness of the language in the statute begs for clarification and guidance and the Department has chosen to offer none. In another, the Department's apparent indifference to a reasonable view of individual differences has resulted in narrow and arbitrary regulation.

The first issue of concern is a lack of guidance in an area where providing it would have assisted states to implement the legislation in a consistent manner. NCLB requires the public reporting of disaggregated achievement data by a variety of subgroups, e.g. race, ethnicity, gender, disability status, migrant status, English proficiency, and status as economically disadvantaged. These same disaggregated results will play a critical role in the determination of whether or not a school is identified and labeled as a "failing school" under the new federal school accountability system. Statutory language does make provisions to exempt states from the requirement to report or use disaggregated data "in a case in which the number of students in a category is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student." The guidance in the rules for protecting personally identifiable information is made clear by reference to requirements of the Family Educational Rights and Privacy Act of 1974. However, the Department has chosen to leave it up to each state to "determine and justify in its State plan the minimum number of students sufficient to yield statistically reliable information..."

Certainly there is no single minimum number that will assure statistical reliability. Some in the measurement community argue that the assumptions and procedures of statistical significance testing should guide our decisions in this area. Others would depend much more on test reliability and decisions guided by concepts of error of measurement. Lacking guidance from the Department, State Departments of Education will need thoughtful assistance from the measurement profession as they craft reasonable standards for judging what constitutes statistical reliability when reporting and using standard-based test results for the various subgroups. This is an extremely important task. And each state must select their standard and provide a justification no later than January 31, 2003.

The second issue concerns an overly prescriptive regulation and the Department being all too willing to provide a number, "0.5 percent" to be exact. The NCLB Act makes it clear that the academic progress of virtually all children in a state must figure into that state's accountability equation. And, this progress is to be judged against one fixed set of challenging state academic standards. There are few exceptions. The proposed rules issued on August 6th do make provision for "students with the most significant cognitive disabilities" to take alternate assessments with associated performance standards reflecting "professional judgment of the highest learning standards possible for those students."

This rule is consistent with the long-standing practice of developing individual educational plans (IEPs) and related assessments for students with cognitive disabilities. However, for purposes of calculating a state or district's adequate yearly progress toward the goal of all students meeting the state standards, the Department has restricted the use of these alternate achievement standards to no more than 0.5 percent of all students in the grade assessed. Under these rules Tacoma, Washington's second largest and most urban district, would be allowed to consider the state's challenging academic standards as educationally inappropriate for no more than 13 children at the typical grade level. For the entire state the number would be only 375 children at a grade level.

The Department, in providing justification for this restrictive standard, argues that it is "(b)ased on current prevalence rates of students with the most significant cognitive disabilities." In other words, children need to be more than two and one-half standard deviations below average in cognitive ability before we can consider the "challenging academic content and student academic achievement standards" mandated by NCLB to be inappropriate for them. In this case, where the Department has been willing to provide guidance about a number, they have gotten it terribly wrong. And those most likely to be harmed by this rule are the very children and families that the administrative rules should protect.

We, as members of the educational measurement community, must make our voices heard on these critical issues, providing guidance and assistance whenever and wherever the opportunities present themselves.

NO CHILD LEFT BEHIND TEST INCREASES UNLIKELY TO BE MET

A new report from the National Center for Research on Evaluation, Standards, and Student Testing at UCLA raises concerns about the way in which states may choose to implement growth targets required by the No Child Left Behind Act of 2001 (NCLB). The Act requires that all students must reach at least the proficient level on their state exams by the year 2014. Each year, schools must increase their percent of students at the proficient or above level by an amount sufficient to move from their 2002 baseline level to 100% over the next 12 years.

"The policy maker expectations almost certainly exceed the ability of schools to make this sort of progress," said one of the study's authors, Robert Linn, a CRESST co-director and current president of the American Educational Research Association. The "proficient" level in most states is set very high.

"One sticking point," said co-author Eva Baker, "is that the new law requires steady increases from year to year between now and 2014." A school can increase sufficiently to meet the target in one year, but if it slips below the target in subsequent years it can fall into the "needs improvement" category. Baker is a CRESST co-director and chairs the National Research Council's Board on Testing and Assessment.

Linn added that scores at the school level are very volatile from one year to the next, making substantial and steady annual progress difficult. Measurement and sampling error contribute to the stability problem. The researchers predict that a substantial number of schools will be incorrectly identified as needing improvement as a result of the volatility of school level scores.

The authors also found that the goal of having all students reach the proficient or above level by 2014 was highly dependent on the difficulty of the state test or where the achievement levels were set. Their comparisons showed that in the year 2000, only states such as Texas with basic skills tests and relatively modest standards had a large enough proportion of their students meeting standards for the *100% proficient or above* goal to be within the realm of what might be possible in 12 years. However, Texas is revising its standards as well.

In states with more demanding tests and more stringent standards it is not unusual to have fewer than 2 students in 5 scoring at the proficient level or higher. For example, in 2000, only 30% of Maryland students scored proficient or above on their state test, the Maryland School Performance Assessment Program.

To verify the Texas and Maryland results, the CRESST researchers compared the tests to the National Assessment of Educational Progress. In Texas, only 25% of students achieved at the NAEP proficient level or above, whereas in Maryland, the figure was 28%. Under the new law, NAEP will be used as to validate performance increases for all states receiving federal Title 1 funds.

The problem is exacerbated by the requirement that all subgroups of students reach the proficient or above level and increase at the required rate, including students with disabilities and English Language learners. If the overall school percent proficient increases enough to meet the adequate yearly progress target, but students with disabilities or English language learners fall short of the target, the school will have failed to meet the target.

The NCLB goals are laudable, conclude the researchers, but even the basic NAEP achievement level would be a challenging goal for most schools to reach in 12 years. The authors recommend a switch from the use of achievement levels, which have had a history of challenges, to more traditional methods for reporting progress.

The full copy of the report may be found on the CRESST web site at <http://www.cse.ucla.edu/>.

Reference

No Child Left Behind Act of 2001, Pub. L. No.107-110, 115 Stat. 1425 (2002).

BEATING THE ODDS II: A CITY-BY-CITY ANALYSIS OF STUDENT PERFORMANCE AND ACHIEVEMENT GAPS ON STATE ASSESSMENTS

The Council of the Great City Schools has prepared this study, *Beating the Odds II*, to give the nation another look at how inner-city schools are performing on the academic goals and standards set by the states for our children. This analysis examines student achievement in math and reading through spring 2001. It also measures achievement gaps between cities and states, African Americans and Whites, and Hispanics and Whites. Finally, the report looks at progress. It asks two critical questions: “Are urban schools improving academically?” and “Are urban schools closing achievement gaps?”

In general, *Beating the Odds II* found encouraging evidence for a second year that the Great City Schools have made meaningful gains in math scores on state assessments. The study also found gains in reading and preliminary evidence that gaps may be narrowing. The findings in *Beating the Odds II* are preliminary and leavened with caution, as they were when we first published these data last year. The nation does not have an assessment system that allows our questions to be answered with certainty.

Still, the data from this report indicate that answers are emerging and that urban education may be establishing a beachhead on the rocky shoals of school reform. Some data look better than others. Progress in math is different from that in reading. Trend lines are not the same from one city to another. Not all grades have improved at the same rates. Not all gaps are closing. But the data indicate progress.

This report is the nation’s second look at how its major city school systems are performing on the state assessments devised to boost standards, measure progress, provide opportunity, and ensure accountability for results. Data are presented on 57 city school systems, city-by-city, year-by-

year, and grade-by-grade on each state test in mathematics and reading.¹ These systems are located in 35 states. Data are also reported by race/ethnicity in cases where the state reports it publicly.

Every effort was made to report achievement data in a way that was consistent with the new *No Child Left Behind* legislation. This was not always possible, however, because most states have not yet updated their reporting. *Beating the Odds II* uses the percentage of students above “proficiency” wherever available, however.

Eight major findings about academic achievement in urban schools emerged from the Council’s study:

- Finding 1: Mathematics achievement has improved in urban schools.
- Finding 2: Gaps in math achievement in urban schools may be narrowing.
- Finding 3: More urban school districts showed math gains in 2001 than in 2000.
- Finding 4: Urban school achievement remains below national averages in math.
- Finding 5: Reading achievement in urban schools has improved on state tests.
- Finding 6: Gaps in reading achievement in urban schools may be narrowing.
- Finding 7: More urban school districts showed reading gains in 2001 than in 2000.
- Finding 8: Urban school achievement in reading remains below national averages.

The report also shows important demographic and financial data. Included are enrollment data by race/ethnicity, poverty rates, percentages of English language learners, and average per pupil expenditures. Statistics are also presented on student/teacher ratios and average school size. Finally, changes in these variables between 1995-96 and 1999-2000 are shown. Data are presented for each city and state.

The full copy of the report may be found on the Council of the Great City Schools web site at <http://www.cgcs.org/>.

MATCHING THE JUDGMENTAL TASK WITH STANDARD SETTING PANELIST EXPERTISE: THE ITEM-DESCRIPTOR MATCHING PROCEDURE

Marianne Perie, Steve Ferrara, and Gene Johnson, American Institutes for Research

A variety of standard-setting approaches has been introduced in recent years, including behavioral anchoring, body of work, bookmarking, and policy capturing. The American Institutes for Research (AIR) has refined an item-mapping-based procedure called *Item-Descriptor (ID) Matching*. The ID Matching procedure emerged in 1993 as a means for re-setting cut scores and re-defining achievement-level descriptions for the 1992 administration of the Maryland School Performance Assessment Program (MSPAP). Steve Ferrara, state assessment director in Maryland at the time, Ross Green of CTB, and Nadir Atash, then of Westat, developed the procedure over two years. Marianne Perie, Gene Johnson, and Steve Ferrara at AIR have refined the ID

Matching procedure and used it to set provisional standards in 1999 for two end-of-course examinations in the School District of Philadelphia's Citywide Proficiency Examinations program and to establish performance levels on a principal's knowledge certification test and on assessments of student achievement in Portuguese and mathematics at grades 4 and 8 for the state of Bahia, Brazil.

Earlier behavioral anchoring work on the 1991 MSPAP influenced two approaches to standard setting: CTB McGraw-Hill's popular Bookmark procedure (see Mitzel et al., 2001; CTB Macmillan/McGraw-Hill, 1991, p. 11–1) and AIR's ID Matching procedure. The Bookmark and the ID Matching procedures share some characteristics. Both require item mappings based on IRT item scalings. Also, proponents of both procedures assert that these procedures reduce the cognitive load for panelists (see Mitzel et al., 2001, p. 250). ID Matching takes this a step further by requiring of panelists a cognitive task that may be more suited to their expertise as educators and curriculum area experts: It requires panelists to match item-response requirements to performance-level descriptions; it requires no probability judgments like those in the Bookmark and Modified Angoff.

In the ID Matching procedure, items are organized from the least difficult to the most difficult based on their IRT scale location. Panelists examine each item (and rubric for constructed-response items), determine the content knowledge, skills, and cognitive processes each item requires, and then match those requirements to one level in a set of performance level descriptions. Panelists start with the easiest item and work their way through successively more difficult items, matching item-response requirements to a performance level description. As a result, panelists produce sequences of items that match one or another performance-level descriptions. A sequence of items in which some items match more closely the next highest performance-level description but others match more closely the lower performance-level description is called the "threshold region." The threshold region is defined by this alternating pattern of matches, and the cut score is placed at the midpoint of the threshold region. In subsequent rounds, panelists adjust cut scores by determining blocks of items that most closely match descriptions of the content knowledge, skills, and cognitive processes of each performance level.

The ID Matching procedure is distinctive in two ways. One, it captures information about panelist thinking. By illuminating sequences of items that clearly match performance levels and sequences of items that are in the threshold between two performance levels, the procedure gives panelists a focus for discussions that help them achieve convergence of judgments. Two, the cognitive process seems well suited to panelist expertise: They match knowledge and skills necessary to answer an item correctly to the knowledge, skills, and processes contained in performance-level descriptions. Unlike some Angoff approaches, ID Matching does not require panelists to envision hypothetical examinees to judge the probability that he they will answer an item correctly. And ID Matching does not require panelists to identify a pair of adjacent items where the probability of a correct response

falls below 0.67 (or another Response Probability value). Research ranging from the 1950s into the 1990s in judgment and decision-making indicates that people are susceptible to judgmental biases and are prone to making errors when judging the probability of an occurrence (Plous, 1993, p. 144), such as the percentage of students likely to respond correctly to an item. Based on this line of reasoning, the item-descriptor matching task seems well suited to the daily professional experience of the content area experts who serve on standard-setting panels.

Despite its similarities to the Bookmark procedure, ID Matching is distinct in at least two ways. One, by requiring the panelists to match item requirements to a performance-level description before determining a cut score, ID Matching provides panelists with a visual that allows them to identify anomalous items easily. Two, the item-descriptor matching process does not require panelists to match items to performance-level descriptions sequentially. Inconsistent sequences of matches between items and performance level descriptions identifies the threshold region, and the cut score is identified as the midpoint of the threshold region—allowing for the estimation error and instability in item parameters. That is, as item parameters shift from a field test situation (when many performance standards are set) to operational testing, this instability is less likely to affect the location of cut scores in ID Matching than in the Bookmark procedure.

More detailed information on ID Matching will be presented at a colloquium sponsored by the Board on Testing and Assessment (BOTA) of the National Research Council on September 18th in Washington, DC.

References

- CTB Macmillan/McGraw-Hill. (1992, June). *Final technical report: Maryland School Performance Assessment Program 1991*. Baltimore, MD: Maryland State Department of Education.
- Mitzel, H. C., Lewis D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Plous, S. (1993). *The psychology of judgment and decision making*. New York: McGraw-Hill.

Brad Hanson is the new NCME webmaster.

**His e-mail address is
Brad_Hanson@ctb.com.**

SOME "OLD FASHIONED" NOTIONS ABOUT ELECTRONIC TESTING

Jon S. Twing, NCS Pearson

Readers of this Newsletter know that many web-based testing solutions available on the market today are completely capable of delivering test questions to students in a variety of formats, collecting student responses to these questions, as well as providing for the scoring, storage and reporting of student results. Most successful implementations of electronic and web-delivered assessments are indeed likely to incorporate a variety of options addressing these implementation issues. However, these "solutions" cannot be independent of the goals and desires of the customers contracting for the assessments and some are troubling in an "old fashioned" way. For example, requiring a dedicated connection between the server and the end user may eliminate the fear of outsiders "hacking" into the system but is likely not to be very practical for large-scale assessments. As such, "state of the art" security may be beyond the reach of many large-scales high-stakes test customers. Our customers and constituents expect the testing system to solve practical measurement problems without limiting the needs of the assessment users or the customer due to constraints imposed by the electronic system.

Perceived Needs. In a nonscientific manner, I have "culled out" some of the perceived needs of people requesting electronic testing. First, it would seem that one of the most important and often asked for solutions is with respect to the issue of turn-around time for student scores. Most assessment users want student scores returned faster and see the use of computerized assessments an excellent opportunity to do exactly that. Second, many users of large-scale assessments see the "on-demand" nature of computerized assessments helping them provide the appropriate assessment for the appropriate occasion. For example, during the fall of the school year, such computerized assessments could help with placement decisions for newly arriving students. In summer school, such assessments could allow for tailored testing following the remediation and/or mastery schedule of individualized instruction. Third, one of the biggest criticisms of current high-stakes achievement testing in the United States is that it narrows the curriculum by focusing on only trivial or easy to measure pieces of the curriculum. Users of assessments see electronic testing as a means to cover more of the curriculum and, with the power of the computer medium, cover it with more "authentic" aspects of student learning. In fact, many users of high stakes assessments see computerized assessments as a stepping stone toward a system which fully integrates assessment with instruction.

Extenuating Circumstances. The perceived needs of the users of online assessments as outlined (faster results, on-demand use, and instructional integration) sound like wonderful goals but are complicated by a host of "old fashioned" problems. For example, at the same time users of assessments are calling for faster and faster return of scores, they are adding more and more open-ended tasks, writing tasks (essays), and performance tasks to the assessments. Clearly the burden of the human scoring required for these tasks is in opposition to

the faster delivery of student scores. The many different "essay scoring engines" currently available will, perhaps, allow for the best of both worlds (e.g., more performance-based test items and faster turnaround). If customer reaction is any indication, however, we have a long way to go before the users of large-scale assessments are likely to embrace these "automatic scoring machines" due to nothing more than their lack of face validity. Another potentially intervening complication regarding the "on demand" use of computerized assessments is the assumption that students will be on different *individualized* assessment tracks. What this means is that if on-demand assessment is available, students are likely to progress to different levels of standards-referenced mastery at different times. Hence, a classroom teacher may be faced with thirty different individualized lesson plans for each student in his/her classroom. As such, policy and procedures to deal with what could be called a "virtual classroom" should be anticipated and planned for. Integrating instruction with computerized testing will also require a different way to look at lesson planning. For example, one of the current criticisms of high-stakes achievement testing is that most remediation is simple "drill and kill" where students practice with similar test items to improve their test taking skills and/or to temporarily improve their content knowledge. Because it is not grounded, once the practice ends the information is essentially useless because it does not constitute real learning, understanding or integration into current schema. Computerized assessment with links to instruction will perhaps provide an opportunity for unscrupulous vendors to make a bad situation worse by pointing to and reinforcing poor instructional habits. Many of us have the "old fashioned" worry of doing more harm than good.

Food for Thought. The practical implementation of electronic testing, while arguably representing the future, is nonetheless generating some "old fashioned" concerns. Many users of large-scale assessments are moving ahead regardless of some of these potential problems. For example, very few large-scale users have the infrastructure to support an electronic testing solution for the entire student population. As such, some schools may have limited participation in electronic testing. This, in and of itself, might not be a problem but there needs to be a plan that for incorporating all schools into the electronic solution and this is likely to be expensive. If the electronic capture of student-generated responses (i.e., open-ended and/or performance tasks) is required, how will these questions be scored electronically? If electronic scoring is not used, most of the power of the electronic medium will be lost. How will we as researchers address the anticipated challenges coming from a perceived lack of "face validity"? What plan will we offer users to demonstrate simple score reliability and defend the validity of the resulting score interpretations? If electronic testing will allow for easier and more frequently implemented "on-demand" testing, what will this do to the size of the item pools required? Will we need more items due to more testing? Will current item development procedures suffice? Will the general public accept automated item and test form construction? These are some of the thoughts that keep an "old fashioned" test builder awake at night.

Special Request

I would like to make a request of NCME members. I've been working on "professional genealogy" of the field of measurement. It seems to me that many of us can trace our professional "lineage" to a relatively small number of early measurement pioneers. If you have a minute, could you email me the name of your mentor and the university from which you received your doctorate? In order for this to work, I need to hear from recent grads and old-timers alike. For most folks, your mentor was probably the person who chaired your dissertation; if not, send me the name of the person who you really feel was your academic mentor in your program. My email address is: jefsmith@rci.rutgers.edu (note only one "f" in the email address). If I can get enough of these together, I'll try to put together an EM:IP piece on it. Thanks!

Jeff Smith