

A MESSAGE FROM YOUR PRESIDENT – JOHN FREMER

THOUGHTS ON THREE TOPICS

In this issue, I will use the wonderful privilege that I have to offer observations about developments in our field, to touch on three areas:

- Telling the story of testing
- What do the testing surveys/studies say?
- My last (?) comment on construct validity



TELLING THE STORY OF TESTING

I want to appeal to my colleagues to use their fine minds and energy to help raise the level of the public debates about testing. It worries me quite a bit that much uninformed and wrong-headed commentary about testing can be found in the seemingly endless string of stories that have been appearing about individual testing programs and about whole areas of testing. One of the particularly troublesome trends that I have noted is the frequency with which standardized tests are routinely labeled “unfair” or “biased” as if this were a well-established “fact.” We have within NCME a large fraction of the measurement-trained folk in our society, so if it is not our job to bring good sense to the testing debate, who else do we think will rise to the occasion?

You could ask, “What can I do as one person to deal with the flood of off-base media coverage?” While I will admit that the task is a formidable one, this is not a task that we have to do entirely on our own. I believe that any member who is willing to speak out will find “kindred spirits,” not just within NCME, but also within other like-minded organizations. AERA and APA, our valued partners in developing the *Standards for Educational and Psychological Testing*, are obvious examples and I believe that we can also find many willing helpers within the other associations with whom we sponsor the Joint Committee on Testing Practices (JCTP), e.g., American Counseling Association, American Speech-Language Hearing Association, National Association of School Psychologists, and National Association of Test Directors. We merely need to reach out and show that we would like a hand in this effort.

So I urge our members to speak up when they encounter biased or misleading reporting about testing. Write to your newspapers and speak up at public meetings. If you are in a position to do so, get on the program of state or regional meetings for a variety of organizations and explain the strengths and weaknesses of educational testing. Our work is so much in the public eye now that you will probably never find it easier to get a general testing piece published or a session accepted for presentation. Your help is really needed to raise the level of the debate and to undercut comments that are without merit but are seductively appealing to people who have no training in measurement.

WHAT DO THE TESTING SURVEYS/STUDIES SAY?

There are shortages of trained measurement staff to fill positions in the testing field and of students to fill our graduate programs, but we certainly have no shortage of surveys and related studies that have looked at attitudes toward various aspects of testing. The most recent one that has come to my desk is actually kind of fun. The Boston College Center for the Study of Testing, Evaluation, and Public Policy asked a sample of Massachusetts’ students to draw a picture of them taking a statewide test. Then the pictures were classified as positive, negative, or no reaction. To me the most interesting part of the study is the design. It helps remind me that there are many ways of looking at the world of testing.

Other studies have looked at attitudes toward high stakes tests or more generally at school testing. You could argue that there are enough perspectives reflected in the results to support almost any position, but I see some consistent patterns. Parents tend to be concerned about their own children and how test results might affect them. They tend to trust the results of tests and have a fairly healthy appetite for information about their own children as well as about the school system. Over-reliance on testing and having too much faith in results is more characteristic of the reactions of teachers, researchers, and commentators than it is of parents. *(continued)*

CALL FOR NOMINATIONS FOR NCME AWARD FOR DISSEMINATION OF EDUCATIONAL MEASUREMENT CONCEPTS TO THE PUBLIC

NCME members are encouraged to nominate an individual or a group for their outstanding accomplishments in disseminating educational measurement concepts to the public. NCME will honor these accomplishments at the annual meeting in 2001.

Deadline for nominations is extended to February 28, 2001.

Please submit nominations to: Steve Ferrara, Chair, American Institutes for Research, 1000 Thomas Jefferson St., NW, Washington, DC 20007, V: 202-944-5431, F: 202-944-5454, (sferrara@air.org).

MY LAST (?) COMMENT ON CONSTRUCT VALIDITY

There is a “rejoinder” from Bert Green on the worry I expressed in the last newsletter about our emphasis on construct validity in our new *Standards* and in much current writing about validating uses of tests. I got almost exactly the same reaction from one of my long-term colleagues at ETS, Michael Zieky. I want to point out that Bert was a very effective member of the *Standards* drafting committee and that Michael Zieky led the ETS effort to revise our own *Standards for Quality and Fairness*. I have no doubt that these senior and very well informed measurement experts have a clear conception of what the construct validation requirement should mean. What I am arguing is the degree of understanding of the other 99.9% of the people who have some role in an educational testing process. To me it is like the sophisticated computer person explaining to me that this latest upgrade or change in software will be a “piece of cake.” At one level this is a readily investigated topic. I invite my NCME colleagues to practice explaining the concept of construct validation to teachers, school board members, test adoption committees, etc. Then we can perhaps use a variant of the Boston College approach and take photographs of the people hearing the explanation. My guess is that we would find ourselves seeking a more common-language term such as the idea of a “coherent account” that you see in Bert Green’s fine comment and also wanting to have some compelling examples of such accounts to bolster our explanation.

If anyone wants to continue this discussion with me, I can be reached at jfremer@ets.org.

THE COMING TESTING BACKLASH?

H. Gary Cook, Wisconsin DPI

A brief reading of many general K-12 educational publications over this past year reveals the emergence of a word we have seldom seen or used in the measurement field, “backlash.” A search of the word “backlash” on the *Education Week on the Web* archive reveals 31 references for just the year 2000. Of the 31 articles, editorials, or letters to the editor, 27 dealt with backlash to tests, particularly state-mandated tests.

The most recent article in this archive (November 11, 2000) chronicles Massachusetts’ largest teachers union’s \$600,000 media blitz against the state assessment. Earlier this year, hundreds of Massachusetts students were reported to boycott state assessments (*Education Week on the Web, April 19, 2000*). To alleged teacher and student angst add individuals who have been barnstorming the nation bashing the use of any standardized assessment. One chief, polemic protagonist is Alfie Kohn. In a recent public speech, sponsored by the Wisconsin Association for Supervision and Curriculum Development (WASCD), Mr. Kohn implied that the use of standardized tests for student or school evaluation was tantamount to genocide. He stated in a September 27th, 2000, *Education Week* article that “[s]tandardized testing has swelled and mutated, like a creature in one of those old horror movies, to the point that it now threatens to swallow our schools whole.”

Not all of the aforementioned *Education Week* articles reported antagonism to standards or tests. An October 11th article titled “Polls Dispute a ‘Backlash’ To Standards” reported poll results suggesting that the large public opposition to tests and standards was “enormously exaggerated.” The results mentioned in the October 11th article are consistent with an earlier study reported by Richard Phelps in the Fall 1998 *Educational Measurement: Issues and Practices*, 17:3. In this study, Phelps summarized the last 25 years of opinion polls about standardized tests. He found that, “[i]n the case of standardized student testing, the public generally favors more of it, with higher stakes” (p. 16).

There are a variety of strongly held conflicting opinions about standardized assessments, particularly high-stakes assessments. One might also argue that there are strongly held conflicting opinions about the scale of the testing backlash, particularly around statewide tests. In Wisconsin, we do not have a wholesale revolt (backlash) in using statewide assessments. But there is certainly a high degree of tension among educators and parents regarding their use. In my experience, much (if not all) of this tension has resulted from either misunderstanding the intended purposes and uses of assessments or miscommunication about the assessments.

As there are large-scale assessments, there also seems to be large-scale misunderstandings about their purpose and use. On occasion we query colleagues in school districts around Wisconsin about the purpose and use of our state assessments. Frequently, educators respond, “Because the state says we have to.” We follow up with, “Why does the state do that?”

NEWSLETTER ADVISORY BOARD

BETTY BERGSTROM, Computer Adaptive Technologies, Inc.
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, National Academy of Sciences
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
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 email: 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.

A variety of answers are offered. Many reflect major gaps in assessment literacy. Thus, we have engaged in a statewide campaign to improve the assessment literacy of our parents, pedagogues, politicians, and public. Wisconsin's Office of Educational Accountability has three full-time staff members dedicated to assessment literacy. Improving the assessment literacy of our publics, we believe, will facilitate a more reasonable notion of test use and better test policy.

Similarly, there is large-scale miscommunication about assessments. In our view much of this miscommunication (at least in Wisconsin) is at the state level. We have failed to communicate to our publics in a way that promotes understanding. We assessment folk have been trained in a second language called psychometric-jargon. Our expectation is that our publics, especially fellow educators, know what we mean. They often don't. Because of limited assessment literacy, the concepts and terms we use are foreign to "non-measurement" folks. We need to fashion our messages in a language understood by our audience. Also, the materials we have developed for public use are often printed in size-12 pica font, on 8 1/2 x 11-inch white paper. They could substitute for great insomnia cures. In this information age, our messages are competing with high quality, graphically interesting messages created by honed marketing professionals. We need to improve our media as well. In Wisconsin, we have contracted with a public relations and marketing firm to assist us in communicating to our publics. We are concerned not only with the message but with the media as well.

We have experienced lower levels of angst regarding state assessments as a result of repackaging and re-fashioning our messages. In many forums, our assessment literacy staff has reported many instances of the "aha" factor. The "aha" factor occurs when people connect to test purposes and test uses - comments like, "Now I understand why we do this." We cannot say that these strategies will eliminate the testing backlash, but we believe that test backlash that results from misunderstanding or miscommunication can be greatly curtailed.

H. Gary Cook, PhD, is the Director of the Office of Educational Accountability for the Wisconsin Department of Public Instruction.

STRESSFUL COEXISTENCE OF LARGE-SCALE AND LOCAL TESTING PROGRAMS

Duncan MacQuarrie, Tacoma Public Schools

Has large-scale student testing at the state and national levels reached its zenith? The problems a number of states encountered while recruiting schools to participate in the state component of the 2000 National Assessment of Educational Progress suggests that it might have.

Despite a highly organized recruiting campaign, titled "All States 2000," last year's state component of the National Assessment eventually included only 40 of the 48 states that originally signed up. The eight that dropped out did so because sufficient numbers of local education authorities could not be persuaded to participate. Growing resistance to participation in the National Assessment is one of the most visible indicators that schools and districts have reached a limit in the amount of

external testing they can tolerate, particularly testing that is perceived to yield little local value.

The National Assessment program, a leader in large-scale assessment since the early seventies, has been an important model for the standards-based assessments that so many states have implemented or are developing. Congress and the U.S. Department of Education have promoted the development of these assessments through a variety of mandates. Many of the standards-based assessments include mixed response formats, i.e., selected response, short answer, extended response, and the direct assessment of student writing. These innovations have increased the complexity of all aspects of the assessments including development, implementation logistics, scoring, and the reporting and use of the results. This complexity is in danger of overwhelming the very systems these assessments are intended to help.

It is not just the complexity of the tests, but the sheer numbers with which local districts and schools are confronted, that are causing problems. The most common levels for state assessments are grades 4 and 8, where the Council of Chief State School Officers' 1999 annual assessment survey showed 37 and 48 states, respectively, test their students. Although occurring in only a sample of schools, these are the grades the National Assessment also tests and the governing board's current schedule calls for some form of the assessments to be administered every year. And both major parties' presidential candidates in 2000 advocated, in one fashion or another, national policies that would increase the amount of large-scale student testing.

Another initiative gaining in popularity is the "value-added" accountability model in which students in all grades are annually tested. Such an approach dramatically increases the amount of testing time required in schools and, if mandated from the state level, would conflict to a much greater degree with the time local districts might wish to devote to their own testing programs.

Increased national and state testing also competes with publishers' needs to periodically conduct norming studies and with school districts' desires to introduce their own testing programs in support of locally developed accountability efforts, such as school improvement programs and student promotion/retention decisions. In my own district, we conduct mandated state assessments at grades 2, 3, 4, 6, 7, 9, and 10. We recently scaled back our local assessment program so it includes only short criterion-referenced tests in grades 1, 5, and 8. Although the schools from our district selected in recent years for the National Assessment have agreed to participate, we have found it necessary this past year to decline at least three invitations to participate in norming studies.

The number and complexity of our assessments may have reached the saturation point. Invitations to participate in voluntary assessments, even for the most noble of reasons, are likely to be turned down.

Congress, state legislatures, and state boards of education need to reconsider the role of state and national testing as a policy tool for promoting education reform. Testing programs are most useful when they are an integral part of local education

programs. While there is a place for mandated large-scale assessments, their current numbers and/or complexity are in danger of driving out important local assessments.

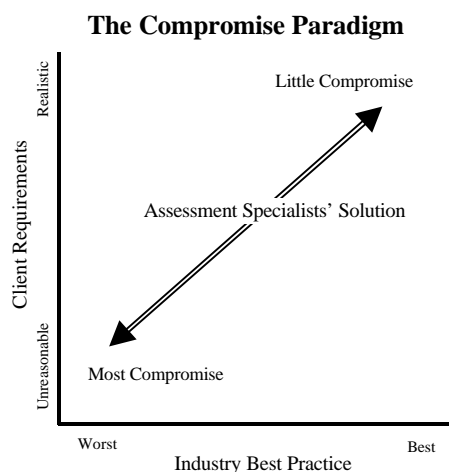
Dr. Duncan MacQuarrie is the former Director of Assessment for Washington State and currently directs the student assessment program for the Tacoma Public Schools in that state.

LARGE-SCALE ASSESSMENT: THE ART OF THE COMPROMISE Jon S. Twing, NCS Pearson

There is little doubt that the use of large-scale assessments, particularly high-stakes assessments, is on the rise. State departments of education have solicited services for such high-stakes testing across the nation (California, Texas, Mississippi, Minnesota, Michigan, and Florida, to name a few). With this increased emphasis on testing, the cries from the trenches regarding the burden such testing places on students, schools, and society in general are also on the rise. District test coordinators comment on a regular basis regarding the challenge of increased testing with limited resources. Add to this burden the increasing political demands to “hold schools accountable” (not to mention the faster turnaround of assessment results) and the life of assessment specialists in local schools is hectic and stressful to say the least.

Unfortunately, the testing burden is not limited to local test administrators. State departments of education are often understaffed and inexperienced when dealing with the shifting burdens of assessment in these politically shifting times. This is usually not the fault of the state agencies, however. Often state departments of education must implement, with equal vigor, well-crafted, meaningful, and insightful legislation along with ill-conceived, dubious, or simply poorly worded mandates.

To reduce this burden, local district personnel, state agency staff, and often politicians and legislators call upon assessment professionals. We are asked to advise, design, and implement solutions that meet the needs of all parties involved. This is no small challenge. In dealing with such challenges, one of the keys to success is understanding the art of compromise. The figure below presents the “compromise paradigm.”



As assessment specialists, it is our charge to deliver measurement advice that maximizes the alignment of the assessment solution to industry standards of best practice. In fact, we are required to do so by our own professional standards. At the same time, however, the circumstances often dictate that the “best” solution is not a viable one, due to unreasonable client requirements. In these situations, it is the responsibility of the measurement specialist to determine a solution that meets the client’s needs (or demands) *and* that is still acceptable when scrutinized under the standards of our industry. When client requirements are realistic and reasonable, little compromise is required in finding a solution to the customer’s needs that meets standards or best practice. These are the solutions that psychometricians publish in peer-reviewed journals. However, when the client’s needs are unrealistic or unreasonable (this might be for a host of reasons, including poor legislative mandates), the solution will require more of a compromise. This typically causes the assessment specialist great anxiety. We must decide and judge for ourselves how far down the vector of compromise we are willing to move, knowing that we will have to defend the proposed solution. These compromises are seldom published in professional journals, but they are, nonetheless, the solutions most treasured by our clients.

A generic example of the compromise paradigm at work can be seen in the area of statistical test-form equating, particularly as it relates to sample size. For example, a legislative mandate to a state agency might be to provide multiple test forms, parallel in both content and difficulty. But the mandate will probably not refer to test-form equating or to the data collection design required to perform such an equating. The assessment specialist may recommend to the state a “typical” common-item counter-balanced design that samples 4,000 students from across 80 districts and 125 schools. Each school would presumably be selected to represent the state in terms of key demographic and background characteristics. The state agency, on the other hand, might suggest a sample of 500 students from one of their “favorite volunteers” to lessen the impact of research on a busy school calendar.

The initial design was probably based on what the assessment specialist saw as acceptable “best practice.” The modified design suggested by the state was probably unrealistic, but motivated by the sincere desire to reduce the testing burden. Clearly, the assessment specialist started from a point of little compromise and will in all likelihood have to move down the “compromise continuum” until an equitable sampling plan/data collection design is agreed upon. Where on this continuum is the point below which the plan/design is no longer defensible? No one knows for sure. That is why it is essential that measurement specialists rely on their expertise, the professional standards of our industry, and our partnership with the client to ensure that the final solution is defensible. This is no small task given that the professional standards are not prescriptive and the sophistication of the state agencies is diverse.

The Association of Test Publishers conference *Computer-Based Testing: Emerging Technologies and Opportunities for Diverse Applications* will be held February 26-28, 2001 in Tucson, AZ. For details, visit <http://www.testpublishers.org/Con2001/index.html>.

ENHANCING NAEP PARTICIPATION

Mary Lyn Bourque

National Assessment Governing Board

Until 1990, the acronym NAEP was synonymous with (what is now known as) the long-term trend NAEP, since that was the primary assessment administered in the National Assessment program. From the early 1970s the federal government collected achievement data in science, mathematics, reading, and writing on a periodic basis. There was an occasional data collection in other subjects such as a computer competency survey in 1986, music in 1972 and 1979, and the social studies in the 1970s and 1980s. However, the first four subjects form the basis of the long-term trend NAEP as we know it today.

Since the implementation of IRT scaling to track trends, the examinee sample size for the long-term trend has been reduced from approximately 75,000 (in 1977 science, for example) to 16,000 in the 1999 administration. This sample size reduction has served the program well in terms of eliciting good participation rates from schools selected for participation in the samples. The sample is selected to be nationally representative, and the number of schools is fairly small (between 750 and 900 total). Consequently, schools know that, with few exceptions, if they are selected in one year, they are apt not to be selected again for another few years. In the past, this encouraged good will and thus fairly high school participation rates. The data from long-term trend, however, is beginning to show that school participation is declining - over nearly 20 years, the weighted school participation rates have dropped almost 20 percentage points in some cases.

NAEP in the 1990s

In 1988 the National Assessment was legislatively authorized to begin a trial assessment collecting NAEP data at the state and jurisdictional levels (e.g., the District of Columbia, Guam, Virgin Islands, and other jurisdictional areas). At the same time, the assessment frameworks for various subjects, including reading, writing, science, and mathematics were revised to make them more consistent with what was happening in America's classrooms, with new directions in curriculum, and with the findings of educational research. These new assessment frameworks were used to develop both a new national assessment and the newly authorized state assessments. At that time the decision was made to leave the long-term trend assessment unrevised, in order to monitor trends in at least the four major subject areas.

By 1990, the acronym NAEP was no longer synonymous with long-term trend: now there were three NAEP assessments, long-term trend, national NAEP (also called short-term trend), and state NAEP (and the latter two had associated field testing as well). Overnight, many more schools were being asked to participate in one or more NAEP assessments.

By 2000, the number of schools sampled for the national NAEP was about 2,500 (including over-sampling of private schools for subgroup reporting); for state NAEP the sample size is about 100 – 125/grade/state, or 5,000 to 6,000 schools across all states (assuming two subjects in two grade levels); for long-term trend the number remains between 750 to 900 across all three grades; and finally the field test samples include about 500 schools. The NAEP assessment schedule

calls for state assessments every other (even) year in reading/writing or mathematics/science, with the odd years being national NAEP in alternative subjects, field-testing, and/or long-term trend. So, on average, anywhere from 3,000 to 6,000 schools or more each year are selected in one of the three operational and field-testing samples.

ALLSTATES 2000

In 1998 the National Center for Education Statistics (NCES) and the National Assessment Governing Board (NAGB) initiated a drive to have all 50 states participate in the 2000 NAEP cycle. ALLSTATES 2000 was a very successful effort since all 50 states seriously considered signing up and indeed 48 states did agree to participate. However, in the end fewer than 40 states will be reported, and some of them will be flagged in reports for response rate problems.

The table below shows some trends in state participation (in grade 8 mathematics) over the decade of the 1990s. State A has a steadily declining rate, and will have its results flagged in the final NAEP reports for not meeting the participation rate criterion. State B also has a steadily declining rate, and in 2000 dropped out of the assessment for the same reason as State A. State C, a state that offered incentives, shows a similar declining pattern.

Weighted State NAEP School Participation Rates in Three Selected States, Grade 8 Mathematics

	1990	1992	1996	2000
State A	86%	83%	80%	72%
State B	98	78	65	--
State C	97	94	86	82

Data taken from Caldwell, N. (2000). Emerging Context for NAEP (unpublished manuscript).

Even states like Florida and Colorado, that have legislation on the books requiring NAEP participation, withdrew from NAEP 2000 at some point because they did not meet the response rate criterion set by NAGB policy (70%). In the case of Florida, there was a direct conflict between the Florida Assessment program's testing window and the NAEP testing window. Principals in that state simply refused to be distracted from their own state program by participating in NAEP. Other states, like Michigan, provided local incentives (\$1,000 per school) to participate. It helped. ALLSTATES 2000 was a concerted effort on the part of federal, state, and local policymakers to encourage NAEP participation. In the end it failed to secure full NAEP participation in about a dozen states. The question is, why?

Reasons for Non-Participation

There are a myriad of reasons why participation in NAEP has become such an issue. Clearly, policymakers at the state level are still committed to a large-scale assessment program like NAEP. For governors and chief state school officers, NAEP represents the "gold standard." For some, it is a validity check on their own testing programs and results. For others, it is a monitor of program improvements and student progress. However, there is no state that is not also involved in, and committed to, their state accountability system. High-stakes state assessment programs have changed the stakes at the district level as well. Consequently, districts are much less

inclined to want to participate in any outside program that (1) has no direct benefits to the schools and staffs; (2) is viewed as intrusive on instructional time; and (3) is labor-intensive for their staffs.

State NAEP Issues

Unlike national NAEP, which is wholly administered by a contractor, the state NAEP places a fairly extensive burden on *local districts* that comes with the *states* saying yes to participate. Federal legislation requires that state NAEP be wholly administered by state and local staff. So, for example, state staff is required to recruit and obtain cooperation from districts and schools selected into the sample and arrange assessment dates with respective school principals. Local staff (assessment examiners) is required to receive examiner training for each assessment, and then administer the assessment to the school sample. They also receive, account for, and return-ship all assessment materials, as well as assume all clerical duties associated with ensuring examinee samples are notified, properly matched to test materials, and tested. The price tag for this (in-kind costs) across all participating states in any one assessment is estimated in excess of \$8 million or 160 full-time equivalents.

Further, the stakes are very high for districts. In high-accountability states, districts not meeting state standards are subject to some kind of consequences, whether it be "on probation," "taken over," or some lesser consequence. Several state accountability systems are tied to graduation requirements, and teacher merit awards. Any activity (testing or otherwise) that infringes on instructional time is rejected. The consequences of a district's failing to meet state accountability standards can be great, there is usually a price to be paid, and districts are not willing to pay this price in the name of altruism.

Finally, the districts receive nothing in return for their willingness to participate in NAEP. By design there are no student scores, there are no classroom results, there are no grade level results, and there are no school or district results. The very most the district will receive when the data are released is the aggregate performance data for students in the state in which the district resides, and a small sample of released test questions that may be used by teachers in their instructional programs, if desirable. These are weak incentives to encourage broad participation.

NAGB Proposals

Since the lessons learned from ALLSTATES 2000, those responsible for NAEP have been discussing and meeting with NAEP's publics to craft some possible solutions.

NAGB received a report in September from one of its committees regarding changes in operations, policy, and legislation that might at least alleviate, if not eliminate, the NAEP participation problem. The eight recommendations included (1) improving/streamlining communications; (2) using intact grade sampling; (3) keying NAEP items to state standards; (4) coordinating NAEP with other state/district testing; (5) providing performance feedback to schools; (6) developing a teacher tool kit (similar to the TIMSS Tool Kit); (7) moving to contractor-administered state NAEP; and (8) providing resources for state NAEP.

In a series of recent focus groups with principals and local superintendents, each of these ideas was evaluated for its potential effectiveness in resolving the participation problem. While none was seen as a silver bullet, the three receiving the highest support included keying NAEP items to the states' standards, providing performance feedback to schools, and using intact grade sampling. The first two it is believed would be responsive to the issue of minimal benefits in NAEP, and would allow NAEP to provide to those who do participate information that would be helpful and supportive of their own state and local programs. The third would reduce disruption in the schools by testing all students in a grade rather than a random sample of students across various classrooms. It is believed that this approach to sampling would help to reduce some of the burden at the school level.

NAGB is actively pursuing each of the eight recommendations. Some would require changes in the legislation, for example, a contractor-administered state NAEP or providing resources for state NAEP such as a full-time equivalent staff person. This is what some would call 'perfect timing', since NAEP will be re-authorized in the next session of the Congress, and changes of this magnitude could be made known to lawmakers in time to potentially become part of the new legislation. Other recommendations such as providing performance feedback to schools (which may also require legislative changes as well) and intact grade sampling are more problematic since they move into a territory never before explored by NAEP, and which, by design, the current NAEP is unsuited to fulfill. The technical issues will need to be resolved if these types of approaches are to be employed in the future.

At its recent meeting concluded just as this *Newsletter* was going to press, NAGB was still pursuing each of the eight recommendations. While it is not yet clear which ones, if any, will be adopted, it is quite clear that if NAEP is to ameliorate or eliminate the participation issue, more than incremental changes will be needed. Many of these changes could be costly (and large budgetary increases are unlikely after four years of level-funding), and at least some will need solid research behind them before moving forward on a redesign plan. This could take several years to plan and implement.

Summer Internship Program

ACT, Inc. annually conducts an 8-week summer internship program for outstanding doctoral students interested in careers with a testing focus. In 2001, the program will run from June 4 through July 27 at ACT's headquarters in Iowa City, Iowa.

Interns are provided a \$3,500 stipend plus reimbursement for round-trip transportation costs. A supplemental living allowance for accompanying spouse and/or dependents is also available. Internships are offered in the following areas:

- **POLICY RESEARCH / PROGRAM EVALUATION / INSTITUTIONAL SERVICES**
- **I/O PSYCHOLOGY**
- **PSYCHOMETRICS AND STATISTICS**
- **VOCATIONAL PSYCHOLOGY**

Application deadline is **February 15, 2001**.

Information and application materials are available at:
www.act.org/humanresources/jobs/intern.html.

REMEMBERING DICK JAEGER



Richard M. Jaeger, NationsBank Professor (Emeritus) of Educational Research Methodology in the School of Education at the University of North Carolina, Greensboro, passed away on October 21, 2000. Professor Jaeger, in addition to serving as NCME President (1986-87), was the recipient of the NCME Career Award (1998) and the AERA E.F. Lindquist Award (1992). The following

tribute is a collection of memories from just a few of Dick's many colleagues and friends.

The field of educational measurement has lost one of its giants. Dick Jaeger's death is a tremendous loss to the profession and a personal loss to his many colleagues and former students who had the pleasure to work with him. Dick was a wonderful colleague who inspired better work and actions from colleagues and students by example and by his thoughtful and detailed reviews of manuscripts. He was committed to educational equity and contributed much to that cause. He leaves a huge void, but many great memories for a large number of the NCME family. *Bob Linn*

Dick Jaeger was an important part of my professional life. I met him for the first time in 1973 when we were working with Mel Novick on applications of Bayesian statistics. Even early in his career Dick made a distinct impression on me. He was insightful, experienced, very clever, and a superb writer. Later, Dick edited several of my papers and chapters in his role as JEM editor and book editor. I was amazed by the care he took over every word I wrote and could imagine the amount of time he must spend on his own writing and research. I saw these same skills demonstrated many more times over the years and with many persons besides myself; he cared about ideas, scholarship, and communication, and he was willing to spend enormous amounts of time to get things right. All of his professional contributions were made under a severe health handicap that few of us knew about and made his accomplishments all the more remarkable. Over the last ten years, Dick, Barbara Plake, Craig Mills, and I worked closely on a number of studies to develop new standard-setting methods. I marveled at his creativity, his problem-solving ability, his patience, his willingness to work long hours, his kindness, and his interest in ideas and the success of others. I will miss his gentle advice and encouragement, and his friendship, and the measurement field will miss one of its most important contributors over the past 30 years.

Ronald K. Hambleton

Other than being the best colleague I ever had, and my best friend, two characteristics of Dick Jaeger that will remain permanently in my memory were the incredible speed with which he wrote, and his remarkable editorial ability. One afternoon in 1990, not quite two years after I arrived at UNC Greensboro, Dick walked into my office and said to me that we (he and I) ought to bid on the recently formed National Board for Professional Teaching Standard's RFP for a Technical Analysis Group. I told him he was crazy; that we would be going up against ETS, ACT, HUMRRO, AIR, RAND, and a

host of other heavy hitters. He said, "Yeah. I'll see you in the morning." The next morning he came into my office with approximately 35 to 40 pages of flawless prose that constituted the backbone of our eventual proposal. We got the grant. Regarding his editorial ability, anyone brave enough to submit a manuscript to his searing and incisive red pen knows how precise (and how brutal!) he was. Any manuscript that had the benefit of his editorial pen was inevitably improved thereby. *Lloyd Bond*

In 1968, I met Dick Jaeger, who was then at the United States Office of Education where he worked with a consortium of states to redesign the federal evaluation reporting system for the ESEA. I was the Florida Department of Education representative to the consortium. It was an exciting time and an exciting project. As part of the project work, Dick came down to work with me in Tallahassee for several days. The last day, we took some rare time off to visit the beach and do some fishing. We got so involved in the fishing that we had to rush back. I drove as fast as I dared (he would have driven faster) and as we approached the airport (still dressed in shorts), we saw his plane taking off. It was the first time he had ever missed a plane! Other than not getting home to Judy, his only disappointment was that it was too late to go back and continue fishing.

James C. Impara

Like many others, I had the wonderful opportunity to work with Dick on a number of projects, most notably the initial Measurement Research Advisory Panel for the National Board of Professional Teaching Standards and later on a standard-setting research grant from the National Science Foundation. In both settings, I was amazed at how productive he was. When everyone on these efforts was working at full capacity, he made us look like we were standing still eating bonbons! His ability to conceptualize clearly important and complicated problems and then generate elegant solutions, albeit sophisticated, permeated all of his work. The measurement field has lost an important colleague and those of us who had the good fortune to work with him will feel this loss greatly. *Barbara S. Plake*

Our paths had crossed at conferences and occasionally we had shared a session, but I had never spent a longer period with Dick Jaeger until I invited him as a keynote speaker at a national conference in the Netherlands three years ago. His work on testing for certification and performance assessment was widely known in this country, and people were eager to listen to him. From every sentence in his address, it was clear how he had fallen in love with these two areas, but also that his special love was his work on teacher certification for NBPTS. After the conference I translated Dick's address for a Dutch journal, and this article has been a valuable source of information to many a colleague here. As it happens, next year I had a conference close to Dick's hometown, and on my way back I spent two days with him. Of course, we talked about new projects and I browsed among the new acquisitions in his library. But this time these usual things were special. From the first minute there was friendship in the air and we enjoyed our interaction immensely. Shortly thereafter Dick retired, and I started missing him at our conferences. My last image of Dick is of him standing in front of his home, one hand down trying to calm his two lively dogs, the other up to wave me a good-bye.

Now that Dick has passed away so suddenly and undeservedly, this image has much deeper meaning to me than I could ever have suspected. *Wim J. van der Linden*

Richard Jaeger provided major leadership to our field. Dick was also a most caring friend, who mentored many (including me). I recall many occasions working beside him at his desk, or beside his hospital bed, in awe as he could compose every word and sentence, and every utterance in a way that provided the model of clarity. This showed in his teaching where students appreciated such crispness, in his writings, in his textbooks, and in his research work. Dick loved sharing his gadgets, his farm, his dogs, and who can forget his enjoyment of pig picking and hush puppies. Along with Judy, Czach and Oz, Dick provided much to me, my family, and to all those who worked with him and were influenced by his work. Dick, you will be thought of fondly, often, and missed much.

John Hattie

In this passage, I will refer to him as *Dr. Jaeger*, as I could never bring myself to call him anything else, even after my days in his classroom had long passed. Though *Dr. Jaeger's* vast contributions to measurement will be celebrated and remembered many times, these words speak to an accomplishment of his that may be overlooked by his peers: *Dr. Jaeger* was a teacher. When I met him, I was unaware of his place in this curious field that was so new to me; however, I was soon in awe; not because of his weighty vita, but because of his ability to teach, to mentor, to nurture, to inspire, to transform. His wisdom and guidance will be missed by many.

Kristen Huff

In thinking about Dick to write these words, I am again overwhelmed by grief at our loss. Dick's joie de vivre equaled his legendary passion for work. Some of the moments we

spent, fleeting but leaving permanent memories: sipping Lagavulin and debating the virtues of single malts; tooling around the back roads in his Mercedes roadster looking for antiques; commenting on the subtleties of a Lafite Rothschild; discussing the woodworking techniques of Norm Abrams. Dick had a true nobility of soul - he was patient but had no patience with the impatient; he was tolerant to a fault but had no tolerance for the intolerant; he was kind except to the unkind. There never will be another. Farewell, STATMAN!

H. Swaminathan

STANDARDS FOR VALIDATION - RESPONSE TO FREMER

Bert F. Green, Johns Hopkins University

In the September, 2000, issue of the *NCME Newsletter*, NCME President John Fremer draws attention to the 1999 *Standards for Educational and Psychological Testing* by writing that the *Standards* have "elevated the concept of construct validation to so high a level that it seems an 'out of reach' goal." But other critics of the *Standards* cry that construct validity has been lost. Plainly, the concept of construct validity has no single meaning.

The Introduction to the *Standards* avers that not every test scale is a construct. Moreover, not every standard applies to every test, and not every type of validity evidence is available for every test. Recognizing this diversity, the *Standards* advocates assembling relevant evidence and forming a coherent account of evidence and theory in support of the particular interpretation of test scores that is being recommended. Although examples would help, the validation task mainly requires thought and organization, a task well within the reach of NCME members, most of whom have written dissertations.



*Happy
Holidays*

