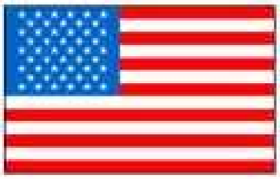
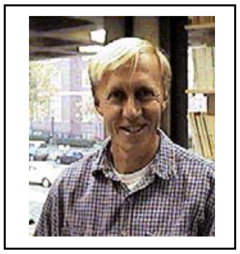
NCME national council on measurement in education



Volume 18, Number 1 NEWSLETTER March 2010

FROM THE PRESIDENT: UPDATE ON NCME ACTIVITIES AND ANNUAL MEETING!



Terry A. Ackerman, University of North Carolina - Greensboro

Dear NCME members,

My year as President as gone by quickly. It has been a very busy year that has been both memorable and rewarding for me. I am so thankful to all of you for allowing me the opportunity to have had this experience.

Update on Current Activities

Wayne Camara (President-elect) and Linda Cook (the newly elected President-elect), and Plumer Lovelace (NCME's Executive Director) and I recently met (March 1st and 2nd) in Washington, D.C. I convened this meeting because I wanted to discuss and layout the NCME agenda and action steps for each committee over the next two years. No president can complete everything he or she would like within their year as president. I believe that to provide a smooth transition and to keep the momentum and NCME's strategic plan moving forward it is essential to work closely with your predecessors. Throughout the year Wayne, Plumer and I have worked very closely together. Our efforts have complemented each other extremely well and I believe we have made a lot of progress. As soon as Linda was elected I felt it was necessary for the four of us to meet and discuss where NCME is currently and where we see the organization moving.

At our meeting we began with our updated strategic plan and organizational goals and talked about action steps for the majority of our committees, especially committees such as the Website, the Membership, and the Outreach. We are in the process of delineating these action steps. Once we've created them, they will be discussed with our Board members and then with each of the committee chairs.

The reason we met in D.C. was because I also wanted to meet with Felice Levine, Executive Director of AERA and discuss the next steps in the renewal of the NCME-AERA contract. As some of you may know, AERA selects the locations and manages the hotel contracts for our annual conference. Plumer and I met separately with Felice. It was our fourth face-to-face meeting since I became president. As Felice had requested, we provided a table of current contract language and noted various services that AERA no longer provides and that have been taken over by our association management company, The Rees Group. We also provided brief suggestions for possible language changes and new services that we would like to request. Felice and I both envision finalizing this contract before our meetings in Denver.

Incidentally, and I know you are wondering, but there were no deer encounters on the trip!

Upcoming Membership Survey

We are in the process of creating a survey that will be sent out to all members. We will be using a lot of the information from the survey to inform different committees about projects they are engaged in. We'll also use some of this information to provide a greater service to our members and practitioners in the measurement. More specifically, we plan to update our website with several new services: 1) A research service in which members can find a listing of members in NCME that are involved (or would like to get involved) in similar research and contact information to facilitate collaboration; 2) A measurement resource service in which practitioners can locate by state and by area of content, individuals who would be willing to consult on the identified measurement topics; 3) information for graduate students concerning the different measurement programs and internship/employment opportunities; and 4) Contact for the Press, in which people from the press could contact NCME members who volunteer to serve as experts on particular topics in measurement.

Our Annual Meeting in Denver

I am very excited about our fast approaching annual meeting, May 1-3, Denver Colorado. Program Chairs, Bob Henson and John Willse, and Workshop Coordinator, Luz Bay, have worked very hard put together a great program. Also, key people from our Management Service, The Rees Group, including Drew Nelesen, our Meeting Planner, Nate Ehresman, Membership Coordinator, and Plumer Lovelace our Executive Director, have also put in many long hours getting ready for Denver.

Webcasting our Workshops

This year, in line with the conference theme, "Bridging the Gap Between Theory and Practice," we will be webcasting four workshops throughout the U.S. and internationally. Within the U.S. we will be involving accountability groups in Maine, New Hampshire, North Carolina, and Alaska. Internationally we will involve practitioners in Guatemala, Chile, Ecuador, Cameroon, South Africa, the Netherlands, and China. Luz Bay, Michael Rodriguez, Plumer and I have been working with a company called Sonic Foundry to put this together. Each site will have a facilitator and practitioners will be able to view both the speaker and their PowerPoints as well as ask questions via text. NCME will cover the costs this year as an experiment. We do understand the issues related to time differences and will closely evaluate our efforts and make recommendations for the future. I think this is so neat. I am really looking forward to this. To me, reaching out to these countries is at the core of NCME's mission. I hope to show pictures of participants at each of the sites at the breakfast.

Partnering with other Organizations

I am in the process of inviting presidents from several organizations to Denver and want to hold a separate meeting with them to talk about ways we can collaborate. These include the Association for Assessment in Counseling and Education (AACE), National Association of Testing Directors (NATD), Institute for Credentialing Excellence (ICE) and the Counsel on Chief State School Officers, CCSSO.

NCME Walk/Run

I have been working closely with Brian French on the NCME run. I will be posting on the NCME Website the highly reliable scoring rules for this competition. This year we'll have a competition for three classes: academia, testing companies, and other (state departments, government organizations etc.). There will be huge, impressive traveling team trophies and theta hat-hats (in some circles more valuable than an Olympic gold medal) for all people who participate. Speed or how quickly you can run the 5-K or walk the mile is not important. The goal is just to get people out talking, walking or running with one another. It should be fun. We'll just see which testing company and which university is in the best shape! Start training and let the trash talking begin.

No-Host Reception

This year again we are having the No-host reception with Division D. We're going to try something new and have measurement people in three different bands play at the reception. The Construct Underrepresentation, the UMass Error Band and from the U. of Kansas, The Skew. All I can say is "it'll be interesting." And yes, as you probably know, each of these bands is coming off of recent world tours.

Capturing the Conference

Plumer has promised to help take a lot of pictures for our website. Did you know that he is an expert photographer? To help Plumer "capture" the meeting we are going to have ten disposable cameras that people can pick up and take with them to sessions and the receptions to take pictures. NCME members can sign them out and then turn them back in at the NCME information booth. This will be an interesting experiment. All pictures will be posted on the NCME website.

Membership Ribbons

Like last year we're also going to have ribbons for everyone, the colors indicating number of years you've been a member of NCME. We also have a special ribbon for new NCME members. We would encourage new members to wear a ribbon so that we can help welcome you into the NCME family.

Conference Facilities

Because the Sheraton may be a bit confusing to get around, our Meeting Planner, Drew Nelesen, will have signage throughout our space with maps and directions. I have also enlisted the help of several of UNCG graduate students to serve as guides to help people find their way around. They'll be pretty visible wearing guide buttons and theta hat-hats.

Breakfast

At the breakfast I've decided to do a few things differently. I will be asking Board members to come early and greet everyone at the door. The Board will not be sitting up on risers and eating our breakfast but rather at tables with new members and everyone else. I've been working a lot on my talk and think you'll enjoy it.

I want to offer my sincere thanks to the many of you who have shared your thoughts, ideas and concerns with me throughout the year. See you in Denver!

Terry

A NOTE FROM THE EDITOR

Thanos Patelis, The College Board

It's true, your editor is way behind trying to keep pace with the wonderful suggestions from the vibrant NCME Newsletter Advisory Board. This will be Terry Ackerman's last article as president of NCME. At the business meeting Sunday, May 2nd, 2010, Terry Ackerman, UNC-Greensboro, will pass the gavel to president-elect Wayne Camara, College Board. This will put your editor of this newsletter in the awkward position of requesting deliverables from his boss! Seriously, we want to thank Terry Ackerman for his leadership of our organization and his openness and information in his columns. We look forward to Terry's continued contributions to our newsletter in his next role of past-president. New to the graduate student column is Dubravka Svetina from Arizona State University. We welcome her column and again thank Carol Barry from James Madison University and current College Board staff member for her efforts last year. In this issue, you will find comments from various members about the efforts to separate consequences from validity theory. I hope you will find the comments interesting and, more important, I hope the comments stimulate some healthy, collegial debates and discussions! Next, Michael Rodriguez, University of Minnesota, offers his experiences from Guatemala. Additionally, Dawn Mazzie, Lincoln (NE) Public Schools provides a piece on the need for pre-service in post secondary teacher education institutions that can help the quality of K-12 assessment. As a spotlight on our members, we have an interview with Ed Haertel of Stanford University and past NCME president. Included in this issues is an overview of the wonderful program that our program co-chairs, Bob Henson and John Willse, UNC-Greensboro, have put together along with a list of the training sessions by Luz Bay, Measured Progress. Last, but not least, Pete Swerdzewski, a Native Colorado, who grew up in Centennial, CO, offers us some nice things to do at our annual meeting in Denver! As always, please drop me an email with suggestions. Sincerely and at your service, Thanos.

THE ANNUAL MEETING OF NCME: FEW TIPS FOR NEWCOMERS AND ANYONE ELSE INTERESTED

Dubravka Svetina, Arizona State University

Our annual meeting is just around the corner. If you spend an hour listening to conversations in most measurement departments across the country (and some abroad), you will most likely hear the NCME's name echoing. Whether it is a professor mentioning to class that the week of April 30th the class time will be "project time," or you and your colleagues are searching for last year's poster template to prepare for your presentation, the NCME is in the air. As we enter into the final weeks prior to the conference, our already full schedules get bombarded by "final" manuscript revisions (hopefully getting in on time for discussants to read them). Add to that squeezing the presentation to 10-12 minutes, handout printing, poster making, and of course figuring out how the heck to transport that big and awkward poster protection tube. Ok, it might be a little bit stressful to prepare yourself for the event; however I consider it to be an extremely rewarding experience. I think it gives me a great opportunity to engage in research talks with people from other universities, build relationships and networking, and since I love to travel, it gives me an excuse to see yet another city. If you are new to the annual meeting, or returning graduate student, I hope you will find something valuable in this article to make your conference experience best it can be.

First time attendee? If this is your first conference (and you are early in your program), most of the sessions are likely to seem like a fast paced foreign language course. My advice is to still attend the sessions and try to get out of them as much as possible. Even if you have not had IRT course yet or ICC means little to you, being exposed to the language and topics that are currently being explored in the field is extremely valuable. Sometimes, the presenters will leave handouts or papers on the tables in the back of the room. Check them out if you are interested, or make yourself a note to contact the presenter for the full paper after the conference.

You cannot make it to every session. This is true weather you are returning to the conference for the nth time or are a first-time attendee. The program is large with a number of great topics that might be of interest to you. Take advantage to scan the middle of the program booklet to review the sessions. Also, try to choose sessions that are of most interest to you or sessions that address topics you would like to learn more about. Sometimes it is easier to search by topic, but sometimes it helps to figure out whose presentation you would like to hear first and then search by author.

Make a schedule ahead of time. It is a good idea to make a schedule of sessions and events you want to attend prior to arriving to Denver, CO. Schedule will be online soon, so you can start planning ahead. Investing some time doing this is well worth it. I typically use a spreadsheet to make my schedule, with time increments of 15 minutes for each of the days. When making my schedule, I write not only the session's name but also the room and hotel/location. This way I can see how much going back and forth from the hotels I'll be doing, and I won't need to flip pages of program to find out where I should be next.

Explore various (graduate student) events. There are several events/sessions that target graduate student population, such as the Graduate Student Issues Committee's Symposium and Graduate Student Poster Session. This year's topic for the Symposium is *The Influence and Impact of Technology on Educational Measurement.* The graduate student poster session is a wonderful opportunity to learn about your fellow colleagues' current research. Potential for emergent collaborations is abundant. I also enjoy several other events, including the NCME's Breakfast with presidential address (one of the largest events, if not the largest, where scholars and students gather together), the Fitness Walk/Run, and NCME and Division D Joint Welcome Reception.

Take time for yourself and make the best of the conference. Attending the conference can be stressful; between presenting, attending events, social gatherings, and networking. You want to bring your A-game, because this is one of the rare opportunities you get to interact with so many great people in the field. While going to sessions, talking with scholars and professionals about research, and polishing your own presentations are likely the main reasons why you are attending the conference in the first place, it is important to take breaks. Even if you take 15 minutes in a day to relax, breath in some fresh air, and reflect on your day and conversations you've had, you will be on a good path to make most of your conference experience.

COMMENTS FROM MEMBERS[1]: THE SEPARATION OF CONSEQUENCES FROM VALIDITY THEORY

Error of Measurement: Validity and the Place of Consequences

Gregory J. Cizek, University of North Carolina at Chapel Hill

For nearly 40 years, what has come to be called consequential validity has been a problem for measurement specialists. According to Brennan "the most contentious topic in validity is the role of consequences" (2006, p. 8).

The consequential basis of test use was formalized as a potential source of validity evidence in the influential chapter by Messick (1989). The notion spread, eventually making its way into the *Standards for Educational and Psychological Testing,* where "evidence based on consequences of testing" was codified as one of five possible sources of validity evidence (AERA, APA, NCME, 1999, p. 16).

From Possible to Primary

Once the camel's nose was into the tent, it eventually wanted title to the place. For example, it is now claimed that "the consequences of an assessment procedure are the *first and most important* consideration in establishing the validity of the assessment" (International Reading Association, National Council of Teachers of English, 1994, p. 17, emphasis added), and consequential validity is the only named, required source of validity evidence for state student achievement testing programs under the No Child Left Behind legislation: "In validating an assessment, the State *must* also consider the consequences of its interpretation and use (USDOE, 2007, p. 39, emphasis added).

Validity as Integration of Evidence

But let us consider the role that consequences are supposed to play in validation efforts. According to Kane, "validity is an integrated, or unified, evaluation of the [score] interpretation" (2001, p. 329). Messick classically defined validation as "an integrated evaluative judgment" based on all available evidence, including theoretical rationales, empirical relationships, and, yes, consequences (1989, p. 13).

Hmmmm... In contrast to these definitions, consider this: no publisher, no researcher, no state, no policy analyst—nobody—has ever produced an "integrated evaluative judgment" about validity in which consequences was included. Why not?

There are many reasons, but the most parsimonious explanation is that the consequences of testing have nothing to do with the accuracy of intended score inferences. Put simply, consequential validity doesn't exist. And, from a practical perspective, no integrated, evaluative judgment including consequences has ever been produced because the evaluation elements (namely, empirical evidence, theoretical rationales, and information about social consequences) cannot be synthesized.

[1] Contributors are listed in alphabetical order.

To illustrate the impossibility, consider the meaning of a total raw score on a test comprising 20 items measuring French vocabulary and 20 items measuring Geometry. An overall index of performance could be calculated, but a meaningful interpretation of, say, a score of 30 on the test is impossible, and no conclusions about the examinee's standing in either subject are supportable. Analogously, evaluative judgments based on integration of any empirical relationships among variables and information about social consequences are similarly uninformative. What integrated validity conclusion—that is, a conclusion about score meaning—could be reached when attempting to integrate disconfirming empirical evidence about the intended inference with information from test use that showed highly desirable social benefits? For example, how would one integrate evidence that an Algebra end-of-grade test is horribly aligned with the relevant content standards, but that, since the introduction of the test, graduation rates have increased? Can't be done. Evidence about the extent to which a test yields accurate inferences about a construct and evidence about the consequences of using the test are not compensatory in any logical sense and cannot be combined into a coherent, integrated evaluation. Any attempted integration confounds conclusions about both the scientific meaning of the performance and the desirability of using the test.

The simple truth is that consequential validity doesn't exist. A measurement error.

Next Steps

OK. Deep breath. Just because consequential validity doesn't exist, doesn't mean that the consequences of testing are unimportant. In fact, they are as important as support for the intended inference. The issue is not that consequences should be rejected altogether; it's that psychometric theory and practice must be reconceptualized so that consequences aren't wedged in with validity where the incorporation diminishes the sensibility of both concepts and precludes the enthusiastic pursuit of evidence bearing on either one. What is needed is a separation of evidence-gathering bearing on the meaning of test scores (that is, validation of score inferences) from evidence-gathering bearing on the desirability of using the test (that is, justification of test use).

A clean break from the fiction of consequential validity also brings the realization that there is much work to be done. There is broad consensus regarding four main sources of validity evidence and well-established traditions, standards and methods for validating score inferences. However, essentially no accepted frameworks or guidelines have been developed to formalize sources of evidence or procedures for justification of test use. One desirable result of differentiating between validity of score inferences and justification of test use is that rigor regarding both efforts will advance. Hopefully, the ultimate consequences of the distinction will be greater alacrity in gathering evidence about score meaning and initiatives that begin to develop formalized expectations for supporting test use.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement, (4th ed.).* (pp. 1-16). Westport, CT: Praeger.

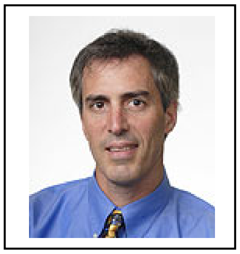
International Reading Association & National Council of Teachers of English. (1994). *Standards for the assessment of reading and writing*. Newark, DE: International Reading Association.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement, (3rd ed)*. (pp. 13-103). New York: Macmillan.

United States Department of Education [USDOE]. (2007, December 21). *Peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, DC: Author.

THE CONSEQUENCES OF EDUCATIONAL ASSESSMENT: WHO SHOULD EVALUATE WHAT AND WHY?

Jenna M. Copella and Stephen G. Sireci, University of Massachusetts Amherst



The Standards for Educational and Psychological Testing define validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, & NCME, 1999, p. 9). Given this widely accepted definition, it is no surprise why many people reject terms such as "consequential validity" and argue that the evaluation of the consequences of a testing program has nothing to do with test validation. However, as many test specialists have pointed out, how we support and defend the "proposed uses" of a testing program necessitates consideration of its effects. After all, it is hard to argue against analyses of the degree to which a test is fulfilling its purposes and whether it may be causing more harm than good.

Although he never coined the term or supported it, many people attribute "consequential validity" to Messick's seminal (1989) chapter in which he argued for a unitary conceptualization of validity. In describing this conceptualization, Messick used "two interconnected facets" (p. 20), which muddied the water enough that some test specialists saw "consequential validity" in the mire. Messick's first facet involved how the test was justified and was divided into (a) an evidential basis and (b) a consequential basis. His second facet involved the outcome of testing and was divided into (a) interpretation, or (b) use. His discussions of the consequential bases of test interpretation and test use encompass virtually all of the issues currently attributable to "consequential validity," and what the *Standards* refer to as validity "evidence based on consequences of testing" (AERA et al., 1999, p. 16).

In our view, the debate regarding "consequential validity" is a debate primarily over nomenclature. We do not support use of this term, but like Linn (1998), Reckase (1998), Shepard (1997), and others, we believe the evaluation of the consequences of a testing program, (or alternatively gathering validity evidence based on testing consequences, as the Standards describe it) is an essential part of any serious validation effort.

Importance of Considering Consequences

The importance of understanding the consequences of test interpretation and use is evident by the two special issues of *Educational Measurement: Issues and Practice* (*EM:IP*)devoted to the topic in 1997 and 1998. But *EM:IP* has not been the only venue for airing concerns about testing consequences. Using the ERIC database, with the search parameters "consequential validity," "consequences of testing," or "test consequences," we found nearly 90 technical reports, conference presentations, and journal articles dating as far back as 1970. Why are testing consequences getting so much attention? Because consequences represent the effects of test use. Sometimes consequences are positive and they are applauded. Sometimes they are negative and they are criticized, or even challenged in court. Like it or not, the study of testing consequences will persevere, because the effects of the use of a test for a particular purpose, whether positive or negative, intended or unintended, is how the worth of a test is ultimately judged by the public.

In considering the two special issues of *EM:IP* and the related literature, one conclusion is clear. Although there is disagreement over whether evaluation of consequences falls under the domain of validity (with Shepard (1997) promoting this notion and Popham (1997) rejecting it), there is complete agreement that evaluating consequences is important. Evaluating intended consequences is clearly important, since doing so is essentially evaluating the degree to which a test fulfills its purpose. But a comprehensive analysis of consequences as validity evidence is much deeper and requires exploring unintended consequences, too (Messick, 1989). Consequences related to test interpretation may involve investigating the effects of labeling our school children as "failing" or "below proficient" on their academic self-concept. Consequences related to test use might involve investigating the effects of a testing program on high school dropout, entry into a profession, or adverse impact in hiring or promotion. It is also important to note that some *unintended* consequences may actually be *positive* (Cizek, 2001), which illustrates the breadth at which consequences should be studied.

A very recent example of the importance of testing consequences was provided by Penfield (2010), who pointed out "several states…and large school districts…retain students in gateway grades primarily on the basis of performance on standardized tests" (p. 110). Validating the use of such tests for this purpose requires longitudinal study of whether the retention decisions are beneficial to the students and to the schools. The consequences certainly could go either way in such a study, and although Penfield did not describe his inquiry as an analysis of consequential validity, it is exactly the sort of validity endeavor Messick (1989), Shepard (1993, 1997), Linn (1998), and others would applaud.

Providing Validity Evidence Based on Testing Consequences

The literature supports our claim that evaluating testing consequences is important (AERA et al., 1999; Cronbach, 1971; Lane et al., 1998; Linn, 1998; Messick, 1989; Shepard, 1993, 1997, etc.). However, what types of consequences should be evaluated and what types of evidence should we gather? Like all aspects of validation, it depends on the purpose of the test and how the test is being used. Seminal validity theorists such as Cronbach (1971), Messick (1989), Kane (1992, 2006), and Shepard (1993) promoted validation as hypothesis testing, with Kane and Shepard both prioritizing the hypotheses to be tested. Thus, if the purpose of a test is to improve student achievement, hypotheses of positive effects such as improved student achievement and improved instructional practices should be investigated as well as hypotheses related to negative consequences (narrowing of the curriculum, dropout, decline in teachers' morale, etc.).

One way to assess negative consequences is simply to listen to criticisms of testing and devise ways of testing if the criticisms were valid. For example, criticisms that tests are biased against subgroups of test-takers have generated hundreds of studies of differential predictive validity, measurement invariance, item invariance, and other explorations of adverse impact. A consequence of these investigations into the legitimacy of these testing consequences is improved measurement practices and increased concerns of fairness throughout the test development process.

Surveys and focus groups of invested stakeholders such as teachers and students is one way to assess testing consequences. Tracking student achievement over time and other indices such as graduation rate is another. However, Lane et al. (1998) suggest a more comprehensive approach that uses multiple sources of evidence "such as evidence obtained through the analyses of classroom instructional and assessment activities" (p. 25). We support such endeavors, and believe that longitudinal studies of instruction—curriculum alignment, and of instruction-assessment alignment might be one productive way to evaluate the effects of tests and curricular reforms on instruction (Martone & Sireci, 2009).

Who Should Gather Validity Evidence Based on Testing Consequences?

Given the importance of studying the consequences of testing, an important question is, "Who should do it?" Likely candidates are test developers (e.g., a test development contractor), testing agencies (e.g., a state department of education), and test users (e.g., a school district using a commercial test). Virtually all partners in the educational process, including measurement professionals, could, and probably should, be involved. However, studying testing consequences is quite a bit harder than computing a coefficient alpha or Mantel-Haenszel statistic. Thus, although the endeavor is important, it is likely to be expensive and time consuming, with few volunteers stepping forward to do it.

Nevertheless, the responsibility remains, and there are good examples of comprehensive assessments of testing consequences (e.g., Lane et al., 1998; Taleporos, 1998). Test developers are responsible for providing evidence that a test measures what it claims to (evidence based on test content), and according to the *Standards*, they are also responsible for warning against inappropriate test use (AERA et al., pp. 17-18). However, test developers cannot be expected to conduct longitudinal studies of tests they develop for one purpose that are ultimately used for another, unanticipated purpose. For this reason, development of clear and comprehensive statements of the purpose and intent of a testing program is critical.

Messick (1989), Popham (1997), and Shepard (1993, 1997) argue that much of the responsibility for investigating the consequences of test use lies with test users. Messick referred to this as an ethical responsibility of the test user, because they are in the best position to evaluate the value implications specific to their setting. In some instances test users adapt existing tests for a new purpose. In these circumstances, the test user is clearly responsible for evaluating testing consequences.

Some authors have proposed the development of frameworks designed to engage multiple parties in the process of investigating consequences of test use. Abu-Alhija (2007) described the concept of critical testing in the arena of large-scale assessment, which is a collaborative procedure used to monitor the unintended and negative consequences of testing programs. Test developers and users, along with test takers, teachers, and the public, are charged with the task of challenging "the content and the uses of tests and to critique the principles and values that are inherent in them" (p. 65). Although this process is proposed to look for negative issues associated with large-scale assessment, it can clearly be extended to include intended consequences as well. Lane et al. (1998) offered a detailed description of a framework for evaluating the consequences of state-wide assessment programs. This framework involves students, teachers, and administrators and collects information at the state level, district level, and school/classroom level.

Much Work to be Done

Although we applaud the efforts of Abu-Alhija (2007), Penfield (2010), Lane et al. (1998), and Taleporos (1998), we note that most of the activity in evaluating testing consequences falls far short of what is expected in a serious and comprehensive validation effort that would justify the use of a test for high-stakes purposes. Educational tests are currently called on to meet lofty goals such as improving student learning and thus improving society. Are they meeting these noble goals? Or, are they causing more problems? Only comprehensive and longitudinal studies of their consequences will provide the answer. Let's get to it!

References

Abu-Alhija, F. N. (2007). Large-Scale Testing: Benefits and Pitfalls. *Studies in Educational Evaluation, 33*(1), 50-68.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20* (4), 19-27.

Cronbach, L. J. (1971). Test Validation. In R.L. Thorndike (Ed.) *Educational measurement (2nd ed*., pp. 443-507). Washington, D.C.: American Council on Education.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*,527-535.

Kane, M. (2006). Validation. In R. L. Brennan (Ed). *Educational measurement (4th edition*, pp. 17-64). Washington, DC: American Council on Education/Praeger.

Lane, S., Parke, C. S., & Stone, C. A. (1998). A Framework for Evaluating the Consequences of Assessment Programs. *Educational Measurement: Issues and Practice, 17*(2), 24-28.

Linn, R. L. (1998). Partitioning Responsibility for the Evaluation of the Consequences of Assessment Programs. *Educational Measurement: Issues and Practice, 17*(2), 28-30.

Martone, A., & Sireci, S. G. (2009). Evaluating Alignment Among Curriculum, Assessments, and Instruction. *Review of Educational Research, 4*, 1332-1361.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement (3rd ed*., pp. 13-100). Washington, DC: American Council on Education.

Penfield, R. D. (2010). Test-based grade retention: Does it stand up to professional standards for fair and appropriate test use? *Educational Researcher, 39*, 110-119.

Popham, W. J. (1997). Consequential validity: Right Concern-Wrong Concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Reckase, M. D. (1998). Consequential Validity From the Test Developer's Perspective. *Educational Measurement: Issues and Practice, 17*(2), 13-16.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Shepard, L. A. (1997). The Centrality of Test Use and Consequences for Test Validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8.

Taleporos, E. (1998). Consequential Validity: A Practitioner's Perspective. *Educational Measurement: Issues and Practice, 17*(2), 20-23.

CONSEQUENCES AND VALIDITY: AN UNEASY RELATIONSHIP

Kurt F. Geisinger, Buros Center for Testing

Since some of the writings of one of our psychometric deities, Samuel Messick (1980, 1988, 1989), there has been considerable discussion of the interplay of the consequences of testing and the validity of testing. Consequences have always had a role in validation (Kane, 2006). Intended consequences after all include criteria, the prediction of which provided the basis for what was probably our first type of validity, what we now call criterion-related validation evidence or the like. Traditionally, the most serious negative consequence of testing was adverse impact on ethnic minority group members—this is what most concerned socially conscious psychometricians , especially in college and graduate/professional school admissions and employment settings (Messick, 1989, 1994).

Debate exists over whether Messick's definition included evidence about all consequences as validity, although there is no doubt he saw such impacts as critically important. On the other hand, other validity theorists have argued validity should only be related to score interpretation (see Kane, 2006, p. 54). I share Popham's (1997) view that consequences are all always important; but importance per se does not transform them into validity evidence!

I believe that we must differentiate the more important considerations as UTILITY, of which validity is but a major part. Utility has long been an issue for industrial-organizational psychologists. Perhaps the most famous study of the utility of tests as primarily used in industry was Cronbach & Gleser (1965). Modern psychological reviews of personnel tests suggest that many aspects need to be included in their evaluation: price, administrative costs, cost of legal actions, costs of scoring, and so on. Many of these factors also impact education. Both intended and unintended consequences of testing should be factors in considerations of utility. Such factors as the impact of testing on protected groups, as well as the impact of testing on instructional time, on students, and on their overall education demand our careful consideration.

The Standards for Educational and Psychological Testing (1999) list consequences as one of five characteristics that can be looked at in validation. They are:

1. Evidence based on Test Content;
2. Evidence based on Response Processes;
3. Evidence based on Internal Structure;
4. Evidence based on Relations with Other Variables;
5. Evidence based on the Consequences of Testing.

However, the Standards also report in the commentary section regarding consequences: "It is important to distinguish between evidence that is directly relevant to validity and evidence that may inform decisions about social policy that falls outside the realm of validity" (p. 16). The commentary then goes into the discussion of group differences—as noted above, historically the key issue in regard to the consequences of testing.

Brennan (2006), in discussing the role of consequences focuses on what is perhaps the most common current definition of validity based on proper interpretation of test scores, stated, "Since it is now almost universally agreed that validity has to do with the proposed interpretations and uses of test scores, it necessarily follows that consequences are part of validity" (p. 8). Does it? Are we concerned about both intended and unintended consequences or does each type have a different interpretation? Brennan has suggested that we consider consequences as intended and unintended, positive and negative, essentially as a 2 x 2 table. Intended positive consequences are virtually always positive indications of validity. Intended consequences that are found to be negative are almost always detrimental to the argument relative to the validity of a test for a particular use (or at least to their utility). However, it is less clear how to deal with unintended positive contributions. My argument is that we must determine whether the consequences relate to the intended uses of the test; if they do, then they impact validity; if they do not, then they impact utility as mentioned above. A simple example is that a particular test may be frightfully expensive. Its use depletes the treasures of the group using it. This negative consequence obviously has nothing to do with validity of the test per se.

Regardless of whether they are validity related or not, we wish to minimize negative consequences of testing. Identifying who is responsible for trying to reduce the unintended negative consequences of a test is not an easy process. Brennan calls for test publishers to enumerate possible misuses directly for users as a method for helping prevent test misuse and also to instruct users on best practices. Can a test publisher ultimately be responsible for the work of a test user? In many if not most cases, such monitoring of test users would be virtually impossible. On the other hand, not many test users have the ability to perform validity studies or studies of test consequences. Does it fully protect a publisher to state that scores cannot or should not be used in a given manner, does it insulate the publisher effectively? Should test publishers withhold test scores from a given user whom the publisher believes will use test results inappropriately? Brennan discussed some limitations in preventing misuse:

"Although there may be a logic for including unintended negative consequences under the umbrella of validity, I would maintain that there is often very little that the measurement profession can do on its own to successfully prevent or remedy such misuse. In this regard, it is worth recalling that the educational testing profession has no mechanism for disciplining those individuals or entities who use test scores inappropriately" (p. 8).

Actually, within the American Psychological Association, there are prohibitions about certain types of misuse, but these instances are probably not the worst misuses. There are actually reasons why we need lawyers and the courts, and some suspected test misuses may represent one such justification! On the other hand, I have always maintained that our profession needs to police itself (and it almost always does!)

In conclusion, I strongly recommend the consideration of a wide variety of consequences—intended and unintended—as part of the decision to use or continue to use a particular test. I urge test users to review a wide range of validity studies of a particular test use prior to their decision to use a test so that they can see how the test impacts as wide a variety of consequences as possible. I also believe that we should return to use the word utility as a broader concept than validity and one that permits us to consider the many impacts of testing that are vitally important but beyond the point of view of validity. On the other hand, I don't like or use the term, "consequential validity;" as many of the uses of the term bring into account things that are not part of validation per se, at least in my opinion. Again, I am certainly not saying that we should ignore such factors, only that their impact should be on utility, not validity.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Brennan, R. (2006). Perspectives on the Evolution and Future of Educational Measurement. In R. L. Brennan (Ed.) *Educational measurement* (*4th ed*.) (pp. 1-16). Washington, DC: American Council on Education/Praeger.

Cronbach, L.J. & Gleser, G. C. (1965.) *Psychological tests and personnel decisions (2nd ed.)* Urbana, IL: University of Illinois Press.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.) *Educational measurement (4th ed.)* (pp. 17-64). Washington, DC: American Council on Education/Praeger.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012-1027.

Messick, S. (1988). The once and future issue of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity*. (pp. 33-45). Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)* (pp. 13-103). New York: American Council on Education & Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 45*, 5-8.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

THE CONSEQUENCES OF TEST USE

Joanna S. Gorin, Arizona State University

That consequences of test use are of critical importance is hardly debatable. However, their place in the validity argument as opposed to supplemental investigation is less clear. As a relatively novice validity theorist and a part-time test developer, I tend to disagree with Messick's (1989) inclusion of evidence of consequences of test use as part of validity examinations. To the extent that appropriate use and interpretation of test scores is derived from the meaning and veracity of test scores as reflections of the construct of interest, the consequences of test use are irrelevant. I fear, however, that removing their connection to validity may lead researchers to ignore consequences all together, which would be detrimental to ethical and fair assessment practice.

THE ISSUE OF SEPARATING CONSEQUENCES FROM TEST VALIDITY THEORY

Ida Lawrence, Brent Bridgeman, Michael Kane and Michael Zieky, Educational Testing Service

Thanks for the opportunity to comment on the issue of separating consequences from test validity theory. I took the liberty of asking three of my ETS colleagues (Brent Bridgeman, Michael Kane and Michael Zieky) to provide their thoughts. Here is our collective response.

We feel that the question does not lend itself to a yes/no answer. The answer really depends on whether the consequences are intended or not, whether they are due to a defect in the test or to some other cause (e.g., ineffective instruction), and whether claims are being made about outcomes.

If claims are made, those making the claims should presumably be prepared to justify them. The responsible party could be the test developer or it could be the test user. Some tests and test-score uses are closely tied to outcomes/consequences, and their validation has always focused on these outcomes (e.g., placement tests, selection tests for employment). The intended consequences (claimed outcomes) of such tests are clearly relevant to the validity argument. On the other hand, unintended consequences may or may not be relevant to the validity of the test. They are relevant if they result from a defect in the test; they are not relevant if they are due to other factors.

The issue of consequences is addressed in several parts of the *ETS Standards for Quality and Fairness* (2002):

- Standard 6.4 (p. 30) indicates that, if relevant and feasible, the validity argument should include "evidence that the program's claims about the direct and indirect benefits of assessment use are supported…." Though the word "consequences" is not used, evidence concerning the benefits of test use clearly is based on the consequences of testing.

- Standard 6.6 (p. 31) requires us to investigate the consequences of assessment "If the use of an assessment results in unintended consequences for a group that is studied." We have to "review the validity evidence to determine whether or not the consequences arise from invalid sources of variance." If they do, we have to "revise the assessment to reduce, to the extent possible, the inappropriate sources of variance." Standard 4.1 (p. 18) also requires information about unintended consequences for studied groups.

- Standard 12.4 (p. 58) requires us to "provide information and advice to help interested parties evaluate the appropriateness utility, and consequences of the decisions made on the basis of assessment scores."

These standards suggest that there are certain circumstances in which including the consequences of testing as evidence of validity makes sense.

But the issue is really quite complicated because there are circumstances where failure to meet consequences does not necessarily invalidate a test. For example, a test may fail to have the intended consequences through no fault of the test. Obtaining the intended consequences often depends on a complicated chain of events. The failure of any link could keep a highly valid test from meeting its intended consequences.

Formative assessment provides an example of this. The purpose of a formative test may be to reduce failures on NCLB assessments. The formative test may work well in identifying students who are at risk of failure, and may even identify a student's weaknesses correctly, but if ineffective remedial instruction is provided, the failure rate will not change. The instruction has to be fixed, not the test. The real question is, can you justify the claim for the test? If the claim is that the use of the test will improve performance on the NCLB tests, then there should be validity evidence to show that such improvement occurs as a result of administering the formative assessment. If the claim is that the use of the test will identify students who are at risk of failure and identify their weaknesses, then there should be validity evidence to show that the test does those things.

The issue boils down to not separating the validity of the test from its use. We all agreed that consequences should be included as evidence for test validity in the sense that claims that are made should be justified. If implicitly one is saying that a test has positive consequences, then the claim should be justified. We also feel that test users should be responsible for what they do, and therefore have responsibility for evaluating consequences, both negative and positive, and then making appropriate decisions based on such evaluations.

References

Educational Testing Service. (2002). *ETS Standards for Quality and Fairness*. Princeton, NJ: Author.

COMMENT ON THE ROLE OF CONSEQUENCES IN VALIDITY EVALUATIONS

Scott F. Marion, National Center for the Improvement of Educational Assessment

The unified view of validity, as represented in the 1999 version of the joint standards (AERA, APA, NCME, 1999) considers the consequences of the assessment as a legitimate and important source of validity evidence (e.g., Lane and Stone, 2002; Messick, 1995; Shepard, 1997). Kane's argument-based approach (2006) further solidified this view with its focus on "plausibility of the proposed interpretations and uses (p.23)".

The focus on use and utility is the key. I believe I understand the desire for moving consequences into an extra validity criterion, in that if the test measures the construct appropriately and leads to technically defensible inferences, then the test maker has done their job. However tests do not operate in a vacuum. Tests, particularly the current generation of summative assessments, function in high stakes accountability systems where the results often have considerable consequences for either those tested or others in the system (e.g., teachers or schools). The test developer may say, "I can't help it if policy makers do silly things with my good test." That is a weak and perhaps irresponsible argument. Anyone who has ever taken an introductory measurement course knows that test design and validation starts with a careful exposition of the purposes and intended uses of the proposed assessment. It is hard to think of a test use without any potential consequences. Even a low stakes classroom assessment has consequences if students are incorrectly denied remediation or other learning supports. If virtually all tests results have some sort of consequences, then it does not make sense to separate consequences from the validity evaluation. Now, if someone uses a test in ways in which it was clearly never intended and documented to be used, resulting in negative consequences, the burden for evaluating consequential evidence should shift from the test developer to the user, but the requirement does not disappear. Anticipating the uses and potential consequences of tests under the next version of ESEA reminds us that consequential evidence must remain an integral part of the validity argument. Anything less suggests that use can be separated from validity and most would agree that just does not make any sense.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). New York: American Council on Education/Macmillan.

Lane, S., & Stone, A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21*(2), 23–30.

Messick, S. (1995). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practic, 16*(2), 5–24.

CONSEQUENTIAL VALIDITY: RIGHT CONCERN—WRONG CONCEPT

W. James Popham, University of California, Los Angeles

In 1997 I published an essay in *Educational Measurement: Issues and Practice* using the above title. More than a decade later, about a week ago, I re-read that essay. Happily, I agree today with everything I wrote way back then. Face it—when you're right, you're right!

Here was the nub of my 1997 argument: Assessment validity is, by a mile and a half, the most important concept in our field. We measure folks so we can come up with valid, that is, accurate, inferences about what's going on inside those people. Phrasing it differently, we use test-takers' overt responses to tests so we can arrive at defensible interpretations about those individuals' covert status such as their knowledge, skills, and attitudes. Validity refers to the accuracy of the test-based inferences we make about people. Validity is not an attribute of tests, but of the inferences we make by using tests. Without validity, educational measurement would be silly.

Nonetheless, the consequences of test usage are important, staggeringly important. When we try to fold test-use consequences into the notion of validity, however, we add murk to an otherwise lucid "accuracy-of-inferences" concept. Most psychometricians won't be baffled by such conceptual clutter, but almost all in-the-trenches educators are sure to be confused by an idea of validity that simultaneously mixes inference-accuracy and the merits of test-use consequences. Let's just keep these two important considerations separate. The potential confusion caused if we were to sanction "consequential validity" is too big a price to pay. It's a consequence that's too costly.

LEARNING TO ASSESS STUDENT LEARNING IN AMÉRICA LATINA

Michael C. Rodriguez, University of Minnesota

The United States Agency for International Development (USAID) granted a 4-year award in 2005 to the *Programa Estándares e Investigación Educativa Guatemala* (Education Standards and Research Program), as part of its Central America and Mexico Regional Strategy to improve educational outcomes. The USAID-Guatemala personnel work closely with the Ministry of Education, local universities, and related government and community organizations, providing financial and technical assistance to improve education outcomes and inform national education dialogues. The project produced national K-12 curriculum standards and standards-based assessments in language arts and mathematics. Throughout the 5 years of the project, USAID-Guatemala and the Ministry of Education encountered significant challenges, some of which could be addressed by current measurement theory and practice, some of which could not.

The project involved consultants to provide needed expertise. Since 2007, I have worked with the USAID-Guatemala team by providing psychometric technical assistance and training. A symposium discussing the challenges and lessons learned in this project and related efforts by others in Honduras and Chile will be presented at the 2010 NCME meeting.

Assessment has a long history. Although selection and placement assessments can be traced back to 2200 BC in China, national assessments of student learning are much more recent. In the USA, the technology of testing (scoring, scaling, equating) has been advanced through national assessment efforts, in part fueled by the NAEP assessments during the 1960s. The 1983 publication of *A Nation at Risk* was a call to arms in assessment. Successive reauthorizations of the Elementary and Secondary Education Act have required greater attention to learning and assessment of student performance, particularly the 2001 reauthorization (NCLB).

Internationally, there is a wide range of experiences in the assessment of learning. In much of Europe, Asia, Australia, and a few other regions, there is a long history of assessment. However, many nations around the world, particularly throughout the Caribbean, Central and South America, and Africa, have limited experience with large-scale assessment. In many nations, education reform efforts attend to more basic needs, including the structural safety of schools, effective resource allocation, professional development of teachers, establishment of curriculum standards, and equitable access to schools.

In a recent analysis of USAID assistance to basic education in the developing world, Chapman and Quijada (2009) reviewed 33 projects that were awarded over $2.3 billion. Some have argued that international development assistance to education efforts is misguided, that addressing more urgent needs like health and food security produces a more certain return on investment. Chapman and Quijada found 15 projects involved test development. The area of test development is where professional expertise is limited, particularly in Central America. Although there are high quality postsecondary institutions throughout the region, there are no doctoral programs in psychometrics, providing no opportunity for local expertise development in this area. International development assistance provides one opportunity to secure such expertise, through technical assistance, technical consultation, and professional development training.

A limited history of national assessment has presented an immense challenge – the language of large-scale assessment is absent in education agencies, schools, communities, and families in Central America. A great amount of institutional capacity building has been required. Fortunately, many countries in the region are experiencing new goal setting and decision making capacity within their respective Ministries of Education. In Guatemala, this has required the creation of additional bureaucracy and regulation making procedures. The technical challenges have been particularly acute, especially given the limited access to professional training opportunities. Overall, local capacity building has been evident and assessment technological advancements are seen as programs move toward modern measurement theory.

The USAID-Guatemala project has developed the current testing system through an IRT measurement model and has provided training for staff within the Ministry on modern measurement theory and advanced statistics. They continue to faces challenges, including the multi-language assessments in early primary levels, extreme limited opportunities to learn, and limited teacher preparation and professional development. Guatemala has a population of nearly 14 million where more than 50% speak one of 24 Mayan languages as their first language. The nation has worked very hard to provide bilingual (Spanish & Mayan) education for the first three years of school but few native Mayan speakers obtain teaching certificates, resulting in stark inequities in education quality. Most of the Mayan population live in rural areas where the schools have very few educational resources compared to those in the urban areas, creating additional inequities. Also, the educational system has serious efficiency problems. About 34% of the students fail their first year; 42% finish primary school; less than 10% finish secondary school. Moreover, teacher preparation occurs at the secondary level, not postsecondary.

The recently developed national curriculum standards were designed to serve several purposes, including establishing clear curriculum and performance goals for each grade and equalizing the quality of education. Before the year 2006, Guatemalan achievement tests where norm referenced and sample based, designed through university-based evaluation projects. Since 2009, the national assessments are administered annually near the end of the school year in grades 1, 3, 6, 9 and 12, addressing Mathematics and Language Arts standards. Operational forms have been developed through a common-item linking design to facilitate equating across years. The tests, linking, equating, and standard setting, have all been supported through the Rasch measurement model. A technical manual was developed during the assessment development process and was used as a guide to evaluate the degree to which each step was consistent with the *Standards for Educational and Psychological Testing*.

USAID staff have addressed several issues that became more complicated because of extreme limited opportunity to learn. For example, assessment design has been difficult with respect to assessing context effects on common items that change position across forms and years. In the context of standards-based tests where there are significant opportunity-to-learn concerns and where a large number of students do not complete the test or tend to skip items, changing the location of items beyond one page may create complications in equating and maintaining item performance consistency and score stability over time.

The choice of IRT measurement model was also complicated. The Ministry of Education considered pattern scoring and the use of a 2PL or 3PL model. For individual examinees, pattern scoring may have a positive effect, while for others, it could have a negative effect. For example, in a few areas, Mayan numbers are covered well and students learn and practice the use of this number system (a first grade standard). In many regions and schools, students do not spend much time learning or practicing the Mayan number system. Item parameters will reflect this by resulting in high levels of difficulty with low levels of discrimination. These items will not discriminate between high and low ability students because performance is not a function of ability but is a function of opportunity to learn. Now, ability scores (person-theta values) will not be a function of the items covering Mayan numbers because the low discrimination will be used to weight these items less than the other items. Students who get all of the Mayan-number questions correct will not get as much credit toward their ability score as students who may get other questions correct and all of the Mayan-number items incorrect. Equating assumes that the same content is being covered from year to year, and although it is not part of the assumptions for estimation, differential opportunity to learn has implications for the validity of equating.

Because standard setting, as a conceptual framework and a process, was new to Guatemala, the first implementation became an opportunity to investigate the methodology. The Bookmark method was selected as the most appropriate method, well suited to setting standards on multiple-choice tests and one that could implemented with sufficient fidelity. It was important to provide validity-related evidence regarding the appropriateness and feasibility of employing a standard setting method in Guatemala. As a learning opportunity, a small study was designed to investigate variability in 3rd grade results from replications of the process, including three independent panels for Language Arts and three for Mathematics. There were significant differences in some cases, including cut score differences as large as 0.44 logits on the Rasch scale. To some degree, participant experiences and perceptions of the process helped explain variation within and between panel results (Rodriguez, Rego, & Rubio, 2009).
The USAID-Guatemala project is an important example of how the educational measurement community has a great deal to contribute to and learn from education reform and assessment efforts in developing countries, particularly Central America. For a general overview of the USAID-Guatemala projects, visit: http://www.usaid.gov/gt/health_education.htm.

References

Chapman, D.W., & Quijada, J.J. (2009). An analysis of USAID assistance to basic education in the developing world. *International Journal of Educational Development, 29,* 268-280.

Rodriguez, M.C., Rego, O., & Rubio, F. (2009, April). *Examining variation in independent replications of the Bookmark standard setting method on two tests*. Paper presented at the annual meeting of the National Council on Educational Measurement, San Diego, CA.

How can post-secondary, pre-service teacher education programs improve the quality of K–12 classroom assessment?

Dawn Mazzie, Lincoln (NE) Public Schools

How can post-secondary, pre-service teacher education programs improve the quality of K–12 classroom assessment? Although I do not have all the answers, my experiences providing professional development to teachers in low performing schools in an eastern state in this area have given me insight on this topic. However, in order to make suggestions for improvement, issues must be identified. I have observed two major factors that negatively affect K–12 classroom assessment: (a) teachers sometimes lack content knowledge and (b) teachers have little or no knowledge of appropriate classroom assessment techniques.

Some teachers lack in-depth content knowledge, which hinders their ability to teach the content to the level of most national standards. When teachers do not completely understand the standards one of three things happen: a) they ignore portions of the standard that are beyond their understanding, b) they create an interpretation of the standard, or c) they ignore the standard completely (Llosa, 2005). This deficit in knowledge also hinders teachers from creating assessments that provide an accurate picture regarding student achievement. To create quality assessments teachers need to have more than a surface-level understanding of the subject matter. As I reviewed teacher-assessments in my previous position, I found that many teachers wrote items only at low cognitive levels probably because that was the extent of their understanding of the content. This is consistent with findings of Sobolewski (2002) who found that 82% of the questions posed by teachers during 80 classroom observations were at low cognitive levels. This lack of content knowledge also restricts the teacher's ability to proceed or adjust instruction as necessary based upon student needs. Having a deep and thorough understanding of the content should be a priority if the goal is to address students' instructional needs.

Teachers also lack knowledge of appropriate classroom assessment techniques. This, coupled with gaps in teacher content knowledge, hinders their ability to make valid and reliable inferences about student learning. Often teachers enter the classroom with a very limited knowledge of assessment techniques (Crooks, 1988; Shafer, 1993; Stiggins, 1992). In fact, their only formal experience with assessment may be as a student during their own schooling. This is a problematic because many times students believe the only reason they are assessed is to determine what grade the teacher will put in the grade book (or report card). Often there is a disconnect for students (and teachers) between the grade assigned and the amount of learning that has taken place. When teachers are not skilled in assessment there is a tendency to assess compliance, effort and behavior rather than the amount of learning that has taken place.

Teacher education programs can work to remedy these issues. First, teacher education programs can require additional subject specific courses that focus not only on the subject matter but also how to convey the information to others. Courses also might consider using samples of student work in the content area in an effort to help pre-service teachers begin to think about how they might approach reteaching the content when necessary. Reteaching learning targets is often a reality, and pre-service teachers need strategies for doing this such that they do not present the same material in the exact same way. National K–12 academic standards in particular content areas can be used to guide the instruction in the post-secondary classroom so that teachers are learning as their future students will learn. Certainly these courses will have to go beyond the K–12 standards to make sure the teacher has sufficient knowledge to facilitate student learning. If teachers are familiar with the standards prior to walking in their own classroom, I believe they will be better able to prepare students to meet those standards and better design assessments to measure student learning of those standards.

Teacher education programs need to require a course or courses dedicated to classroom assessment as a requirement for graduation. Although some institutions offer and require pre-service teachers to complete assessment courses, these institutions are rare (Schafer & Lissitz, 1987, 1988). In these assessment courses, pre-service teachers also need to be given the opportunity to practice these techniques with students. Giving pre-service teachers opportunities to actually practice administering assessments, collecting samples of student work, and using results for decision making will give them a better idea of what works and what needs to be changed in both the assessment itself and what may need to be changed in instruction. One of the most powerful activities I completed when I was in graduate school was an activity where I had to create an objective assessment and then administer it in an actual classroom of students. The results of that assessment enlightened me to student thought processes that I had not been expecting; things that I thought were clear were, in fact, very unclear. For this particular assignment, I had to use the students in another teacher’s classroom but I think it would have been more powerful if I had been the one responsible for instruction. I think this kind of activity should occur at least a couple of times during the pre-service training. (Actually, I think reflection on assessment results should occur anytime an assessment is administered but for the purposes outlined here I am suggesting only a couple of formal reflections.) The first formal reflection should occur during a practicum or internship where the pre-service teacher is observing a classroom. They would observe the instruction leading up to a particular assessment, observe the administration, the scoring, and then review the results. Then reflect on the results obtained to figure out what went well and what should be changed. I believe this practice should be incorporated in student teaching when the pre-service teacher is responsible for the instruction. Administering assessments to students, analyzing the results, and reflecting on what worked well and what needs to be changed, and developing a re-teaching plan would be a powerful process for pre-service teachers. Moreover, the education professors can provide a support system for students in reviewing student work (what is the work telling us) and guiding students where to go next (how do I re-teach?).

In addition to requiring a course in classroom assessment, post-secondary institutions can provide onsite assessment training to all professors so that they have the skills necessary to create classroom assessments that provide them with data they can use to make valid inferences about student learning. Thus, increasing the overall knowledge and understanding of all graduates, not just pre-service teachers. This will facilitate professors modeling appropriate assessment techniques in their own pedagogy classes. This process will help pre-service teachers understand the power of appropriate assessment and feedback. It is like the old Chinese Proverb “Tell me and I'll forget; show me and I may remember; involve me and I'll understand." If teachers are involved with good assessment techniques, even as students, they will understand the power of these techniques and are more likely to use them in their own practice. If improving classroom assessment is a priority for pre-service teacher education programs, applying at least some of these suggestions is very important. We need graduates who have a strong grasp of the content and who are knowledgeable about the use and impact of good assessment techniques in the classroom. When teachers enter the classroom without proper content and assessment skills, they may focus on lower level skills in the standards and create assessments that do not align to the desired learning targets. Often these assessment results contradict results on large-scale assessments. When this happens, teachers, parents, and students are receiving conflicting information about student achievement. When teachers are trained in creating assessments that are aligned to the established learning targets and yield valid and reliable results, learning can be accurately documented and students can be lifelong learners and as a by-product, they will also be better prepared to achieve on large scale assessments.

References

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*, 438-481.

Llosa, L. (2005). Assessing English learners' language proficiency: A qualitative investigation of teachers' interpretations of the California ELD Standards. *The CATESOL Journal, 18*(1), 7-18.

Shafer, W. & Lisstiz, R. (1987). Measurement training for personnel: Recommendations and reality. *Journal of Teacher Education, May/June*, 57-63.

Shafer, W. & Lisstiz, R. (1988). *The current status of teacher training in measurement*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Shafer, W. (1993). Assessment literacy for teachers. *Theory Into Practice, 32*(2), 118–26.

Sobolewski, K. B. (2002). *Gender equity in classroom questioning*. Unpublished doctoral dissertation, South Carolina State University, Orangeburg.

Stiggins, R. (1992). High quality classroom assessment: What does it really mean? Instructional Topics in Educational Measurement, 12, 211–215.Rodriguez, M.C., Rego, O., & Rubio, F. (2009, April). *Examining variation in independent replications of the Bookmark standard setting method on two tests*. Paper presented at the annual meeting of the National Council on Educational Measurement, San Diego, CA.

SPOTLIGHT ON THE PEOPLE WHO MAKE OUR ORGANIZATION GREAT – ED HAERTEL

Thanos Patelis, The College Board

Why do people continue to come to our annual conference? Why do we have so many members that once they become members and come to the conference continue to do so every year for decades? Well, it’s not only the quality, innovation, and utility of the content, but the people themselves. So, in an effort to get us to know the people, each issue will offer excerpts from interviews of the people who make our organization great.

This interview is with Ed Haertel currently at Stanford University and past president of NCME. In an effort for us to get to know Ed, we asked him a number of questions and the excerpts of his responses are provided below.

How did you get into this field?

Ed: I started as an undergraduate math major at the University of Wisconsin. During that time, I had an hourly job working at the Wisconsin Research and Development Center. After I received my bachelor’s degree, I started working there full time as a technical specialist in data processing. It was there where I become interested in going on to graduate school. So, I applied to graduate programs in philosophy and education. The graduate programs in philosophy did not show much interest, which sort of made sense since I had only taken a few courses in the subject. But, the education program at the University of Chicago was very interested and made me a wonderful offer that I accepted, in the Measurement, Evaluation and Statistical Analysis (MESA) Program. The program represented general methods. After finishing there, I taught program evaluation among other areas. Then, a position at Stanford University became available. The job description called for psychometrics, so that’s what I did! So, as with so many other folks, I followed opportunities as they emerged and fell into this area.

If you weren’t doing this what would you do?

Ed: I really don’t know. I guess I would be doing something in philosophy or more likely, computer programming.

What advice would offer a graduate student who is thinking about psychometrics?

Ed: I have four pieces of advice for graduate students:

1. Consider getting a masters’ degree in statistics.
2. Take some psychology courses, since the work that we do intersects psychology and statistics.
3. Look for opportunities offering relevant experiences.

4. In keeping with #3 above, participate in summer internships. There are so many nowadays at ACT, College Board, CTB McGraw-Hill, ETS, HumRRO, Measured Progress, RAND, among others. NCME has a list of these and I encourage graduate students to seek them out: www.ncme.org/careers/internships.html

When not teaching or researching, what do you do or like doing?

Ed: Well, I like hiking, cooking, crossword puzzles and mathematics puzzles.

What would you say has been one of the biggest innovations in psychometrics in the last decade or two?

Ed: There are two innovations that I would like to mention:

1. Bob Mislevy's Evidence Centered Design (ECD) framework. I think this expands our thinking in test validity and psychometrics. ECD represented a maturation of the field, offering an overarching framework that embraces a range of testing purposes and philosophies.
2. The clarification of the framework of validity that moves the field beyond the technical aspects to the social impact and consequences. Some milestones in the evolution of our thinking about test validity include Messick's 1989 chapter in the third edition of *Educational Measurement*, the treatment of fairness and validity in the 1999 *Standards* (Standards for Educational and Psychological Testing) up to Michael Kane's chapter on validation in the 2006 edition of *Educational Measurement*. Kane made the concepts represented in Messick's work practical and accessible. With Messick, you kind of never knew when you were done. Kane offers a schematic for scoring, generalization, extrapolation, and on to decision making with some nice examples. There are many others who contributed to advancing these ideas (e.g., Lorrie Shepard).

When you go to conferences, how do you pick what sessions to attend?

Ed: I look at who's presenting, or I go to award or sessions that synthesize information or major ideas. I look for sessions that offer big ideas. This is partly a reflection of where I am in my career, although I encourage everyone to look for these types of session. However, when I was a grad student, I looked for sessions that provided a lot of information, so I can get into the nuts and bolts of the issues I was most interested in at the time.

Who has been a significant influence in your professional life?

Ed: First, Benjamin Bloom and David Wiley were initial influences in my professional life. Both were on my dissertation committee. Next, Darrell Bock was a major influence on me. Finally, Lee Cronbach was an influence when I arrived at Stanford. He retired the year I arrived at Stanford, in 1980. I feel very fortunate to have had these remarkable people as influences in my life.

NCME Annual Meeting
2010 Annual Meeting and Training Sessions
April 29-May 3, 2010
Denver, CO, USA

Our program co-chairs, Bob Henson and John Willse, North Carolina University-Greensboro, have worked hard with our membership to prepare a wonderful program for our annual meeting in Denver, CO. Below please find the highlights of our meeting below. Please also join us in extending our deepest appreciation for all of their efforts! Thank you!!

If you haven't already, please go to the following link to register: http://www.ncme.org/meeting/index.cfm

For details of the training sessions, please go to the NCME web site. Thanks are due to Luz Bay, Measured Progress, for putting together a wonderful selection of workshops! Here's a listing of the titles:

- Quality Control in Test Development, Scoring and Reporting of Test Scores
- Data Visualization Using R
- Skills Diagnosis with Latent Variable Models: Theory and Practice, the Theoretical Component
- Bayesian Analysis of Item Response Models: Theory and Methods
- Beyond NDE: Understanding and Working with National Assessment of Educational Progress (NAEP) Restricted-Use Data
- Test Equating Methods and Practices
- A Practitioner's Introduction to Linking and Equating
- Skills Diagnosis with Latent Variable Models: Theory and Practice, the Practical Component
- Developing and Sustaining a Campus-Wide Commitment to Assessment of Student Learning in Higher Education
- State Assessment for Students with Disabilities: Getting it Right
- Tips for Graduate Students: Advice for Finishing School, Obtaining a Job, and Starting a Career
- A Brief Introduction to IRT Parameter Estimation Techniques
- Assessing New Constructs Using New Measures
- Score Drift: Why District and State Achievement Results Unexpectedly Bounce Up and Down from Year to Year
- Moving from Art to Science: An Item-Writing Course for Test Developers, Researchers and Teachers
- Comprehensive Approaches to Validate Construct Invariance and Test Comparability for Federal and School Accountability Reporting Purposes
- Impacting Learning Through the use of Formative Assessment

Program Highlights

Presidential Address: Bridging the Gaps, Terry Ackerman

Career Award Address: Defining and Controlling Errors of Measurement
Moderator: Michael Kolen
Presenter: Michael Kane
Discussant: Robert Brennan

Committee-Sponsored Symposia:

DIVERSITY ISSUES AND TESTING COMMITTEE
Ensuring Equitable Representation of English Language Learners in NAEP: Reactions to the Technical Advisory Panel Report to NAGB on Uniform National Rules for Including and Accommodating ELLs in NAEP
Organizer/Moderator: Charlene Rivera
Participants: Carlos Martinez, Jo O'Brien, Cornelia Orr, Sharif Shakrani, Deb Sigman, Katherine Viator

NATIONAL ASSOCIATION OF TEST DIRECTORS
Validity Issues for Interim Benchmark Assessment Systems
Organizer/Moderator: Jack Monpas-Huber
Participants: Judith Arter, Marty McCall, Lorrie Shepard
Discussants: Pamela Moss, Catherine Taylor

GRADUATE STUDENT ISSUES COMMITTEE
The Influence and Impact of Technology on Educational Measurement
Organizer: Mary Roberts
Moderator: Kimberly Swygert
Participants: Lisa Harris, Richard Luecht, Kathleen Scalise, Joe Willhoft

Invited Symposia:
Assessment of Learning in the Context of Educational Reform: Experiences from America Latina
Organizer/Moderator: Michael C. Rodriguez
Presenters: Michael Fast, Lorena Meckes, Mario Moreno, Fernando Rubio

Update on the Revisions to the Standards for Educational and Psychological Testing
Organizer: Barbara Plake

Moderator: Michael Kolen
Presenters: Laura Hamilton, Joan Herman, Barbara Plake, Denny Way, Laurie Wise
Discussant: Steve Ferrara

Measurement in Higher Education
Organizer/Moderator: Donna Sundre
Presenters: Peter Ewell, Gary Pike, Richard Shavelson, Donna Sundre, Tom Zane
Discussant: Lorrie Shepard

Common Core Standards and Coordinated State Assessment
Organizer/Moderator: Wayne Camara
Presenters: Wes Bruce, Pascal Forgione, Brian Gong, Suzanne Lane, Robert Linn, John Tanner

Are You Being Served? Operational Difficulties in Serving Real and Perceived Needs of State Assessment Clients
Organizer/Moderator: Luz Bay
Presenters: Luz Bay, Daniel Lewis, Diane Henderson-Montero, Paul Nichols
Discussant: Robert Brennan

View from the Top of the Mountain
Organizer/Moderator: Terry Ackerman
Presenters: Robert Brennan, Ron Hambleton, Robert Linn, William Mehrens, Barbara Plake, Lorrie Shepard, Wendy Yen

An Application of Assessment Engineering to Multidimensional Diagnostic Testing in an Educational Setting
Organizer/Moderator: Richard Luecht
Presenters: Mark Gierl, Jacqueline Leighton, Richard Luecht
Discussants: Steve Ferrara, Kristen Huff

Graduate Student Poster Session
This 13th annual poster session of NCME's Graduate Student Issues Committee provides an opportunity for graduate students to share their work and receive feedback from professionals and their peers.

NCME Fitness Run/Walk
Monday, May 3, 2010
5:40 a.m. - 7:30 a.m.

Organizers: Brian French and Jill van den Heuvel

See old friends and meet new ones while running a 5k or walking a 2.5k course on Denver trails. Commemorative t-shirts will be given to all participants (even if you don't wake up in time to make it!).

AN NCME'ERS GUIDE TO VISITING DENVER
THINGS TO DO IN DENVER, CO

Pete Swerdzewski

Great Places to Eat and Drink

Around the Convention Hotels…
The convention area has a number of national chains and a few local gems. Most are located on Denver's 16th Street Pedestrian Mall, which runs from Colorado's capitol building in Civic Center Park to Denver's Lower Downtown area, known today as LoDo. Not only does the 16th Street Mall have free WiFi, but it also has a complementary shuttle bus that runs up and down the street – just hop on at no charge at any designate stops.

> ***Paramount Café***
> One of Denver's classic downtown gathering places. Great burgers, great beer, and excellent outdoor seating. Paramount is also known for its live music, which often features local bands with a unique Colorado vibe.
> *16th Street Mall across from the Denver Pavilions (corner of 16th and Glenarm).*
> *Open daily for lunch and diner; kitchen open until 1am every day of the week.*

Marlowe's

One of the more well-known restaurants in the downtown area, Marlowe's has been a downtown Denver institution for over 25 years, which by Denver standards is a long time (remember, Colorado is only a little over 130 years old!). Marlowe's is the place to go for nice continental cuisine with a local flair. Although billing itself as a chophouse, Marlowe's also has excellent seafood, pastas, and salads.
16th Street Mall across from the Denver Pavilions (corner of 16th and Glenarm), next to the Paramount Café.
Open daily for lunch and dinner; breakfast served on weekends; excellent happy hour.

Denver Pavilions – Home of Maggiano's Little Italy, the Hard Rock Café, Lucky Strike, and Corner Bakery Café

Although not unique to Denver, the restaurants in the downtown Denver Pavilions are close to the convention hotels and cater to conference-goes. Most welcome large groups and can accommodate time constraints. After dining at one of Denver Pavilions' establishments, consider heading to the top floor of the Pavilions for a memorable—and possibly wild—time at the Coyote Ugly bar (yes, it's just like the movie).
16th Street Mall between Glenarm and Welton.

In LoDo…

A short ride on the free 16th Street Mall shuttle bus and you'll find yourself in the center of Denver's nightlife, dining, and entertainment area. LoDo is known for its live music and expansive indoor/outdoor bars. If the weather is good, grab a Fat Tire brew (a Colorado classic), sit outside, and enjoy views of the Rocky Mountains. Make sure to check out the Wynkoop Brewing Company (18th and Wynkoop, across from Union Station), Denver's oldest and considered by many to be Colorado's premier brewpub.

In Larimer Square…

Located just off the 16th Street Mall between the convention area and LoDo is Denver's Larimer Square, a popular place to eat, drink, shop, and explore. Known for its upscale cuisine, Larimer Square boasts French, Mediterranean, continental, and Western restaurant options. And there's even a Starbucks. Consider heading to Larimer Square for a nice dinner away from the convention craziness.

And the Uniquely Denver Options…

Some restaurants can only exist in Denver. Among them are the Buckhorn Exchange, the restaurants at the Brown Palace Hotel, and The Fort.

The **Buckhorn Exchange** is officially Denver's oldest restaurant, and could certainly hold the recognition of Denver's most adventurous establishment. Located a short light rail ride away from the convention area, the Buckhorn Exchange has a decidedly Coloradan menu that includes rattlesnake (quite unusual), bison (good but gamey), and rocky mountain oysters (don't order them unless you know what they are…trust me on this one).

Located right in the downtown area, the **Brown Palace Hotel** hold some of Denver's finest restaurants, including the upscale Palace Arms and the more subdued Churchill Room and Ship's Tavern. Consider a restaurant at the Brown Palace if you'd like to entertain clients, recruits, or your own taste buds.

For those with access to a car and a desire to get out of the downtown area, venture to **The Fort**, consistently billed as one of the most impressive restaurants in the Denver area. The Fort features upscale Colorado fare, including buffalo, elk, lamb, and trout. Not only is the food at The Fort impressive, but the fort itself it an old-style adobe building that is sure to impress anyone venturing to this unique establishment.

Great Ways to Spend Your Time

Things to Do in Downtown Denver…

Whether you have a free hour or a free day, downtown Denver has an extensive mix of things to do and see—there's something for everyone in Denver.

Stroll up and down the 16th Street Mall. The mall has food, shopping, and entertainment venues that are perfect to explore during a break between sessions or after a long day of presenting research.

Read a book at the Tattered Cover. Denver's most loved bookstore, the Tattered Cover has a huge assortment of books to explore, and numerous nooks in which to explore them. *16th and Wynkoop*.

Take a tour of the Colorado capitol building. Tours of Colorado's impressive capitol building are offered throughout the day and include a look at the rotunda and highlights of Colorado's past and present. Look for the placard on the 13th step of the building's front staircase that marks one mile above sea level, and while outside search for flakes of gold that have fallen off the building's dome – yes, it's real gold! *Located at the end of the 16th Street Mall in Civic Center Park.*

Take in some culture. Denver has one of the country's largest fine arts complexes, the **Denver Center for the Performing Arts**. The DCPA host everything from touring Broadway-style shows to small instrumental and performance pieces; be sure to check out their Web site for what's playing *(located next to the Denver Convention Center at 14th and Curtis).* The **Denver Art Museum** displays local and internationally recognized works of art in its impressive new building designed by Daniel Libeskind *(Located across from Civic Center Park on 13th Street)*. For live music, head to LoDo, Larimer Square, or upscale Cherry Creek North (*a few minutes by taxi south of the convention area*). Or, for those with access to transportation (a car or a cab) and a sense of Colorado adventure, head to the **Grizzly Rose**, one of the nation's premier country music dance halls. Although cowboy boots and a hat are not required, they are certainly the norm at this well-known spot for unforgettable two-stepping *(Located approximately 15 minutes north of the convention area by car).*

Explore Denver's less-traveled locales. Check out the fish at the Downtown Aquarium. Learn about the history of Denver through the eyes of the "Unsinkable" Molly Brown (of Titanic fame) at the **Molly Brown House**. Learn something new at the **Denver Museum of Nature and Science**. Shot 'til you drop at **Cherry Creek Shopping Center** and **Cherry Creek North**.

Things to Do Outside of Downtown Denver…

Although Denver has lots to do, if you have a car you can explore the many sites that make Colorado unique and interesting.

Even if there's not a concert going on, **Red Rocks Amphitheater** in Morrison, Colorado is a great place to visit. Red Rocks has hosted performers ranging from Dave Matthews to John Tesh in its naturally occurring red sandstone venue. Visitors are free to explore the amphitheater and surrounding area from the early morning to late at night. *Located 25 minutes by car from Denver*.

Home to the University of Colorado and the Flatirons mountains of Chautauqua Park, **Boulder, Colorado** is renowned for being one of the most livable cities in America. Explore the shops and restaurants of the Pearl Street Mall, take a tour of the Celestial Seasonings Tea Company, or drive a few miles down the road to the Coors Brewing Company in neighboring Golden, Colorado. *Located 45 minutes by car from Denver.*

For the adventurous type, **head up to the high country** and visit famous mountain towns including Georgetown (*1 hour by car*), Frisco/Dillon (*1 hour 30 minutes by car*), Breckenridge (*1 hour 45 minutes by car*), *Glenwood Springs* (*3 hours by car*), or Aspen (*4 hours by car*). These mountain towns all have unique Colorado mining pasts and make great day or weekend trips. For something a little closer to downtown Denver, consider heading to the gambling town of Central City, Colorado, which recently began offering a more extensive selection of slots and table games in its historic pioneer-themed casinos (*1 hour by car*).

Send articles or information for this newsletter to:

Thanos Patelis
The College Board
45 Columbus Avenue
New York, NY 10023

Phone: 212.649.8435
Fax: 212.649.8427
e-mail: tpatelis@collegeboard.org

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.