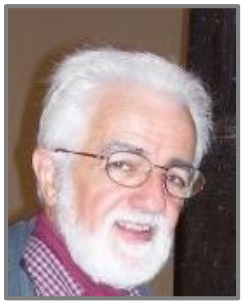
NCME national council on measurement in education

VOLUME 22, NUMBER 4 NEWSLETTER December 2014

FROM THE PRESIDENT

Laurie Wise, HumRRO



The holiday season, from (American) Thanksgiving through New Year's Day provides an important opportunity for us to reflect on all of the things for which we are thankful. I'm hoping that, like me, you will include NCME in this list, both for the useful things we accomplish together and the help we receive from others in advancing our own careers and contributions.

Increasingly, we are asking test users to be explicit in their goals in administering a test and to develop a "theory of action" as to how particular uses of tests will lead to advancement of these goals. In a somewhat analogous way, we might reflect on what we hope to get out of our membership in NCME and on specific chains of actions and outcomes that will lead us to these expected benefits. In hopes of stimulating your thoughts on this topic and in preparation for another NCME Board strategic planning discussion, here are a few of my own thoughts.

I have three general goals in belonging to NCME. First, I hope to continue to learn from others so that I can be more effective in developing tests and in advising others on their use. Second, I hope to help others in working to improve educational measurement and to carry on work in this field long after I retire. Finally, I would like to contribute to evidence (data) based policies that improve education and learning for all children leading to our enhanced ability to solve increasingly complex physical and societal problems.

The theory of action for the first goal is fairly simple. I go to annual meeting sessions to hear what others are working on and read our (and other) journals. A lot of what I know about our field (post graduate school) has resulted from these actions.

My membership dues and contribution to the NCME Mission Fund go to sustain the organization so that NCME can continue the opportunities for others provided by NCME activities. Service on NCME committees is another way to sustain and improve NCME's ability to help others develop knowledge and skill in our areas of work. Annual meeting presentations and journal articles are yet another way of helping others develop professionally. These are the activities that my theory of action leads me to believe will contribute to my second goal.

I think the theory of action for the third goal is related to why educational measurement is important in the first place. Creating tools to diagnose individual student needs as well as to use in evaluating the effectiveness of different instructional strategies is a very necessary step to evidence-based improvements in instruction and student learning. The alternative is an endless cycle of educational fads that may not ever lead to significant advancements. Of course, many of us just enjoy the mathematical puzzles inherent in measurement challenges, from automated item generation through form assembly, scoring, and reporting.

Those are some of my thoughts about the importance of our work. What are yours? If you agree that what we do contributes significantly to educating all to provide them with opportunities and to allow them to help address our many societal problems, please be sure to renew your membership and plan to attend our Annual Meeting in Chicago this April. Also consider volunteering for committee work and, as your circumstances permit, contributing to our Mission Fund.

In the meantime, I hope everyone's holidays were merry and bright!

FROM THE EDITOR – THANKS!!!

Susan Davis-Becker, Alpine Testing Solutions

I hope everyone is having a great time starting off the NEW YEAR!!!! In this issue we have wonderful content from our NCME community. NCME President Laurie Wise kicks off the issue with some thoughts on his goals for our organization. Dianne Talley wraps up her year as graduate student columnist by sharing some thoughts and insights on conducting replication work. Our spotlight member for this issue is Heather Buzick from ETS. We have several special content features in this issue including a Legal Corner from S.E. Phillips, several perspectives on the practical side of validity, and thoughts on the value of standard setting activities – all great reads for the winter break! There are also a number of announcements related to activities going on at NCME including committees and events related to the annual meeting.

With this issue I am completing my term as NCME Newsletter Editor. It has been a great three years getting to serve in this role! I want to express my gratitude to my advisory committee who has helped me identify ideas and authors for content. I am also very grateful to Alpine Testing Solutions for allowing me to dedicate time to this role and Jennifer Paine, our managing editor, who helped me review and prepare each issue of the Newsletter. Finally, I want to sincerely thank all NCME members who have contributed content to the Newsletter over the past few years – the Newsletter is a success and value to the organization because all of you were willing to volunteer your time and share your thoughts. THANK YOU!!

With that, I am very excited to turn over the editorship to Heather Buzick (see Spotlight, this issue) – Welcome Heather!

GRADUATE STUDENT CORNER: WRITING REPLICABLE METHODS

Diane Talley, University of North Carolina, Chapel Hill

A fellow graduate student and I recently attempted to replicate a study as part of a course project. What started as an attempt to understand the model tested in the study, turned out to be a lesson in how to write a methods section of a paper or publication such that the study described could be replicated. Without doubt the readers of this column have taken research courses and learned the importance of writing a clear and sufficiently detailed methods description. However, even the best effort in documenting methodologies may be insufficient for study replication. The aim here is to emphasize the relevance of replicability and posit ways in which the authors of research studies may improve it.

The methods section of a research study provides the key indicators used to assess the validity of the research (Gall et al., 2003, APA, 2009). The sampling methods, chosen instruments, and statistical models used for analysis determine how well research questions are addressed and whether the results are generalizable. The conclusions of the study should be supported by the methods used. For methods to carry this heavy burden, they must be clearly and thoroughly presented.

When a study significantly contributes to the body of literature, replicating a study, conducted either by the original researchers or by external researchers, can be very valuable. In the former case, researchers have the advantage of having been through the process and may not notice an existing lack of clarity or specificity in the methods section of the original research. In the latter case, replicating someone else's research can be a great challenge, particularly for a complex study.

How, then, can a complete and clearly worded documentation of methodology be achieved, while not confusing the readers with superfluous information? There are, not surprisingly, many thoughts on this. The following are a few suggestions:

Review exemplars. The more research we read from high-quality, peer reviewed journals, the more we learn about how the experts in our field manage this task. It also increases awareness of the standards required by industry publications, which presumably encourage (if not require) replicability.

Go forth and replicate**.** The obvious way to verify the thoroughness of a study's documented methodologies is to actually try to replicate that study. Using someone else's research methods to conduct a study can provide invaluable insight into how to adequately document your own research. My colleague and I found that even with what appeared to be a thorough methods description in the study we attempted, there ultimately were important gaps in the information. We found ourselves wondering how the researcher had conducted some parts of the study and what software was used.

Review your work thoroughly. This may seem self-evident, but particularly when time is tight, it's easy to forget to give your methods section a thorough review. As you review your methods section, consider whether you could replicate the study with only the information you provide. Are there missing steps?

External review. Ask a fellow graduate student or mentor to review your work. Request that they specifically consider whether they could replicate your work based on what you have written. If they're indulgent, have them repeat back to you the process for conducting the study, looking for misconceptions and missing important steps.

Remember the software. It is essential to exhaustively detail the names and versions of all software used in the study. As I have found, results may vary considerably using different software or even versions of the same software. If a study is to be truly replicated, the same software is advisable. If a software package is obscure or even somewhat unknown, provide instructions on how and where other researchers may obtain it.

Implications for work in Psychometrics

Not limited to research methods, replicability is also critical in applied psychometrics. There are occasions when a research body, such as a standard setting or an equating study, may need to be replicated to support the validity of its conclusions. This replication may be conducted by someone other than the original psychometrician, or perhaps even by a different testing organization, making thorough documentation of the process, statistical methods, and tools used in the original study imperative. Without such documentation, validity of test scores based on this research can be called into question or discounted. The *Standards for Educational and Psychological Testing* (APA, AERA, & NCME, 2014) provide clear guidance in this regard.

Final thoughts

Research is conducted to increase combined knowledge and solve the practical problems of our field. Clearly articulated, replicable methods can make a strong contribution toward this goal. It is, therefore, valuable to attempt different approaches to improving our writing of research methodology and consider how we may use this skill in our future careers as researchers and psychometricians.

References

American Psychological Society (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Gall, M. D., Gall, G. J., & Borg, W. R. (2003). *Educational research: An introduction*. Boston, MA: Allyn and Bacon.

SPOTLIGHT ON THE PEOPLE WHO MAKE OUR ORGANIZATION GREAT: HEATHER BUZICK, ETS

How did you get into the field?

My first academic interest was in mathematics, in primary and in secondary school, and then I majored in math at Bucknell University. Since mathematics had always been a constant in my life, I actively sought a second major during my undergraduate years. Art history and computer science eventually gave way to economics and in the end I wrote an Honors thesis not in math, but in economics, on merit pay for teachers. After college, I completed a MA in economics, finding interest in human capital models and production functions. The accumulation of these experiences prefigured my career in psychometrics, a space which I began to more fully inhabit during my years at the University of Maryland.

If you weren't in this field, what would you do?

Many career paths have offered their allurements, but I would submit my application for architect by day and drummer by night.

What advice would you have for graduate students who want to get into this field?

I think that adaptability is an important quality in our field and that having both technical and communication skills together fosters this trait. Acquiring a strong set of foundational technical knowledge and skills, which includes research design, statistical modeling, and computer programming, is necessary to be productive in a field that is both evolving and diverse. The ability to simply and accurately communicate about the work we do is essential; because our field is made up of individuals from many backgrounds, this helps to be able to share ideas among researchers, practitioners, and policy makers.

What do you like to do for fun outside of work?

I have two young sons so I spend most of my time with them. I also like visiting with family and friends who live nearby. In the summer, going to the beach or the pool is my favorite activity. If I had unlimited time, I would play golf, spend a month on the Outer Banks, and take trips abroad with my husband.

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

I began graduate school in psychometrics only a decade ago, so I can't speak to changes beyond that on a personal level. Advances in technology have had a widespread impact on our field, across the areas of assessment, measurement, scoring, and statistical modeling. For example, we can now train raters over the network, ping teacher's cell phones to remind them to log their activity, create state longitudinal data systems, estimate complex models with Bayesian approaches, tailor assessments and accommodations to individual needs, and use open-source and proprietary software to estimate more complex models.

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

Typically I look for topics related to major projects that I'm working on at the time. I like to see how different people are pondering the same things I'm thinking about.

Who has been a significant influence in your professional life?

I have been fortunate to encounter creative, caring role models at every turn of my career. My thesis advisor at Bucknell, Catherine O'Connor, my dissertation advisor at the University of Maryland, Greg Hancock, and Cara Laitusis, my colleague, mentor, and manager at ETS—all have inspired me to think about argumentation and solving problems in new ways.

COMMENTARY FROM OUR MEMBERSHIP ON STANDARD SETTING

Thanos Patelis, Center for Assessment

Standard setting is a fundamental component of assessment with the results affecting examinees directly and serving as topics for many discussions, arguments, and policies.

As many educational assessments are being revised and new ones developed, standards are being set. In an effort to continue our discussion within the measurement community and with hopes of informing test users and policy makers, we have solicited comments from our membership on standard setting. We hope this commentary stimulates ongoing conversations and inspires us in both our research and practice.

Empirical Data in Standard Setting

Wayne Camara, ACT

The Smarter-Balanced Assessment Consortium (SBAC) recently approved achievement level cut-scores and projections of student performance across four performance levels. SBAC reportedly used data from ACT assessments and NAEP to inform decisions (Gewertz, 2014; SBAC, 2014).

A number of issues have been raised about the process and results. Most of the issues are not surprising and not of concern. First, between 32%-43% of students were estimated to be at levels 3-4 on assessments based on field trial data. Such results may show far fewer students are proficient or college ready than results based solely on judgmental processes conducted in states, but are in line with results from national assessments (ACT and NAEP) and state efforts using empirical data in setting cut scores. Second, there are legitimate concerns that results were based on field trial data. It is likely that students and educators were less motivated to do their best in a field trial. We know that when any new assessment program is introduced initial results

often under estimate subsequent student performance. Using field trial data rather than waiting for the first year of operational data is not likely to make a significant difference, and the initial standards should be revisited in 2018 when the first cohort of students will have completed freshmen courses, if not earlier using a common student design with external data (Camara & Quenemoen, 2012). Finally, there has been criticism that using achievement levels to report proficiency is misleading and results in a significant loss of information. Amen. When we translate a continuous scale score into four achievement levels or combine multiple indicators of school quality into an A-F grading system we have purposely obfuscated the data and context that goes into teaching and learning. Less information may be desired by policy makers, but as measurement professionals we should be making this argument. However, in all fairness, this is a criticism of accountability requirements which seem to be designed to simplify complex issues, rather than a criticism of the consortium.

In my view, such approaches which include empirical data in content-based standard setting are a vast improvement over previous efforts where states set cut scores based solely on judgments of educators and the only empirical evidence provided panelists was information on consistency of ratings and impact on the students classified at each level. While there has been extensive research conducted on standard setting in the past (e.g., Cizek, 2012), we have continued to see judgmental processes result in vastly different outcomes (and impact) across states and often present a different picture than provided from results of national tests such as NAEP and the ACT. I have assisted or conducted standard setting processes in five states and three national testing programs using empirical data; each time the data have been incorporated in slightly different manners. But in each case, final results were more closely related to results from predictive validity studies than content-based, judgmental processes. There are a variety of ways to incorporate empirical data in standard setting, but when the outcome of interest is college success, content-based judgments should have a narrower role or perhaps no role at all (Camara, 2014). The ACT college readiness benchmarks have been established based on criterion-related evidence between the relationship of ACT scores and freshmen grades. One can legitimately quibble with whether a 50% probability of B is the appropriate threshold, but when a cut score is based on a large nationally representative sample of students and their subsequent academic performance across a large number of 2 and 4 year colleges, the impact of judgment in the definition and outcome is much narrower than traditional content based standard setting. This summer, ACT established four performance levels and benchmarks for ACT Aspire in grades 3-10 from its first operational administration. The cut scores will be revisited and adjusted if needed next spring and as additional longitudinal data become available across grades connecting ACT Aspire with performance on the ACT and college. In 2015, the ACT will be administered as a state assessment to virtually all 11th graders in 19-20 states. Results from these states will comprise a quasi-national reference population and allow state-to-state comparisons and longitudinal trends to track college readiness, college success and to establish empirically based standards and predicted paths for college readiness in lower grades (Allen, 2014). Such efforts which are based on predictive evidence have been used for employment tests for decades and are more appropriate and persuasive for assessments of college readiness than processes which privilege content based judgments alone.

References

Allen, J. (2014). *Development of predicted paths for ACT Aspire score reports*. Working Paper-2014-06. ACT: Iowa City.

Camara, W. J. (2014). *Employing empirical data in judgmental standard setting processes*. Paper presented at the Annual Meeting of the Society for Industrial and Organizational Psychology, Honolulu, HI.

Camara, W. J., & Quenemoen, R. (2012). *Defining and Measuring College and Career Readiness and Informing the Development of Performance Level Descriptors (PLDs).* Commissioned white paper for PARCC. Available at http://www.parcconline.org/sites/parcc/files/PARCC%20CCR%20paper%20v14%201-8-12.pdf

Cizek, G. (2012) *Setting performance standards: Foundations, methods and innovations (2nd ed.).* New York, NY: Routledge.

Gewertz, C. (December 3, 2014). Consortium sets high bar for its common-core tests. *Education Week*, 34 (13).

Smarter Balanced Assessment Consortium (2014). *Smarter Balanced states approve achievement level recommendations*. Retrieved from http://www.smarterbalanced.org/news/smarter-balanced-states-approve-achievement-level-recommendations/

Established Best Practices

Kurt F. Geisinger, Buros Center for Testing, University of Nebraska-Lincoln

Science is defined by the *Random House College Dictionary* as "a branch of knowledge or study dealing with a body of facts or truths systematically arranged and showing the operation of general laws." Using such a definition, the study of the standard setting on educational tests is a science or at least a budding science. Cizek's (2007, 2012) various books lay testament that we have studied standard setting systematically and are cognizant of numerous considerations affecting the various standard setting methods. We have established best practices, but we must reconsider such processes in light of our understandings of the nature of judgment.

The Smarter Balanced Assessment Consortium recently conducted a preliminary standard setting on their grade 3-8 field test results. It appears that a variant of the Bookmark procedure was used to set the three cut scores. Almost 500 teachers, school leaders, higher education faculty members, parents, business, and community leaders participated in this process, the majority of whom were educators. The system was designed to help assure fairness to a variety of student groups (e.g., those with disabilities). Students were to be grouped into one of four ordinal groups on the basis of their test results: novice, developing, proficient, and advanced. Based on these results, the preponderance of students in the varying states would not meet the proficient categorization or higher, a result that presumably reflects demands for increased standards of academic performance.

Our profession has established best practices, and using the Bookmark method with well-trained judges would appear to be one of those approaches. However, we do need to be aware of the following assumptions.

(1) That judges can conceptualize minimally competent students (or students who approximate each of the four score categories).
(2) That judges know how the students at various levels of the score continuum would score on the examination.

These assumptions are "shaky" at best. The evidence for the ability of even well-trained educators being able to make these judgments is weak at best (Impara & Plake, 1998; Hambleton & Jirka, 2006; Lorge & Diamond, 1953). Discussion among judges tends to make their judgments more consistent and to provide the judges more confidence in their decision; it does not necessarily make them more valid. Looking at item statistics is an activity that few other than psychometricans engage with any regularity. I would feel much more confident in the results if we used either a contrasted groups or a borderline examinees approach. Teachers commonly consider the specific achievement levels of the students they teach. My approach would begin with teachers being trained on the meaning of the score levels. Next, teachers would pre-identify students proficiency level if they believe them certain, without access to their test scores. Using the test scores of these exemplar students, the cut scores would be determined. These judgments are those that I believe teachers relatively more competent to make and I would have more faith in such cut score results. That being said, Smarter Balanced nevertheless followed best practices. Commonly accepted practices, however, are not always the most valid practices and I fear that the foundations on which our standard-setting houses stand are often built on sand.

References

Cizek, G. J. (Ed.) (2012*). Setting performance standards: Foundations, methods, and innovations (2nd ed.).* New York, NY: Routledge.

Cizek, G. J., & Bunch, M. (2007). *Standard setting: A practitioner's guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Hambleton, R. K., & Jirka, S. J. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. M. Downing & T. M. Haladyna (Eds.). *Handbook of test development* (pp. 399-420). Mahwah, NJ: Erlbaum.

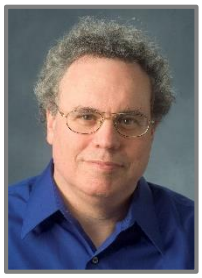
Impara, J. C., & Plake, B. S. (1998) Teachers' ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69-82.

Lorge, I., & Diamond, L. K. (1954). The prediction of absolute item difficulty by ranking and estimating techniques. *Educational and Psychological Measurement, 14*, 365-372.

Thorndike, R. L. (1982). Item and score conversion by pooled judgment. In P. Holland & D. Rubin (Eds.), *Test equating* (pp. 309-318). New York, NY: Academic Press.

The Value of Standard Setting: A Personal Meandering

Neal Kingston, University of Kansas



For the first 14 years of my career I avoided any involvement with research or practice on standard setting. I viewed standard setting as arbitrary. Further, I knew that the creation of categories from continuous (or near continuous) data threw away information and that from a statistical point of view this categorization is always a bad thing.

I then became the Kentucky Associate Commissioner of Education responsible for a testing program that was legally required to have multiple performance standards. This led me to think more about the psychological and policy reasons for having such standards.

Despite the statistical concerns about the appropriateness of standard setting, the question remains about whether the categorization of test scores into performance standards is more likely to provide desired educational outcomes than focusing stakeholders solely on test scores. There are two reasons why the use of performance categories might be desirable. First, some stakeholders might be weak quantitatively (unable to understand means, but able to understand counts and percentages). Second, performance categories might serve as goals to motivate students and teachers.

Almost two decades ago Ed Reidy and I hypothesized these two points might be true (Kingston & Reidy, 1997), but I was unaware then and remain unaware now of any empirical research about stakeholders' understanding of state testing results. There is a rich literature about the impact of low numeracy in general (American Institutes for Research, 2006) and in the health care area specifically (Ancker & Kaufman 2007). Regarding the second point, performance categories as goals, the use of levels (in addition to scores) is a common motivator in many massively multiplayer online role playing games. Scores go on forever, while levels present goals to be accomplished.

There has been much research over the last 50 years about how to conduct standard setting, yet almost no research has occurred about why we should use standard setting for presenting state assessment results. Though I believe the use of performance standards is useful, I do not know that to be the case. More research is necessary and I urge the Newsletter readership to move the field forward in this regard.

References

American Institutes for Research (2006). *A Review of the Literature in Adult Numeracy: Research and Conceptual Issues.*

American Institutes for Research, Washington, DC. (Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=5&ved=0CD8QFjAE&url=http%3A%2F%2Fwww2.ed.gov%2Fabout%2Foffices%2Flist%2Fovae%2Fpi%2FAdultEd%2Fnumlitrev.doc&ei=QjigVPH4JMjgggSq44KoCg&usg=AFQjCNFki3GE-2ejmQrdYan2fhBz6EZDjQ&sig2=VYKYihqB280KqbCVB72Yhw 12/28/2014)

Ancker, J. S., & Kaufman, D. (2007). Rethinking health numeracy: A multidisciplinary literature review. *Journal of the American Medical Informatics Association, 14*(6): 713–721. (Retrieved from http://www.ncbi.nlm.nih.gov/pmc/journals/76/ 12/28/2014)

Kingston, N. M., & Reidy, E. (1997). Kentucky's accountability and assessment systems. In J. Millman (Ed.). *Grading teachers, grading schools: Is student achievement a valid evaluation measure*? Thousand Oaks, CA: Corwin Publishers.

Standard Setting is a Tool

Jerry Melican and Deanna Morgan, The College Board

It is frequently necessary to define, as best possible, the minimal levels of knowledge and ability required for licensing or certification or to define proficiency levels in education resulting in cut-scores. There are two widely held tenets about standard setting: there is no "true" cut-score and standard setting is arbitrary based on human judgment. Taken together, these sentences place a humbling responsibility on standard setters. The tenets do not eliminate the usefulness of setting cut-scores for licensure/certification or education but they do imply that any standard setting study be performed in a professional and transparent manner allowing users to understand the strengths and limitations of the interpretations based on the cut-scores. Further, the score reports and documentation must be clear and accurate to avoid misuse, overuse, and misinterpretations.

Standard setting studies are a tool to solicit information, in the form of a recommendation, from representatives of relevant constituencies regarding appropriate locations to establish cut-scores separating categories. In education, for example, standard setting is a tool used to translate the content knowledge and skills information about what students need to know and be able to do into the context of the numerical test score and how much students need to know and be able to do. Since human judgment is required to provide the information and the interpretation of that information, the quality of the information depends on assembling representatives with appropriate knowledge of content and the population of test takers. These panelists must have or develop an understanding of the qualifications required to be classified in one category versus another, receive training on the standard setting method, and indicate that they understand the process. The method must be chosen to meet the needs of the specific test and the materials used for training and collecting panelist input must be carefully prepared. Any statistical information required must be appropriate and accurate.

Central to use of this tool is the identification of the content knowledge and skills that are required for student success in the next grade or as they move into a career, or college. The strong connection and grounding to what students need to know and be able to do along with the involvement of subject matter experts is the real strength of the standard setting study as a tool in the cut score recommendation process. Once again, the documentation must include the delineation of the abilities required at each category to allow users to understand and accept the utility of cut-scores.

The Bookmark method, for example, requires the item characteristics used to develop the Ordered Item Booklet so it is important to describe the sample upon which the parameters were generated. During many decision-making deliberations the frequency distributions and percentages of field test takers in each category are used to evaluate the impact of the recommendations on student performance again requiring accurate description of the sample.

Each type of information mentioned from the selection, qualifications, and training of panelists and decision makers to the interpretation of results must be communicated fully, accurately, with appropriate caveats and limitations, if the tool is to be useful and used. For example, having a cut-score does not mean every person above the cut-score will be successful and every person below will not be successful. The complete explanation of how particular cut-scores may be interpreted, whether for certification/licensing or education, depends on the particular interpretations to be used for the scores but should be explicitly stated as part of the documentation supplied to test takers and other users to avoid confusion and over-interpretation. In education, cut-scores provide a useful starting point for interpretation to then be buttressed with additional information regarding interpretation (e.g., estimates of false positives and negatives) and with other external information such as courses taken and grades.

Research-Based Evidence

M. Christina Schneider and Erika Hall, Center for Assessment

Setting performance standards on large-scale assessments is central to test score interpretation. To support valid user inferences, standard setting procedures must be aligned to the primary purpose of an assessment. From our perspective, recent changes in the use and interpretation of many K–12 assessments necessitates increasing the use of research-based evidence in the standard setting process and changing the role of educator panels.

The Common Core Standards were developed to "prepare students for college, career, and life." As a result, for many assessments aligned to these standards the primary goal of standard setting has shifted from educators identifying the threshold between content-based performance levels to researchers identifying scores that predict "readiness" or a specified "likelihood" of success in post-secondary endeavors. Policymakers understand this, but to maintain face validity, they often convene educator panels to validate recommendations. If cut scores are intended to support empirically-grounded claims, educator participation at this stage of the process may not be necessary. Rather, it may be optimal to conduct the research and proactively communicate the evidence supporting the cut scores. Cut scores, however, should not be the only output from a standard setting. The process needs to be expanded to include supports that help stakeholders better understand how much knowledge and skills students must possess as a result of where cut scores are located on the test scale.

The process should focus educator panels on developing high quality performance level descriptors once the cut scores are known. Such descriptors should define what "readiness" means and describe student achievement in ways that are identifiable in the classroom. Such a process should serve as a foundation for making expectations transparent and help teachers identify resources and supports necessary to move student learning toward policy goals.

Merits of Standard Setting and Cut Scores

Andrew Wiley, Alpine Testing Solutions Inc.

There are very few guarantees in life, but one assurance is that in the world of educational assessment, there will be some controversy regarding every key decision made for a program, particularly when those decisions involve determinations of student achievement. There has been significant discussion regarding the merits of sett
ing achievement levels for educational assessments as well as how the cut scores representing these levels are developed. The state of Vermont recently raised a significant amount of controversy when it released a statement announcing that it did not support the cut scores that the *Smarter Balanced* consortium had developed.

One of the primary criticisms, and the one we will discuss here, is that the process for setting cut scores is arbitrary and cannot be trusted. While there certainly can be valid criticisms directed at educational assessments and how policymakers implement cut scores, the critique that cut scores are set arbitrarily does not give enough consideration to the rigorous process traditionally followed in setting cut scores. While no one who works in the field of educational measurement would deny that the process of standard setting is a mixture of both art and science, to dismiss the process as arbitrary does not sufficiently appreciate the science behind the process.

While many standard setting methodologies rely on the judgment of subject matter expert panelists, the cut scores are not recommended to policymakers in a vacuum. These panelists engage in extensive review of the test itself and of any relevant data before providing any cut score estimates, and have multiple opportunities for discussion on the nature of the assessment and its intended uses. As outlined by Kane (1994, 2001), all standard setting activities should be able to demonstrate that they have followed a rigorous procedure that supports the procedural, internal, and external validity of the process used to establish the cut scores. Through the procedural validity, evidence is presented that shows that panelists had the opportunity to review and discuss the test in question, and to also discuss the meaning and implications of each cut point being established. Through the internal validity, evidence is presented that demonstrates that panelist reached a reasonable consensus on the appropriate cut scores for each test. Finally, with external validity, evidence is presented that demonstrates whether or not the cut scores are consistent with other similar uses and provides results that are reasonably in line with the external data available.

In the end, just as with many components of our entire education system, there are judgmental components in determining cut scores by both content experts and policymakers. However, these judgmental processes are enhanced through systematic and rigorous procedures that lead to cut scores that reflect the best judgments of a sample of representative panelists in combination with contextual factors considered by policymakers.

References

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*, 425–461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.

THE PRACTICAL SIDE OF VALIDITY

As noticed during the 2014 AERA/NCME annual meeting, two of the best attended sessions were on the subject of validity. As a field, we continue to theorize and debate how validity should be conceptualized, documented, and evaluated. Although we may disagree on the best psychometric approach to considering validity, we may all agree that it can be challenging to explain the concept of validity to lay audiences including why it is important, what it means in practical terms, and why it is an ongoing effort as part of the test maintenance process. In an effort to pursue this important topic, I reached out to some colleagues in the field who have worked with a variety of groups and ask how they approach this type of task. I would encourage everyone to reflect on these types of issues and share how they tackle such challenges when working with lay audiences.

Avoiding the Misuse of Scores

Jerry Melican, The College Board

Whether in educational or licensure/certification testing, people generally understand the initial concept of validity in that the scores should be measuring what the test is intended to measure. They therefore understand at a conceptual level the steps that need to be taken at the outset, such as defining the test specifications, writing items to specifications, and performing fairness reviews. The remarkable amount of effort and expertise required to generate a test that provides valid scores is not as readily understood, and it is therefore necessary to recognize and build on the knowledge the client brings to the program.

As noted it seems to me that to effectively explain validity to prospective or current clients it is necessary to capitalize on their current understanding and/or to understand areas of confusion for them. At some point I may use a PowerPoint presentation but early on I find it helpful to have an introductory conversation with the person doing as much or more talking about their experience, needs, and thoughts. A lot of my contributions tend to be more "practical" than theoretical or definitional; that is, this introductory discussion does not include a lecture on validity unless the conversation suggests that is what is called for.

The practical part of my contributions include short descriptions of collecting content specifications, item writing and purposes of reviewing, and relating items back to the specifications. At this point, the initial idea of "validity" is being established on common ground to be used as advanced organizers. It then becomes possible to discuss the less obvious steps required to develop a testing program such as generating items at appropriate levels of cognitive ability, developing preparatory documents and practice tests for the test takers, instituting test protocols and training proctors, providing appropriate levels of score information for the score reports, and continuing statistical analyses.

There is one area that crops up regularly once a testing program has been established which involves uses and misuses of scores. Once a test is available, it is not uncommon that well-meaning individuals conceive of additional uses. For example, a client may want to use the tests that were developed for entry into a field as a test for determining promotions as well. Clients do not understand that there are additional steps necessary to determine whether it is possible to use the scores for the new purpose. With regards to uses of test scores that have not undergone a validity review my favorite example is when Dr. Gregory Anrig of ETS denied use of the National Teachers' Examination (NTE) to Arkansas to evaluate practicing teachers. This was a long time ago, 1983, but stands as a great example of the profession standing up for appropriate test use. Understanding this concept is helpful in developing the Q&A for the testing program which is a major part of the external communication plan.

Another common issue for licensure/certification testing is when users want to compare scores for test takers who have exceeded the passing standard. That is, a client user might want to hire a person with a higher score than one with a lower score even though both have "passed" the test. It is not immediately clear to clients that the test specifications have been generated to provide maximum discrimination in the area of the cut-score with lower levels of discrimination at other areas on the scale.

This latter part of the conversation concerning uses, misuses, and abuses may take the tack of helping the client or the person responsible for public explanations develop a Q&A. In generating the Questions interactively a lot of learning – on both sides – occurs. It is important then as the documentation is developed that the uses, misuses, and possible abuses be recognized and text generated. The client uses this text and training materials to prepare all users for proper uses of the test scores and to preemptively avoid misuses and abuses. It is common to have punitive measures for users who overstep the accepted uses of the test scores.

Author note: I have not been called upon at the College Board to provide explanations of validity to members or clients so the preceding paragraphs were based on previous experience mostly in licensure/certification.

Making It Personal

Donna Sundre, James Madison University

I have found that it is fairly easy to entrap faculty into thinking carefully about validity issues when we make the discussion personal to them. They care deeply about their academic programs, but their real investment is in individual students—their graduates. It's incredibly important for our faculty to understand validity concepts and to be convinced that our methods align with their conceptualizations—largely because there is increasing demand for USE of DATA. Everyone wants them to SHOW how they are using assessment results to improve their programs. Actual evidence of improvement in student growth and development is the new mandate. I do NOT want faculty using assessment results to make inferences about curricular quality or individual students that lack validity for those purposes. It takes time to do this correctly, and we have to give faculty the time to get this right.

When I want to know about the heart and soul of what they want their program to deliver….when I want to know how they hope students will be transformed as a result of their exposure to their academic program, I can ask them to tell me about a certain someone. I assure you that every faculty member can immediately respond to this query with detail and enthusiasm. "Tell me about a few students who you are most proud of and why." I have never asked for this kind of information without hearing about the most important goals and objectives of our academic programs. With a little probing, I will hear glorious details about how these objectives are manifested in behaviors, competencies, values, attitudes, and dispositions. These go well beyond the table of contents of a stack of text books—these are the real objectives that we pursue in all of our graduates. This is what we hope for and strive for, and yes, this is the heart and soul of validity. This is the noble quest that must be honored to conduct assessment that truly recognizes the importance and value of the degrees we award—at every level. This is why so many faculty are committed members of academe. If we can make this come alive by being explicit about what we really strive for, we will be successful. Anything less will lack validity and result in an assessment compliance mode.

These goals and objectives are not easily explicated or measured, but it is certainly worth the time and effort. I know that we are on the right track when faculty are deeply engaged in this work. This should be one of the most intellectually stimulating and important efforts we will engage in. This is what validity is all about.

Determining What Is Useful and Appropriate

Brett P. Foley, Alpine Testing Solutions

I believe that it is especially important for program leaders and policymakers to have a basic understanding of validity, because it is they (along with salespeople) who are often responsible for extending the uses of otherwise good tests into unjustifiable realms. However, it's difficult to fault these stakeholders for having some confusion about the concept of validity, given that some of the greatest scholars in our field still debate its definition.

I consider it a small, personal victory when I can convince a stakeholder that validity is a property of the uses and interpretations of test results, rather an inherent property of the test itself. One challenge is dissuading the stakeholder from the notion that "valid" is synonymous with "high quality." Doubtless, we all want to produce high quality tests. However, test quality is a necessary but insufficient condition for determining whether or not a use/interpretation is valid. We come closer to the mark when we can convince stakeholders that a better (though admittedly simplistic) operational definition of "valid" might be "useful and appropriate."

Given such a simplification, the concept can be reinforced through analogies: In the same way an exquisitely beautiful European sports car, while undoubtedly of high quality, would be neither useful nor appropriate for helping your roommate move his couch, an exquisitely beautiful (European?) math item would likely be neither useful nor appropriate for measuring a student's reading skill. The "useful and appropriate" pseudo-definition also helps to convey the need for supporting evidence for each intended use and interpretation. For example, it may seem useful (or, at least, convenient) to employ a test of a student's mathematical achievement to evaluate a teacher's math teaching skills. However, what is less clear is whether or not such a use is appropriate. In order to have confidence in the appropriateness of this use, we need some strong supporting evidence. Thereby, the "useful and appropriate" interpretation of validity helps to get stakeholders thinking in a way more in line with the goals of the psychometric community.

In summary, while the debate as to the "true" definition of validity may continue in perpetuity (or at least as long as measurement scholars still need tenure), we can help improve stakeholders understanding of validity by gently moving them towards the interpretation of validity as the useful and appropriate uses and interpretations of test results.

LEGAL CORNER: RECENT DEVELOPMENTS IN DISABILITY LAW[1]

S.E. Phillips, Consultant

One relatively recent and another very recent development in disability law have changed the landscape of professional admissions testing and may also put pressure on professional licensure testing programs to adopt similar policies. These developments include the ADA Amendments Act (2008) and a recent settlement agreement in a California case challenging the policy of the test provider to deny certain nonstandard test administrations and to annotate scores obtained with extra time on the Law School Admissions Test (LSAT).

Background on the ADA Definition of Disability

The Americans with Disabilities Act (ADA, 1990) extended the Section 504 of the Rehabilitation Act (1973) provisions prohibiting discrimination against persons with disabilities in programs receiving federal funding to all public and private entities. The ADA and its associated regulations defined disability as an impairment that substantially limits a major life activity. Major life activities included functions such as caring for oneself, performing manual tasks, walking, seeing, hearing, speaking, breathing, learning, and working. An activity was substantially limited when the conditions, manner or duration with which a person could perform the activity were restricted in comparison to most people. In *Sutton v. United Airlines* (2000), the U.S. Supreme Court found that applicants for the position of global airline pilot whose vision was 20/20 with corrective lenses but who were unable to meet the airline's requirement of 20/100 uncorrected vision were not disabled under the ADA. The Court held that disability determinations should be made based on a person's functioning with correction and that the applicants were not disabled because they were not substantially limited in any major life activity when their vision was fully corrected. On the contrary, the Court found they only were unable to meet the requirements of one specific job. Moreover, the Court held that agency guidelines specifying the evaluation of disability based on the individual's hypothetical uncorrected state were impermissible because the ADA had not given the implementing agencies the authority to modify its general definitions. Two years later, in *Toyota Motor Manufacturing v. Williams* (2002), the Supreme Court adopted a strict standard for deciding when an impairment substantially limited a major life activity. In that case, a person's carpal tunnel syndrome was found to interfere only minimally with performing manual tasks of central importance to the daily lives of most people. The Court further explained that a person claiming a disability based on a medical diagnosis must also provide evidence of substantial limitation(s) specific to that individual.

ADA Amendments Act (2008)

The ADA Amendments Act (ADAAA) was passed in response to a report by the National Council on Disability (NCD, 2004), a federal agency charged with collecting and analyzing information about the effectiveness of the ADA. The report was critical of a series of U.S. Supreme Court decisions (including the *Sutton* and *Toyota* cases discussed above), arguing that the definition of disability and requirements for coverage had been inappropriately narrowed contrary to the intent of Congress and the objectives of the ADA.

The purposes of the ADAAA as enacted by Congress were to reinstate a broad scope of available ADA protection by

- Rejecting the holdings of *Sutton* and its companion cases evaluating disability as corrected by mitigating measures;
- Reinstating the broad definition of disability applied by the Court in *Sch. Bd. of Nassau County v. Arline* (1987) finding a history of hospitalizations for infectious tuberculosis sufficient to establish a record of disability under Section 504;
- Rejecting the holding in *Toyota* that the terms *substantially* and *major* in the ADA definition of disability "need to be interpreted strictly to create a demanding standard for qualifying as disabled" as an "inappropriately high level of limitation" for ADA coverage (42 U.S.C.A. § 12102(b)(5)).
- Stating the intent of Congress that determinations of whether a person's impairment qualifies as a disability covered by the ADA "not demand extensive analysis" and the primary focus of litigation "should be whether entities covered under the ADA have complied with their obligations" (*id*).

Substantive changes in the ADA codified by the ADAAA included

[1] Adapted from Phillips, S.E. (in press). Legal issues for credentialing examination programs. In S. Davis-Becker & C. Buckendahl (Eds.), *Testing in the Professions: Credentialing Policies and Practice.*

1. Replacing references to discrimination "against a qualified individual with a disability" with discrimination "*on the basis of disability;*"
2. Expanding the definition of disability to
    a. Add eating, sleeping, standing, lifting, bending, reading, concentrating, thinking, communicating, and operation of bodily functions (immune system, normal cell growth, digestive, bowel, bladder, neurological, brain, respiratory, circulatory, endocrine and reproductive) to the list of major life activities;
    b. Define *regarded as having a disability* to include persons subjected to ADA prohibited actions because of an actual or perceived impairment irrespective of actual limitation but not including transitory (duration ≤ 6 mo.) and minor impairments;
    c. Favor broad coverage of individuals;
    d. Include impairments that are episodic or in remission if, when active, a major life activity is substantially limited;
    e. Make disability determinations based on an individual's uncorrected state disregarding the effects of mitigating measures (medications, magnification, prosthetics, hearing devices, mobility devices, assistive technology, reasonable accommodations, auxiliary aids or services, and learned behavioral or adaptive modifications) EXCEPT ordinary glasses or contact lenses that fully correct visual acuity;
3. Disallowing selection criteria based on uncorrected vision unless such criteria are job-related and consistent with business necessity (valid for the intended interpretations and uses);
4. Requiring *reasonable accommodations* for individuals with actual impairments or records of actual impairments but not for individuals qualifying for ADA coverage solely based on being *regarded as* having an impairment; and
5. Specifying that the DOJ and EEOC have the authority to issue regulations that implement the ADAAA definitions of disability and rules of construction.

The ADA Amendments Act also amended Section 504 to incorporate the ADAAA definitions of disability and specified an effective date for the ADAAA of January 1, 2009. Because Congress expressed an intent for the courts to apply a lenient standard to the determination of disability in order to shift the focus from qualification of the impairment to whether the appropriate accommodations were provided, more examinees are now protected under the ADA. Virtually any recognized medical condition evaluated without correction appears to qualify for a testing accommodation if it limits one of the many listed major life activities and is linked to test taking skills. However, because this expanded qualification of a greater variety of impairments as disabilities creates an incentive for struggling examinees to claim a disability to obtain assistance (e.g., extra time) they believe will raise their scores, it continues to be important for examinees to provide sufficient documentation to ensure that diagnoses are provided by appropriate professionals and are consistent with credible corroborating evidence. In addition, documentation needs must be balanced against provisions in the ADAAA Regulations (2010) stating that documentation requirements must be reasonable, limited to the need for the accommodation or auxiliary aid and elicit information about past accommodations received in similar testing situations so such precedents can be given considerable weight. But the ADAAA did not change the need for a causal connection between the impairment and the specific testing activity that is substantially limited, the duty of the examinee with a disability to explain how the requested adaptation(s) will address the specific limitations the impairment causes, the expectation that a testing adaptation will be provided only when the tested construct has not been fundamentally altered, or the requirement for individual determinations of appropriateness and effectiveness.

It has taken time for cases arising under the 2008 ADAAA to create written decisions. As of this writing, only one credentialing case has applied it (*Jenkins v. NBME,* 2009). That case was remanded back to the trial court for reconsideration of a 2002 decision denying extra time to a medical licensure examinee who was judged not to qualify as disabled under the strict *Toyota* standard. Although he was a slow reader, the court found the examinee had failed to show that his reading impairment prevented him from performing any major life activity, such as reading newspapers, as well as most people. Since the student was seeking extra time for future exams to be administered after passage of the ADAAA, the appeals court ruled he was entitled to have his case reconsidered under the new ADAAA disability qualification standards. Beyond this case, the terms of a recent settlement of a law school admissions test (LSAT) case under the ADAAA may impact other professional admissions and credentialing programs.

The LSAT Settlement (2014)

In a class action lawsuit, a California consumer protection agency alleged that LSAT scores obtained with extra time were being annotated in violation of the amended ADA and California's more expansive civil rights act (*DFEH v. LSAC,* 2012). The agency also cited a Department of Justice regulation requiring examinations to *best ensure* measurement of the individual's achievement rather than reflecting the individual's impairment. An earlier California case involving annotation of scores with

extra time on another professional program admissions test had held that testing entities have an affirmative duty to provide test administrations for examinees with disabilities that minimize the effects of the impairment while enabling the fullest possible demonstration of the tested content (*Breimhorst v. ETS*, 2000). This decision led to a settlement with ETS and decisions by the College Board and ACT to discontinue annotating scores obtained with extra time on their professional and college admissions tests.

In the LSAT case, the agency was seeking to extend this ruling to the admissions test for law school. The court concurred, ruling that the test provider had not met its burden of demonstrating that its test *best insured* that the abilities of examinees with and without disabilities were measured equally. The court acknowledged that the precise requirements of the *best ensure* standard were not clear but held that it was "more exacting than a reasonableness standard" (p. 869). The court also agreed with the agency that score annotations discouraged examinees from requesting testing accommodations and punished those whose requests were granted. The court declined to dismiss the case and the test provider chose to settle the case out of court to avoid a trial.

The settlement was announced in May 2014 and required the LSAT test provider to pay $7.73 million in civil penalties and damages to compensate approximately 6,000 examinees denied testing accommodations nationwide over the previous five years. The test provider also agreed to permanently end all score annotations for extra time, automatically grant most accommodations received previously by an applicant on a postsecondary admissions test, and to implement the recommendations of an expert panel convened to identify best practices for evaluating testing accommodation requests. Neither the court nor the settlement appeared to have considered whether LSAT scores obtained with and without extra time are actually comparable or whether it is fair to expect a test provider to honor a decision made by another testing entity at an earlier time based on unknown evidence and criteria. Further, it appears that the possibility (and limited evidence) that scores obtained with extra time might over-predict college or professional program success because reading fluency and processing speed are relevant to the ability to handle the reading load and work speed expected at the college level or in the professions was not considered.

Although this settlement may be cited in future cases, it is important to note that a settlement does not have precedential value in court because its terms apply only to the parties who agreed to it. Nonetheless, there seems to be a trend in which testing entities are sued or threatened with lawsuits and subsequently acquiescence to granting extra time to examinees with disabilities while foregoing score annotations that would alert users, such as colleges or professional programs, of the possibility that the scores have different meanings and represent a somewhat different construct than for examinees tested under standard time limits. To the extent scores obtained with and without extra time are not comparable, providing extra time to examinees with disabilities without annotating the scores amounts to affirmative action for persons with disabilities, contrary to the amended ADA which requires only the removal of construct irrelevant barriers.

One of the reasons given for removing score annotations is to avoid identifying an examinee as disabled to programs that might use that information to discriminate against such applicants. However, treating scores obtained with and without extra time as comparable implies that processing speed is construct irrelevant. If so, a fairer solution might be to allow all examinees to choose between taking the test with (1) standard time limits or (2) with extra time and the score annotation "tested with X% extra time." This policy would allow struggling examinees to demonstrate what they know and can do under relaxed time pressures but would not identify an applicant with a score annotation as a person with a disability. The college or program receiving the scores could then make fair comparisons among applicants who took the test under the same conditions (with or without extra time) and could decide individually whether processing speed was or was not important for their particular programs.

Conclusion

Time will tell whether the LSAT settlement terms and the amended ADA definitions will impact credentialing programs or remain an admissions testing issue. The two situations may be viewed differently because the credentialing consequences are dissimilar and potentially more serious. Admitting an applicant with a disability who tested with extra time to a college program where processing speed matters may result in the applicant dropping out, an outcome that wastes the applicant's time and money and occupies a slot that could have been offered to another applicant. It might also put pressure on programs or instructors to change or waive normal course requirements that are particularly difficult for the applicant. Alternatively, the applicant may overcome the disadvantage and successfully complete program requirements. However, if the applicant will be required to pass a credentialing test following completion of the program but will not be permitted the same testing adaptations on the credentialing test that were provided on the admissions test, it might better serve the applicant to be considered for admission based on a test score without the adaptations than to be unsuccessful later on the credentialing test after spending significant time and money on training.

On the other hand, allowing extra testing time and licensing a candidate who lacks the processing speed necessary to function safely and effectively in a profession may put members of the public at risk of unnecessary and irreversible negative outcomes, a result that is unacceptable and may cause the unsuccessful professional to face legal action and/or credential revocation. In addition, it may not be possible to adequately compensate members of the public for their injuries and the professional whose credentials are suspended or revoked may then find it extremely difficult and expensive to change occupations. Moreover, a score annotation would provide no protection because the decision is to either license the candidate or not and the public has no access to the scores. An important consideration in deciding whether processing speed or any other factor is construct relevant for a credentialing test is the support provided by the job analysis information. A job analysis survey in which a substantial majority of surveyed job incumbents report that performance under time pressure and in distracting environments is a frequent and important aspect of their jobs may provide the test provider with sufficient corroboration to refuse certain test adaptation requests. Put more bluntly, would you want to engage a doctor or lawyer who needed double time, a separate room, a reader and extra breaks to pass the required credentialing exam?

References

Americans with Disabilities Act [ADA], 42 U.S.C. § 12101 *et seq*. (1990).

ADA Amendments Act (ADAAA), 42 U.S.C.A. § 12101 *et seq.* (2008).

ADAAA Regulations, 29 C.F.R. § 35.101 *et seq*., § 36.101 *et seq.* (DOJ, 2010).

Breimhorst v. Educational Testing Service [ETS], No. C-99-3387 WHO (N.D. Cal. 2000).

Dept. of Fair Employ. & Hous. (DFEH) v. Law Sch. Admissions Council (LSAC), 941 F.Supp.2d 1159 (N.D. Cal. 2013), Consent Decree (Settlement), Case No. CV 12-1830-EMC (N.D. Cal. 2014).

Jenkins v. NBME, No. 08-5371 (6th Cir. 2009).

National Council on Disability. (2004). *Righting the ADA*. Washington, D.C.: author.

Sch. Bd. of Nassau County v. Arline, 480 U.S. 273 (1987).

Section 504 of the Rehabilitation Act [Section 504], 29 U.S.C. § 701 *et seq.* (1973).

Sutton v. United Air Lines, 527 U.S. 471 (1999).

Toyota Motor Manufacturing v. Williams, 534 U.S. 184 (2002).

ANNUAL MEETING COMMITTEE UPDATE

Terry Ackerman, University of North Carolina – Greensboro

The Annual Meeting Committee continues to meet about once a month via conference call. Currently the committee is working with Leah Knope from The Rees Group (TRG) to create an Annual Meeting Handbook. Leah is the staff archivist at TRG. The Committee has broken down the NCME Annual Meeting into 23 events or tasks that occur throughout the meeting. These include things such as the Board of Directors Meeting, the Fun Run, the Breakfast and Business Meeting, Housing, Graduate Student Poster Session, etc. Leah is helping us complete a standard form for each event. That is, for each event we are noting such things as:

- The timeframe and length of the event
- How the effectiveness or success of the event is evaluated
- What is the past history of the event
- Are there certain procedures/protocols that must be followed
- The roles and tasks of people responsible for the event
- Timeline of preparation for the event (who does what and when)
- Onsite meeting setup checklist
- Any recommended changes or issues from the evaluation that need to be resolved before the next meeting

Note that cost or budget for the event is not addressed in this Handbook, but is reviewed by the Board of Directors as part of the annual meeting planning budget.

Leah prepares a first draft for each event/task. These drafts are put up on Basecamp for Committee members to review. Then, on our monthly call, the committee goes through the information Leah has prepared. We discuss each event and, if necessary, make changes to the document. Once "approved" it becomes part of the handbook. Our goal is to complete the handbook and present it to the Board of Directors at the annual meeting in Chicago. This document, once completed, will be extremely helpful to future presidents and Board members, program chairs, as well as to the NCME management company.

HIGHLIGHTS FOR THE 2015 MEETING

Ye Tong & Jennifer Randall, conference co-chairs
Caroline Wiley, training chair

Invited Speaker: John King

Commissioner John B. King, Jr. was appointed Commissioner of Education and President of the University of the State of New York (USNY) in May 2011. USNY comprises more than 7,000 public and independent elementary and secondary schools; 270 public, independent and proprietary colleges and universities; 7,000 libraries; 900 museums; 25 public broadcasting facilities; 3,000 historical repositories; and 436 proprietary schools.

Dr. King is a strong voice for education reform, and he was a driving force in New York's successful Race to the Top application. A former high school teacher and middle school principal, Dr. King has earned a national reputation for his vision and commitment to education reform. Dr. King earned a B.A from Harvard University, an M.A. from Teachers College, Columbia University, a J.D. from Yale Law School, and an Ed.D. from Teachers College, Columbia University.

Satirical Session: Contemporary Problems in Educational Measurement
Moderator**:** Kevin Sweeney, The College Broad

- Solving 22nd-Century Measurement Problems
  - Ellen L. Ripley, Nostromo Inc.; Robert Neville, U.S. Department of Education; Elroy Jetson, Spacely Space Sprockets; Christopher Pike, NASA
- An NCME Invited Debate: Godzilla vs. Fairtest: The Rematch
  - Anne T. Exam, Fairtest; Dr. Godzilla, University of Tokyo
- Joint Committee on Fair Testing Practices
  - David Williams daughter, Acid Tests, Inc.; Kristen Puff, Regis Philbin Research Fund; Neal Kingdomcum, Yonkers University; Ellen Fortress, misCount, LLC; Ric Elect, University of North Antarctica
- Certifying Psychometric Competence
  - Andrew Wiley, Alpine Testing; KT Han, Council of Cheapskate School Officers
- Detecting and Prosecuting Cheaters on Educational Exams
  - Ellwood U. Cheet and Jake K. Opy, Joliet Correctional Facility; Robert Crook, Bored of Medical Examiners
- Assessing College Readiness: Non-cognitive Factors
  - Gil Andromeda, Even Higher Education Research Consortium; Mary Petunia, Educational Testy Service; Highfive Elephantmat, Professional Procrastination Service; Sparky Torres, PARCC Inc. Lot.

The Importance of Instructional Sensitivity: A Colloquy Among Combatants
Participants:

- Jim Popham, University of California Los Angeles
- Neil Kingston, University of Kansas
- Jon Fremer, Caveon Test Security
- Denny Way, Pearson

The Myth of Equal Measurement Units in Educational Testing
Participants:
- Derek Briggs, University of Colorado Boulder
- Wim Van der Linden, CTB McGraw-Hill

Moderator:
- Terry Ackerman, University of North Carolina Greensboro

NCME-NATD Symposium: Implementing the Common Core Assessments at the District and School Levels: Voices from the Field - Overcoming Challenges, Making it Work

Participants:
- Didi Swartz, Chicago Public Schools (Illinois)(PARCC)
- Melanie Stewart, Milwaukee Public Schools (Wisconsin)(SMARTER Balanced)
- Dale Whittington, Shaker Heights Public Schools (Ohio)(PARCC)
- Brad McMillan, Wake County (NC) Public Schools (SMARTER Balanced)

Organizer:
- Zollie Stevenson, Jr., Howard University/NATD President-elect

Moderator:
- Elvia Noriega, Richardson Independent School District (TX)/NATD Secretary

NCME Diversity and Testing Committee Symposium: Exploring the Implications of the "Fairness" Chapter of the 2014 Standards for Educational and Psychological Testing
Moderator:
- Meagan Karvonen, CETE, University of Kansas

Presenters:
- Laurie Wise, HumRRO

Perspectives from a Co-Chair of the Standards Development Committee
- Linda Cook, retired

Perspectives from a Co-Chair of the Fairness Chapter
Discussants:
- Edynn Sato, Pearson

Reflections from a Test Contractor
- Peggy Carr, NCES

Reflections from NAEP
- Brian Gong, NCIEA

Reflections from an Organization Providing Technical Assistance on State Assessment Systems

Invited Session: Advances in Score Reporting
Moderator/Discussant: Ron Hambleton
University of Massachusetts Amherst
Participants:
- Sandip Sinharay, CTB McGraw-Hill
- Shelby Haberman, ETS
- John Behrens, Pearson
- April L. Zenisky, University of Massachusetts Amherst

Invited Session: Measurement and Implementation Challenges in Early Childhood Education
Moderator/Discussant: Michael Rodriguez, University of Minnesota
Participants:
- Alisha Wackerle-Hollman, University of Minnesota
- Megan Cox, Minnesota Department of Education
- Ryan Kettler, Rutgers, the State University of New Jersey
- Scott McConnell, University of Minnesota
- Kristen Huff, Regents Research Fund

Invited Session: Standard Setting in the Common Core World: PARCC and SBAC Experiences
Moderator: Leslie Keng, Pearson
Discussant: Lauress Wise, HumRRO
Participants:

- Michael Bunch, Measurement Inc.
- Enis Dogan, PARCC Inc.
- Julie Miles, Pearson
- Joe Willhoft, Smarter Balanced Assessment Consortium

Invited Session: Quality Focus: Experiences from a Number of Assessment Programs
Chair/Moderator: Judith Monsaas, University of North Georgia
Participants:

- Henry Braun, Boston College
- Kristen Huff, Regents Research Fund
- Marianne Perie, CETE, University of Kansas
- Joe Willhoft, Smarter Balanced Assessment Consortium
- Gloria Zyskowski, Texas Education Agency

NCME ARCHIVES COMMITTEE UPDATE

Annie Davidson, CTB

The ad hoc Archives Committee has been working closely with the Board on an initiative to create an online repository for Annual Meeting papers and presentation slides. Initial plans for the repository needed adjustment due to the transition to our new management vendor in the coming year. We will collect voluntarily-submitted 2015 papers and presentation slides through a temporary email system after the Annual Meeting. These documents will then be made available to members online in a new system developed later in the year. Look for updates in the New Year!

The committee moves forward with the comprehensive archives plan to include both digital and physical materials. In the longer term, we envision a webpage on the NCME website that provides description of the archive, including but not limited to the Annual Meeting paper repository, as well as search capability. We will work closely with the Website Committee and the new management vendor to develop the database and interface as time and budget permit. Physical archival material will also be available to membership and housed with the management vendor.

Stay tuned for more on how you can contribute to the NCME Archives in the coming months! In addition to a request for your 2015 papers and/or slides, we will collect other artifacts and documents to feed the comprehensive archive. Of particular interest are digital materials not published on the website or in journals (e.g., obituaries, photographs, personal communications, out-of-print publications, any other materials you deem worthy). Make a directory now and save materials you think are valuable to the organization. We will be in touch when ready to receive them.

NCME FITNESS WALK/RUN CORNER

Brian French & Jill van den Heuvel, Co-Coordinators, NCME Fitness Walk/Run

We hope your fall season has been outstanding. Brian just completed his first snowy trail run over Thanksgiving. There is no better way to prepare for the few months of winter and the holiday season like snowy runs in the woods. We are in the planning stages for the NCME Fitness Walk/Run in Chicago. I am sure many of you remember the windy runs along the waterfront from years past. John Corrigan of JMC Partners will be assisting on-site this year. We await permit approvals from the city for a run along the water front. The proposed course should allow walking to the course instead of relying on buses.

Please remember to register for the run when you register for the conference. And encourage a friend to join!

Keep moving and Happy Holidays!

CALL FOR DISCUSSANTS: GRADUATE STUDENT POSTER SESSION

Please sign up to serve as a discussant for the GSIC poster session during this year's Annual Meeting in Chicago, IL. As a discussant you will be asked to provide valuable feedback to graduate student presenters on their papers before or during the electronic presentation. For discussants who may not be able to make the conference, you have the option to provide feedback by email, Skype, or phone to your presenter. The papers are 8 to 10 pages in length, and you may sign up for as many presenters as you wish. You may sign up as a discussant here:
https://docs.google.com/spreadsheets/d/13-3fDC-Lf5bt7vzEm04Q4CcqjftA3sXqWRIg3_nieWo/edit?usp=sharing
or by contacting Jason Herron at Jason.p.herron-1@ou.edu.

Thank you for your continued support of NCME's graduate population.

To get the NCME Newsletter four times a year (March, June, September, and December) go to http://ncme.org/publications/newsletter/

Send articles or information for this newsletter to:

Susan Davis-Becker
Alpine Testing Solutions
6120 Loma Circle
Lincoln, NE 68516

Phone: 402-483-5898
e-mail: susan.davisbecker@alpinetesting.com