FROM THE PRESIDENT
Laurie Wise, HumRRO

A lot is going on in our neighborhood. Here are updates on a few of the things NCME is doing, followed by some key developments in educational measurement to which many of our members are contributing.

Some of the key NCME activities in which I and other Board members have been involved include:

- **2015 Annual Meeting Program.** Over 600 paper and session proposals were received and are now under review. We are currently planning to bring back discussants for the individual paper sessions to provide some integration and perspective for these presentations. Also, the Archive Committee is working to establish a repository for papers and presentations from the 2015 Annual Meeting. Stay tuned for highlights of the 2015 program as they emerge.

- **Contract for Annual Meeting Support.** We have negotiated a contract extension with AERA for support and coordination of our Annual Meetings in 2015 and 2016. Our Annual Meeting Committee, chaired by Terry Ackerman, will consider and recommend options for Annual Meeting arrangements starting in 2017. See a separate report in this edition of the Newsletter for a report from that committee.

- **Strategic Plan.** The NCME Board spent time at our July meeting discussing activities designed to further our mission, including (a) ways to increase support for development of measurement professionals (e.g., sessions or a conference on emerging issues targeted for graduate students), (b) ways to support new knowledge in educational measurement (e.g., more journal articles), and (c) ways of expanding contributions to policy and practice (e.g., assessment literacy videos). My hope is that the Board will expand, evaluate, and prioritize activities to support these three aims and report back to the membership at our annual meeting.

- **Contract for Management Support Services.** Our current contract with the Rees Group expires in June, 2015. The Board decided it would be prudent to put out a request for proposals to be sure we continue to receive the best support possible for the cost. Rich Patz is heading up an evaluation team and the Board will review proposals at our October 7 meeting in Princeton.

- **Standards for Educational and Psychological Testing (Test Standards).** The 2014 version of the Test Standards have now been published and is available for purchase. Several members of the Joint Committee that created the new version provided a Capitol Hill briefing on the new Standards on Friday, September 12.

- **Test Standards Management Committee.** APA, AERA, and NCME each appoint a representative to manage funds generated from sales of the Standards for Educational and Psychological Testing and manage periodic revision of the Standards. We are negotiating with AERA and APA management to revise our agreement on how the Management Committee operates.

In the broader world in which we work, each of the Common Core Standards Assessment Consortia is finishing analyses of field test data collected last year and preparing to go operational as soon as December. Many states are also rushing to develop or purchase their own assessments aligned to the Common Core or to their individual state standards. A few states are continuing their current assessments of their existing content standards. Also, debate over the use of student assessment results in teacher evaluations continues. We hope to highlight progress and remaining issues on these topics at our Annual Meeting in Chicago.

The U.S. Department of Education will soon release draft peer review guidance that states must follow to meet validity, reliability, and fairness criteria for their accountability assessments. A link to the draft guidance will be posted on the NCME website once the draft guidance is released.
FROM THE EDITOR
Susan Davis-Becker, Alpine Testing Solutions

Happy Fall to everyone! This NCME Newsletter has some great information and insights from around our industry. The issue kicks off with an update from our president Laurie Wise including progress on key NCME initiatives. Dianne Talley presents her third graduate student column highlighting the importance and role of professional serve in our development within the field. Our Spotlight member for this issue is Billy Skorupski who shares a little bit about himself and his professional development. In our first feature content of this issue, we are fortunate several presenters from an AERA 2014 session on validity share their perspectives with us – very thought provoking. Our second feature content is a legal corner by S.E. Philips addressing standards for intellectual disability which underscores the importance of the implications of our work. Finally, we have several important updates regarding the annual meeting from Terry Ackerman (chair of the annual meeting committee), Brian French and Jill van den Heuvel (co-chairs of the Fun Run), and Ye Tong, Jennifer Randall, and Caroline Wiley (conference co-chairs and training chair).

GRADUATE STUDENT CORNER: A REFLECTION ON SERVICE
Diane Talley, University of North Carolina, Chapel Hill

While recently reflecting on my ongoing graduate school experience, and on what I want from this final phase of my education, I was reminded of the oft-emphasized obligation of service. This is a term whose definition and precise application to my graduate studies has been elusive. I have heard it spoken in terms of academic service, community service, public service, and service learning. I’ve searched university websites and student handbooks in vain for a clear singular definition of service as it should apply to graduate students. Though there is an emphasis of service learning and community involvement for undergraduates and academic faculty (in fact, there are entire books written on these topics) little reference appears to the service obligations of graduate students.

The University of North Carolina Graduate Student Handbook characterizes research and teaching assistantships as service. Though both certainly provide benefit to the university and the student; is that as far as it goes? Do we as graduate students have no further obligation to service in our communities than what is directly related to our academic programs and assistantships? I would like to share an enlightening lesson I recently received regarding service, and some thoughts on what our community of service is and how we can serve that community.

The concept of service was clarified for me in a conversation with Dr. Lynn Blanchard at the Carolina Center for Public Service. This organization works to help students and faculty of UNC Chapel Hill serve local and global communities through engaged scholarship.

Two important points emerged from this conversation. First, service is not to be confused with volunteerism. Volunteerism is what we do when we paint houses for Habitat for Humanity or spend a Saturday at the local soup kitchen. Service, however, exemplifies the idea of applying our knowledge, our scholarship, to benefit our community. According to Dr. Blanchard, the goal of academic service is to “use the energy and expertise of the university to solve problems” within a community. That community can be defined in many ways depending on the particular field of study.

Dr. Blanchard also emphasized the concept of reciprocity. We have something to give to our community through engaged scholarship and this engagement in return enriches our educational experience and prepares us for future careers as professors, researchers, and psychometricians. Our community is thus defined beyond the university; to the professional community we will serve for the duration of our careers. Engaged scholarship is, thus, an application of our learning, research, and intellectual creativity to the benefit of that community.

How, then, can we serve both our university and the measurement community during our time as graduate students? And how do we do so with the concept of reciprocity in mind?

It is true that our ultimate goal in graduate school is to complete our degrees. Though not altruistic, it is indisputably in the best interest of the university, the measurement community, and ourselves to achieve this goal. Thus, as we consider how to serve our communities, we must ensure that such service aligns with our own academic and professional goals. Following are ideas to consider:

Serving the University
Since this institution is our current home and base of learning, service to the university seems a logical starting point. We are all part of a community of scholars that include fellow graduate students, faculty, and occasionally undergraduate students. Teaching and research assistantships are one way to serve both the university and your education. Presenting your work both within and outside your particular programs is another option. Graduate student government provides an additional service opportunity that not only aids the university, but benefits other graduate students through your representation of their interests, and provides leadership experience that is valuable for future careers in and out of academia.

**Industry Conferences and Organizations**

Attending and participating in industry conferences is a great way to be an engaged scholar. It serves the community and allows you to gain experience and develop relationships with others in the industry that may help you in conducting research and starting your career. It also has the potential to broaden your world view through the knowledge and insights of a diverse group of scholars. Participation benefits the measurement community when you share your theories and research, engaging other scholars in your academic interests and actively addressing problems that the community is facing.

Participation in a conference is not limited to presenting. There are often opportunities to work on conference committees, review submissions, assist with the registration, or run conference events. These activities are excellent networking venues and provide much needed assistance to the conference organizers.

Getting involved with organizations such as NCME and AERA beyond the annual conferences can also provide service opportunities. Serving on committees or contributing to industry publications (this newsletter, for example) can be beneficial. Other measurement-related organizations that may also provide opportunity include the Association of Test Publishers, the Institute for Credentialing Excellence, and the International Test Commission.

**Research**

Research, whether conducted as part of the graduate program (i.e., dissertation), through an assistantship, or independently, is a critical part of our graduate education that also serves the interests of our universities and professional communities. Assistantships are particularly helpful since they provide both academic and, in many cases, financial support to graduate students. In return, the support that these assistantships provide to funded research projects is invaluable to the university and may also help to address practical issues in the measurement field. Independent research is also beneficial to the university as any quality publication associated with the university may increase its reputation, in addition to the contribution that the student is making to the measurement community and their own resume.

**Internships**

Internships are an excellent way to engage with other scholars/practitioners, learn from their experiences, and contribute to the work being done in the measurement community. Internships may be research focused, conducting studies that solve theoretical and practical measurement problems; or they may be applied, such as working with test developers and psychometricians in a testing organization or licensing board in the actual development of examinations.

Internships not only serve the measurement community, but may also provide the opportunity to engage with communities outside of this field. Psychometricians often work with subject matter experts across a variety of industries (medicine, law, accounting, etc.), leading to a greater shared understanding of measurement across industries.

**Final Thoughts**

Though we are not bound in any contractual manner to serve our universities and our field of study, the potential benefit to these institutions and to ourselves is substantial and worth consideration. There are undoubtedly many ways to serve beyond those discussed here. But perhaps this will inspire you, as it has me, to consider service in your own graduate experience.
SPOTLIGHT ON THE PEOPLE WHO MAKE OUR ORGANIZATION GREAT:  
BILLY SKORUPSKI, ASSOCIATE PROFESSOR, UNIVERSITY OF KANSAS

How did you get into the field?
As an undergraduate at Bucknell University (Lewisburg, PA), I started as a psychology major interested in doing research. I had always been good at math, but I didn’t like the idea of being a math major. I hadn’t really discovered statistics as a field of study yet. I chose psychology so I could do quantitative research, but still have a “human element” to the work. While taking psychology courses, I branched out and took an educational psychology course in the school of education. I quickly discovered that I was even more interested in educational research. I learned that Bucknell has a Bachelor’s of Science degree in educational research, so I ended up double-majoring in that and psychology. As an undergrad I took classes on Research Methods, Statistics, and even one on Tests and Measurements. That really opened up my world. I realized that measurement affects every single science, so studying how to do it well was vitally important. I remember thinking, “so, I don’t have to choose a ‘substantive area?’ I can just study methods?” I was so happy! I decided that I was going to be a psychometrician, and I’ve never looked back.

If you weren’t in this field, what would you do?
That is a tough question. I really like computer programming, and problem solving in general, so maybe something along those lines. I love science. I’ve often thought that I probably would have been pretty happy as an engineer of some sort. I’ve also always been very interested in astronomy. All of those are fairly quantitative pursuits that deal with measurement and data modeling, to some extent…so I guess I chose a pretty good field for myself!

What advice would you have for graduate students who want to get into this field?
The coursework in psychometrics is obviously very important; there is a lot to learn. Take all the classes you can, and especially learn some programming! But everyone in school knows the classes are important. What’s perhaps even more important are the learning activities that take place outside the classroom. I always encourage students to get involved with research projects as much as possible. Approach your advisor or another professor with research ideas, or ask if s/he could use some help with a current or future project. Likewise, engage your fellow students in conversations about research. I love it when I see our graduate students collaborating on projects they have conceived and implemented on their own! Lastly, I strongly encourage graduate students to apply for summer internships. These are amazing (paid!) learning opportunities that get you involved with research, build your professional network, and provide insight into what your later career might be like.

What do you like to do for fun outside of work?
I am a long distance runner, so that takes up a number of hours a week. I love running. It takes away my stress, clears my head (I do my best computer programming when I’m out on a run), makes me feel fit, and lets me get away with having a less-than-perfect diet! I also love playing music. I play piano, guitar, and bass; for the last few years, I have almost exclusively been playing piano, usually about an hour a day. My wife and I are both musicians, it’s how we met in college. Our house has this crazy music room, in what would be the dining room in a normal house. We have a piano, electric piano, trombone, bass and guitars, amplifiers, and a full drum set. Our house is never quiet! Part of that is we have two wonderful children, Ike (6) and Millie (4), so spending time with them (often playing and listening to music) is our primary fun non-work activity!

What would you say has been one of the biggest innovations in psychometrics in the last decade or two?
I think the biggest innovation in psychometrics has been the accessibility of Bayesian methods, particularly Markov Chain Monte Carlo (MCMC), for specifying and estimating complex measurement models. These methods are by no means brand new, but software and computer processing speed have only fairly recently caught up with the computational demands. As a result, there has been an explosion of new measurement models over the last ten years or so. Some models are multidimensional and/or hierarchical. Many have been generalized from dichotomous items to include polytomous ones. Some allow the population to be specified as a mixture of latent classes, and include combinations of fixed and random effects. Measurement models can now easily accommodate collateral information in the form of covariates (perhaps at different levels) for
explanatory purposes. A lot of these approaches exist in the research literature, but are not part of operational practice...yet. Some are ready to solve problems we don’t even have...yet. But the great thing is that psychometricians, myself included, are able not only to dream up these cool models, but are now also able to program our own algorithms for estimating their parameters and demonstrating their effectiveness.

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

I pick sessions based on who the speakers are, not necessarily the topics. I certainly like to hear what prominent members of our field have to say about measurement. But I also like to attend sessions where I know the speakers are really dynamic and interesting. You can’t really learn all that much during a conference presentation (I tell this to students all the time). Conference presentations are really advertisements for the paper; they are too short for a speaker to say everything. I like to hear speakers that understand this and can effectively communicate the major themes and findings, making me want to read the paper. It doesn’t hurt if a speaker has a sense of humor, either!

Who has been a significant influence in your professional life?

My professors from graduate school at UMASS Amherst have been hugely influential. Ron Hambleton taught me IRT, Hariharan Swaminathan taught me multivariate statistics, Steve Sireci taught me test construction, and Jane Rogers taught me computer programming. My academic advisor, Lisa Keller, taught me how to get research done! These people have all shaped how I think and the work I do. I was also extremely fortunate to participate in a summer internship at the National Board of Medical Examiners, working with Howard Wainer. Howard and I became fast friends and work really well together. I learned and continue to learn so much from him. He’s largely responsible for my being a Bayesian! Since my internship, over a decade ago now, we have always kept in contact and continue to work together. (We are currently working on two different projects!) Howard is both a mentor and a friend.

INNOVATIVE VALIDITY APPROACHES FOR HIGH-QUALITY ASSESSMENTS: AERA SESSION FOLLOW-UP

At the AERA 2014 conference, several members presented their perspectives on innovative validity approaches following a recent publication, Criteria for High-Quality Assessment (Darling-Hammond et al., 2013). In this Newsletter, we are fortunate to have several of the authors and discussants share their perspectives from this intriguing session. For those who are interested, a recording of the entire session can be found here:


Valid Systems of Assessment to Support Intended Purpose

Joan Herman, CRESST

Validity, as we all know, is not a state of being. Rather it is an accumulation of evidence that justifies the use of assessment scores for specific purposes (AERA, APA, NCME, 1999). It does not reside in a single indicator, but rather is advanced as an argument, using multiple sources of evidence, to substantiate specific claims about the precision, accuracy, appropriateness, and consequences of particular score inferences and use (AERA, APA, NCME, 2014; Kane, 2001; Linn, Baker, & Dunbar, 1991). As a corollary, we also know that a single test cannot serve all purposes. Different kinds of assessment are needed to support accountability and improvement. With massive investments being made in a new generation of assessments, we need to apply these important lessons, which are well articulated in three important new reports: the recently released revisions of the Standards for Educational and Psychological Measurement (AERA, APA & NCME, 2014); Criteria for High Quality Assessment (Darling-Hammond et al., 2013); and Developing Assessments for Next Generation Science Systems (Pellegrino, Wilson, Koenig, & Beatty, 2014).

These efforts converge in making the case for multiple criteria for evaluating state tests and for substantiating the validity argument that state assessments can serve their intended purposes. Here I want to highlight three criteria, drawn from Darling-Hammond et al. (2013), on which state assessments traditionally have been very weak: assessment of higher order cognitive
skills, high fidelity assessment of critical abilities, and use of items that are instructionally sensitive and useful (https://edpolicy.stanford.edu/publications/pubs/847). In a nutshell, these criteria boil down to the extent to which state assessments reflect the range and richness of now current college and career ready (CCR) standards for student learning and are useful for gauging and improving learning. If state assessments do not dramatically improve their performance on these criteria, research indicates that students’ learning and achievement of CCR standards will suffer.

Developing Assessments for the Next Generation Science Systems focuses on the assessment of next generation science standards, but provides vital cues for improve performance on these criteria in any subject area (see http://www.nap.edu/catalog.php?record_id=18409). The report suggests that we need to get serious about designing coordinated systems of learning-based measures, including those explicitly designed to support classroom teaching and learning and those designed to support accountability. We need to get beyond the idea that external, on-demand assessments alone can get us where we need to go in either assessing all that is important for student learning or in supporting teaching and learning.

Instead, the report advocates that states combine two types of external assessment strategies: on-demand assessments developed or selected by the state and administered at a time mandated by the state; and classroom-embedded assessments, developed or selected by the state or district, but administered at a time to fit with local curriculum content and sequences. Recommendations are roughly akin to the summative assessments being developed by PARCC and Smarter Balanced summative assessments—mixed item formats with extended written responses and/or performance tasks—but with the suggestion that matrix sampling might be considered, depending on intended accountability use.

The most novel thrust of the report lies in its advocacy for the inclusion of classroom embedded assessment as part of state monitoring and/or accountability systems. What might these look like? Certainly there are any number of possibilities. Here are a few examples:

- Replacement units (model curriculum materials with embedded formative and summative assessments) developed outside of the classroom (by state or district), where students’ end of unit performance could contribute evidence of learning for monitoring and/or accountability as well as classroom purposes.
- Banks of rich performance tasks, developed outside of the classroom that teachers could use at key points in their curriculum to provide evidence of student learning for both formative and summative purposes.
- Portfolio collections of work samples, with task types specified by the state or district, to demonstrate what students have learned over the course of the year.

Critical in all is that they reflect standards that cannot be well assessed by the on-demand component; are educative for teachers and students in communicating the kinds of instruction, assessment, and deep learning expected by CCR standards; can be embedded as part of on-going curriculum and instruction; and provide formative data to support learning improvement and summative data for teachers and monitoring purposes.

Certainly these recommendations echo suggestions we have heard before and carry with them thorny technical and comparability challenges. But some existence proofs do exist internationally and in some student accountability systems. I wonder whether and how we—NCME—can help to move this agenda forward and help to bring stronger, richer assessment systems to fruition?

References

Or
Suggested citation:


### Functional and Transformational Validity

**Eva L. Baker, UCLA/CRESST**

When validity and quality are joined in a discussion of testing, the first topic usually addressed is the purpose(s) of the examination, followed by excruciating or nominal efforts to present evidence that the inferences drawn from results are appropriate. Our team was interested in the next steps following a fall in performance on new tests: the validity element of instructional sensitivity (Linn, Baker, & Dunbar, 2001).

Given the typically technical nature of the quantitative argument made on behalf of validity, it is easy for non-specialists to get lost in the detail, with a likely result of acceding to “experts” validity claims. The development of new standards, whether State generated or part of the Common Core State Standards, brings with it the requirements for new measures. Given the breadth and depth of the Standards, it would be most desirable for extensive periods of engagement on examinations for students to demonstrate their progress, even if only for a few of the Standards. As salient as the Standards and assessments have become (“Are they too hard?” “Do they depart too much from revered tradition, e.g., fiction?”), they embed even more dramatic implications than simply novelty and challenge.

First, they are required explicitly to contribute to multiple purposes beyond student learning, for instance, to play a role in teacher evaluation. More importantly, the teacher evaluation goal requires an additional, new purpose for the test, that is, that teacher knows how to analyze its content, its data—common errors, and develop instructional appropriate approaches to improve students’ performance the next time around. I opine rare data on this point is available. Given the start of the court case in Atlanta alleging rampant cheating by educators to achieve high test scores, it is critical that the test designers and analysts provide formative cues for teachers even for state assessments, so they are not left to desperate and immoral methods to protect their jobs. Even without the headline of adult cheating, test prep itself cheats students. While it may prepare them in perseverance and tolerating boredom (two important life-long skills), preparation to pass a restrictive sample of items does not an education make.

Despite the wealth of specifications prepared by vendors, in our Bill and Melinda Gates study of a new State assessment, we found that careful identification of features of the item or task, either the demands in the stimulus, the intervening thinking needed for the response, and the response, differentiated from wrong answers required more granular thinking. By looking at attributes related to the assessment (content alignment, task requirements, for instance how much and what kind of text or graphics were used, specific cognitive demands such as types and degree of problem solving), we were able to identify attributes for meta-tagging items. This analysis of features would allow reporting not only for total score, but also across samples of items that shared certain features, for instance, to be able to discern the performance of students on problem solving elements. Most of this work happened in accessible language rather than in lovely, imponderable sets of equations. The data itself identified sets also associated with growth. Thus, the analyses of the annual State assessment data provided a beginning point for the review of instruction with an eye to helping students do better. Across grades and subjects, the features accounted for an average of .50 of the variance associated with performance.

Paired with the innovative analyses done by Cai and Choi, we were able to disaggregate items for which there was no movement over time from those where we could see performance growth (grades 3/4; 7/8; math and English Language Arts). We look to undertake additional work to see whether the item classifications are attributed to poorly developed items, those that almost wholly depend upon out of school factors, or those for which more adequate instruction could yield improved performance. We are hoping to follow up this work with studies of classrooms where we think we might find evidence of the latter, and are prepared to describe, given needed replication, the replacements for specifications and items or tasks that will increase the sensitivity of the measure to instruction and thereby, its legitimacy for use in educational evaluation. We called this “validity”
approach functional, for although dependent upon careful description and analyses, its greatest good will be found in its ability to guide productive change in teaching and learning. Everything summative is formative again.

**The Role of Hierarchical Item Response and Diagnostic Classification Models in Functional Validity Studies**

*Li Cai, CRESST/UCLA*

With support from the Bill and Melinda Gates Foundation, and in partnership with a State Department of Education, CRESST conducted qualitative, quantitative, and integrated analyses in early 2014 of multiple years of summative assessment items in both math and ELA, and longitudinal raw item response data, both prior to the State’s adoption of the CCSS and after standards implementation. Preliminary results were presented at an invited session at AERA in Philadelphia in April, 2014. From the analyses, CRESST has begun the development of a functional framework for assessment development, validation, instructional design, and improvement. There are several noticeable new features in our approach worth highlighting.

First, one of the validity research questions that we hoped to answer whenever new content standards are adopted and implemented is whether we can isolate the effects of standards based instruction on student performance on summative assessment items. In standard measurement models (e.g., unidimensional item response models), item parameters such as difficulty are assigned the role of describing an item’s psychometric property and how one might expect a student with a certain level of achievement to respond. While entirely satisfactory for many purposes, the standard approach does not explicitly account for the potential effects of standards based instruction on student performance, and instead merges prior knowledge or other individual differences into a single unitary dimension of achievement. Our technical approach combines recently developed hierarchical item factor analytic tools with diagnostic classification modeling to disambiguate the contributions of students’ prior knowledge or individual differences non-specific to the standards from those that are potentially more related to instructional improvement due to standards implementation efforts. We are aided by a distinguishing feature of newly developed assessments in that the items are specifically prescribed to provide measurement of mastery of knowledge or attributes as specified by the test blueprints aligned to the new standards. The test design thus provides the item x feature matrix needed for the diagnostic classification part of our models, which is convenient for confirmatory validation purposes. Furthermore, by defining higher-order dimensional latent variables from the attributes, we obtain estimates of relative location of individuals on both prior differences as well as differences potentially attributable to standards implementation and consequential learning.

Second, it is almost proverbial that score drop is observed whenever new assessments are put in place to support adoption of new standards. We submit that typically a relatively large proportion of observed variation in raw item responses cannot be explained by observed or latent variables related to the implementation of standards. Therefore, to use scaled scores from standard modeling approaches would amount to trying to sail a ship with tons of deadweight. We argue that our technical approach that hierarchically decomposes item-level performance into instructionally relevant components should tend to yield improved sensitivity in detecting changes in student performance toward the criterion of college and career readiness because it amounts to statistically controlling for a background covariate highly related to the outcome in an evaluation study. We are aided by the availability of longitudinal data so we can compare student performance across grades. Indeed our results show positive gains in learning.

Third, we believe that our models operate at an intermediate level between highly aggregated domain-level scaled scores (which are difficult to act on instructionally) and raw item-level performance (which may be very tempting to be teaching to, but counterproductive in the larger sense). This way assessment design and instructional design related feedback can be provided. The attributes or features in our models are either content related (e.g., as defined by the CCSS), or cognition or linguistic features that may generate feedback either for improving the instructional sensitivity of the items, or can be combined to larger units of aggregation (e.g., classroom-level) for instructional feedback. In other words, the likely results from our approach explicitly encourage actionable feedback but discourages teaching to the test.

Finally, we believe that while our first forays have been encouraging, there remains a great deal to be learned and discussed. For instance, a major shortcoming of examining student outcomes alone without corroborating information about instructional practices potentially renders all of our logic circular: we are not certain whether the observed differences are due to the presence of lack of an item’s instructional sensitivity or due to the success or failure of implementation. As new Consortia-developed assessment begin to roll out in the 2014-15 school year, and more data become available, we believe there will be plenty of potential for new research that may interest the members of NCME.

**Discussant Perspectives**
The Political Economy of Assessment
Jack Buckley, College Board

I first read Darling-Hammond et al.’s “Criteria for High-Quality Assessment” with great interest last year while working at the U.S. Department of Education. The authors distill much of the best recent thought on assessment into five guidelines that seem eminently reasonable (though, I suppose, one could quibble about international benchmarking truly being necessary). In particular the fifth standard, that assessments be valid, reliable, and fair, is a central tenet of our profession. So why, then, does it seem so unlikely that, as a nation, we will be successful in introducing such next-generation, high-quality, assessments into our classrooms? Why are many testing industry professionals skeptical of the viability of any new high-stakes assessment that deviates too much from the status quo of both cost and features? The trivial answer is: economics and politics. But the problem is worth considering in greater detail.

One way to understand why our assessment system looks the way it does is to consider the political economy of another important part of American life that few are satisfied with: our healthcare system. Public policy scholar David Wilsford has argued that healthcare in the United States—and in many other nations—is the result of “path dependence,” a process wherein early decisions establish constraints that prevent large scale, sweeping change later on, or at least make it highly unlikely. As a result, even a significant change like the Patient Protection and Affordable Care Act is less a social revolution than a bundle of incremental adjustments to a complex system.

Another reason why systems like healthcare—and, I would argue, educational assessment—look the way they do is because competing forces operate under specific “rules of the game” (or, to a political scientist, “institutions”) that further constrain the range of possible and likely outcomes. Combine these institutional rules with path dependency, and the prospects of radical change are bleak.

Who are the actors that shaped the early decisions—and created the institutional rules of the game—in educational assessment in the United States? Far enough in the past, say in the 1920s when the SAT was introduced, the major players were a limited set of elite colleges, their feeder high schools, and some academic researchers. More recently, around the passage and implementation of the testing provisions of No Child Left Behind, the range of involved actors is orders of magnitude larger: Congress, successive Presidential administrations, state government (both executive and legislative), the court system, school district leaders, mayors, a wide range of business interests, unions, advocacy groups, professional organizations, and more have joined the fray and made assessment into one of the most highly politicized areas of domestic policy.

Although there are many areas of educational assessment, each with its own set of institutional rules and actors, a quick look at one of the most important recent initiatives—the Race-to-the-Top Assessment grant program (RTTA) is instructive. When the Obama Administration introduced the RTTA program, part of their theory of action was that consortia of states would be able to produce higher-quality assessments than the status quo, where states enter into individual contracts with a limited set of vendors to build tests to idiosyncratic state standards. These new assessment consortia would, partly through federal grant funding but also through their novel governance structure, be freed of the constraints that make today’s typical state tests look like they do. Furthermore, it was hypothesized that this new model would serve as an incentive for new entrants to the testing marketplace—primarily technology companies that would act as a force for disruptive innovation, increasing efficiency and driving down costs.

I believe, however, that the architects of this policy underestimated both the path dependent nature of assessment and the difficulty of revolutionary change in this area. Few new firms entered the market to compete for parts of this work and, as projected costs per student looked high, many of the most innovative parts of the new assessments were jettisoned and other aspects, like the high-quality and validity studies agenda called for by Darling-Hammond et al., look less and less likely for these tests.

Other educational assessments, however, have the sort of validity studies infrastructure that the authors call for, notably the National Assessment of Educational Progress (NAEP) and, I would argue, the leading college entrance examinations—ACT and SAT. NAEP’s program of validity evidence and research, in particular, is unmatched in the field. The National Center for Education Statistics, the National Assessment Governing Board, and the NAEP contractors all fund, sponsor, and/or participate in dozens of technical panels and advisory committees that bring together leaders in our field to advise the government, conduct original validity research, and generally support the high quality of the program. Likewise, the major college entrance assessments are built upon decades of validity (and other) research, from content surveys to longitudinal predictive data collection and modeling.
So why can this sort of validity infrastructure exist for some educational assessments but be so difficult to implement for others? I believe a major reason is that they operate under different institutional rules of the game. Take the National Assessment: a high level of national political attention on NAEP at its inception made it essential to demonstrate the highest levels of reliability and validity that were feasible under cost and technological constraints. Moreover, compared to most state testing programs, these constraints were not severe. Cost has been less of a constraint for NAEP because of its federal nature (although there is never enough available funding to execute the ambitious plans of the Governing Board) and technology has been less of a constraint for several reasons including that NAEP’s status as an Interesting Problem for researchers has ensured a steady flow of outstanding human capital.

The most sure way of ensuring that the nation’s next round of high-stakes K-12 assessments are high quality is to closely examine the prevailing set of institutional rules and adjust them as necessary to create the necessary environment. This means paying attention to the details of state and federal legislation, but also to regulation and guidance—particularly guidance around aspects of state assessment like federal peer review. It also means rethinking the cost structure of the assessments to ensure that they are feasible for states but that features like a robust validity agenda move from “nice to have” to an integral part of steady-state operations but perhaps funded from a different source than per pupil state charges.

Building high-quality, high-stakes educational assessments is not impossible, but we won’t get there by playing under the same rules and hoping things get better. We need to think carefully about why our tests look like they do and how to structure the politics and economics to be more conducive to a better future.

Comments on an AERA Presidential Session

Douglas Becker, HMH-Riverside

I was indeed honored to be part of the AERA Invited Session — “Innovative Validity Approaches for High-Quality Assessments: An Interaction.” I extend my thanks and appreciation to Joan Herman, Eva Baker, and their co-presenters for inviting me to participate in this session. As noted in the program guide, this symposium was organized to address key validity criteria involving both assessment features and inferences drawn from data, and how these criteria can support high-quality assessments for measuring complex thinking and applicable domain learning. From my perspective, however, there really wasn’t much focus on innovative validity approaches as suggested in the session title, nor was there sufficient opportunity for interaction. Rather, the session provided a platform to advance a perspective toward assessment largely reflected in the Criteria for High-Quality Assessment (Darling-Hammond et al., June 2013). Just as an aside, the attendance was fine but not overwhelming at this session, and this is no small feat given it was scheduled opposite the NCME Breakfast Meeting.

It was worthwhile that the session afforded an opportunity for some of the authors of the Criteria for High-Quality Assessment to present their views and portions of the associated work that has been conducted in support of their positions. And I would like to say right up front that I have no conceptual problem with the idea of Criteria or the spirit in which they are presented. My comments really hit on two main points: (1) Let’s choose our language carefully, and (2) It’s going to take all of us.

I think it is important to point out some language from the session proposal that really aroused my interest and helped motivate me to participate in the session. The opening paragraph of the proposal stated, “A number of thorny questions loom, including whether the assessments will measure up both to the spirit of the standards and to the complex performance that many standards specify. These questions are all the more important because of the financial straits of many States. The attendant fear is that a cheap test will similarly cheapen and undermine the intention of the CCSS.” (bold added)

My first reaction was this: As a practicing member of the greater test-publishing community, I have never worked for or with a test publisher, nor have I ever worked on a K-12 assessment contract for a state or district, where cheap or inferior assessments were developed in response to a mandate to assess state standards. But then I calmed down. And, I arrived at my first point. We really do need to choose our language carefully. We really don’t need the academic, research, and test publishing communities taking (cheap?) shots at each other. Additionally on this point, there needs to be some clarity as to what is actually meant when we speak of assessments. The aspirational side of the CCSS would like to have us think of a well-integrated SYSTEM of assessments—formative, benchmark, and summative—all working together and espousing all the ideals that the CCSS put forth. Reality, however, is an entirely other matter, and the conversation quickly switches to (and focuses on) the summative and accountability components of an assessment system. Once this happens, aspirational meets practical, and we end up on a road that is less than what the CCSS would desire. So we need to be clear whether we are talking about an assessment system or a component of an assessment system—it is easy to villianize an assessment while forgetting the
My second point is pretty basic: It’s going to take all of us. It’s going to take academics, researchers, policy wonks, SDOEs, and even test publishers, all working together, to move the needle and achieve the dream of next-generation assessment systems. So the push for developing and promoting criteria around high-quality assessment can’t be one-sided; in fact, it should capitalize on the strength of existing relationships as well as seek input from experts and innovators from a variety of communities. As just one example, the Association of Test Publishers (ATP) has worked cooperatively with the Council of Chief State School Officers (CCSSO) to publish two editions of Operational Best Practices (CCSSO, June 2013). The reason I highlight this collaboration is simple—when the rubber hits the road and a K-12 assessment program needs to be implemented, it’s the SDOEs and the test publishers who by and large get the job done. Nor should the K-12 assessment community think that they have to go this alone. Outstanding work by our measurement colleagues in the Clinical, I/O, and Licensure/Certification communities can certainly serve to inform CCSS work. The National Board of Medical Examiners, for example, has decades of published research on computer-based case simulations and their application to high-stakes assessment. And yet, I can’t recall seeing this research referenced in any CCSS assessment work. We need to be willing to look beyond our K-12 education walls.

In summary, I have no conceptual problem with the idea of Criteria or the spirit in which they are presented. But the Criteria can’t belong to one group exclusively. So I will choose to view the Criteria as an invitation to engage in further conversation. Through carefully chosen and purposeful language, and with a broad spirit of inclusion, we can hope to achieve the dream of developing and delivering high-quality next-generation assessment solutions.

Note: Thanks to a few of my close colleagues for their helpful reviews and comments.

---

**LEGAL CORNER: SETTING STANDARDS FOR INTELLECTUAL DISABILITY**

S.E. Phillips, Consultant

What do IQ tests and convicted felons have in common? Identification of intellectual disability. A recent U.S. Supreme Court decision, Hall v. Florida,1 has addressed the question What role should IQ scores play in the identification of offenders with intellectual disabilities ineligible for capital punishment? and is the subject of this Legal Corner column.

IQ test scores from individually administered instruments such as the Stanford-Binet and the Wechsler have long been used in education to assist school psychologists in identifying students with IQs in the average range or above whose academic achievement is significantly below their cognitive ability (students with learning disabilities). IQ test scores have also been used to identify students whose cognitive functioning is substantially below average (e.g., IQs below 70 indicative of what has traditionally been labeled mental retardation) qualifying for alternative educational standards and placements. In both cases, performance standards related to IQ tests established a rebuttable presumption of disability for which additional supporting evidence was required for confirmation.

The judicial system has also used IQ scores to assess intellectual disability. Intellectual disability is considered a mitigating factor when sentencing a convicted felon because persons with intellectual disabilities are considered less culpable than the average criminal even if they can distinguish right from wrong. According to the Court, the notion of reduced culpability for an offender with an intellectual disability is based in part on the “evolving standards of decency in our maturing society” and the fact that a cognitive impairment renders the imposition of capital punishment less defensible as retribution for past crimes and less likely to act as a deterrent for future crimes. Until 2002, convicted felons with intellectual disabilities in states with capital punishment could introduce evidence of the disability as a mitigating factor during sentencing. But if the aggravating factors of the crime outweighed the mitigating factors presented by the defense, a convicted felon with an intellectual disability could be sentenced to capital punishment. Such was the determination of the Florida courts at the time Hall was tried and convicted of murder.

**Capital Punishment of Individuals with Intellectual Disabilities**

---

Standards for intellectual disability took on added importance in 2002 when the U.S. Supreme Court ruled in *Atkins v. Virginia* that the Eighth and Fourteenth amendments barring cruel and unusual punishments rendered capital punishment (but not lifetime incarceration) unconstitutional for convicted felons who are intellectually disabled. The Court stated that capital punishment violates the inherent dignity of individuals with intellectual disabilities and serves no legitimate purpose of deterrence or retribution because individuals with intellectual disabilities “have a diminished ability to process information, to learn from experience, to engage in logical reasoning or to control impulses” (p.320). The Court further held that the “diminished capacity of [individuals with intellectual disabilities] lessens moral culpability and hence the retributive value of the punishment” and that prohibiting capital punishment “protects the integrity of the trial process” because individuals with intellectual disabilities are at increased risk of wrongful conviction from false confessions, being poor witnesses, and being unable to meaningfully assist their attorneys (p. 319-321). The effect of the *Atkins* ruling was to greatly increase the significance and importance of intellectual disability by transforming it from a mitigating factor in sentencing to a factor precluding capital punishment under all circumstances.

After reviewing the existing criteria used by states to determine intellectual disability and finding them inconsistent, the *Atkins* Court left it up to the states to determine how IQ scores and other evidence would be used to reliably ascertain which defendants were intellectually disabled. At that time, the use of a threshold IQ value in some states versus the estimation of IQ score confidence intervals in other states indicated a likelihood of continued inconsistency in findings of intellectual disability based on similar circumstances. It took until 2014 for the *Hall* Court to revisit the appropriateness of the standards set by states for intellectual disability determinations. In particular, the *Hall* case challenged Florida’s interpretation of its statutory standards for demonstrating intellectual disability for a defendant who had been sentenced long before the *Atkins* case was decided but whose capital punishment had not yet been administered.

**The Florida Statute**

The Florida statute defining intellectual disability was enacted just prior to the Court’s ruling in the *Atkins* case. It provided

> [Intellectual disability] is significantly subaverage intellectual functioning existing concurrently with deficits in adaptive behavior and manifesting during the period from conception to age 18. The term “significantly subaverage general intellectual functioning,” for the purpose of this section, means performance that is two or more standard deviations from the mean score on a standardized intelligence test specified in [Agency rules]. The term “adaptive behavior” for the purpose of this definition, means the effectiveness or degree with which an individual meets the standards of personal independence and social responsibility expected of his or her age, culture group, and community. [Agency rules shall] specify the standardized intelligence tests [that are acceptable]. (Florida Statute 921.137).

The Florida statutory definition of intellectual disability specified three criteria that a defendant was required to demonstrate by a preponderance of the evidence (>50% probability): (1) significantly subaverage general intellectual functioning, (2) deficits in adaptive behavior, and (3) onset prior to age eighteen. Agency rules specified the use of a Stanford-Binet or Wechsler IQ test. Given a mean of 100 and a standard deviation of approximately 15 on both IQ tests, two or more standard deviations below the mean signifying subaverage intellectual functioning was understood to mean an IQ score of 70 or below. The Florida Supreme Court interpreted the statute as having three conjunctive requirements such that if the first requirement was not satisfied (i.e., the offender’s IQ score was above 70), there was no need for further inquiry.

Adaptive disability was a matter of subjective judgment and considered other relevant evidence of past performance, environment and upbringing, including testimony about such factors as medical history, behavioral records, school tests and reports, past behavior, family circumstances, inability to learn basic skills, and inability to adjust behavior to changing circumstances. Onset prior to age 18 was generally established through school records and the testimony of educators, family, and friends regarding behavioral indicators of adaptive disability during childhood.

**Factual Background of the Hall Case**

The evidence of Hall’s culpability for the crimes charged was apparently unrefuted. The crimes originated in 1978 when Freddy Hall and his younger friend, Mack Ruffin, decided to rob a convenience store. Hall and his accomplice needed a car for their crime so they drove to a local supermarket to find one to steal. In the parking lot, Hall forced a 21-year-old pregnant newlywed into her car and drove her car to a wooded area outside town. Ruffin followed in his car. The two men then beat, raped and killed the woman and afterward drove her car to the convenience store they intended to rob. As they exited the store, Hall shot an investigating officer in the parking lot and took his gun. The gun used to kill the woman was later found under the body of the officer. Hall and Ruffin escaped in the stolen car and a chase ensued. They abandoned the vehicle, fled on foot and were

---

caught soon after. Hall was convicted of murdering the woman and sentenced to capital punishment. A long series of appeals followed culminating in the appeal of Hall’s capital punishment to the U.S. Supreme Court on the grounds that it was unconstitutional because of his alleged intellectual disability.

The focus of the challenge was on the Florida Supreme Court’s interpretation of its statute as requiring an IQ score of 70 or below to establish the subaverage functioning prong of its 3-part test for qualifying a defendant as intellectually disabled. Hall was not given the Stanford Binet, but he was given several Wechsler IQ tests for adults (WAIS-III). The jury that originally recommended capital punishment had been presented with Hall’s multiple IQ scores above 70 and testimony describing severe adaptive disability as mitigating factors. In the adaptive behavior testimony, Hall was characterized by slow speech, slow learning, making simple mistakes, poor school performance, mental abilities of a preschooler and being physically and mentally abused by his unsympathetic mother. The prosecution countered with a list of seven aggravating factors that the jury and judge found to substantially outweigh the possibility of intellectual disability. The sentencing court was skeptical about the evidence of intellectual disability, stating that the expert testimony failed to explain how an individual with Hall’s alleged serious adaptive behavior disability “could formulate a plan whereby a car was stolen and a convenience store was robbed.” The court also stated that even assuming the accuracy of the expert testimony, Hall’s disabilities could “not be used to justify, excuse or extenuate [his] moral culpability [in this case]” (p. 1992-93).

After the Atkins case was decided, Hall’s sentence was reviewed and reaffirmed. Given IQ scores that did not meet the intellectual disability threshold of 70, the Florida court saw no reason to continue the inquiry into factors related to Hall’s adaptive disability. The Florida court also rejected Hall’s claim that his IQ should be read as a range of scores from 67 to 75 (apparently a 95% confidence interval around Hall’s lowest IQ score of 71 admitted by the trial court). However, the resentencing court found Hall intellectually disabled as a mitigating factor and gave it “unquantifiable” weight.

Challenging the IQ Standard for Intellectual Disability

Hall’s defense team presented two reasons why the threshold score of 70 was unfair. First, they argued that it did not take into account the inaccuracy and unreliability of IQ scores. They argued that any IQ score falling within two standard errors of measurement (SEM) of the cut score should qualify Hall as meeting the standard. Using evidence that standardized ability tests generally had SEMs of about 2.5, they argued that an IQ score between 70 and 75 should be considered to meet the first conjunctive criterion and that the extensive witness testimony describing Hall’s adaptive disability prior to age 18 should then be sufficient to satisfy the second and third criteria and qualify Hall as having an intellectual disability under the statutory definition. However, the Florida court rejected this argument, observing that the statute did not use the word approximate, nor did it reference the standard error of measurement (SEM) of IQ scores.

The second reason given for unfairness of the threshold standard centered on an argument suggesting a compensatory rather than a conjunctive standard. That is, the defense argued that when there is overwhelming evidence of adaptive disability, that evidence should compensate for an IQ score slightly above the cutoff score. This argument focused on the premise that intellectual disability should not be determined by a single IQ score. However, the Florida statute did address the issue of possible misclassification based on a single IQ score by specifically permitting consideration of IQ scores from multiple retests. At Hall’s trial, the court had admitted into evidence four IQ scores of 71, 72, 73 and 80 (average IQ score of 74). Hall claimed two additional IQ test scores below 70, which the court refused to admit stating that there was insufficient evidence of their validity. One of those scores, an IQ of 69, was contained in the report of a deceased doctor who had not provided sufficient documentation of its derivation and was obviously unavailable for cross examination. The Court did not consider the issue of how multiple IQ scores should be combined to address the criterion of subaverage functioning, indicating only that it was “complicated.”

Apparent Contradiction

In 1991, the sentencing judge who imposed capital punishment found Hall intellectually disabled, but in 2010 the same judge found Hall not intellectually disabled. The reason for the apparent contradiction was that the standards for intellectual disability were based on different purposes in the two instances. In 1991, intellectual disability was a mitigating factor considered by the judge alongside overwhelming evidence of aggravating factors. But in 2010, the judge was applying the Florida statutory definition of intellectual disability under the Atkins framework that had not existed in 1991. Additionally, the prosecution had not contested Hall’s evidence of intellectual disability in 1991 because conceding it would not change its minimal weight

---

3 Hall was also convicted of murdering the officer but his sentence of capital punishment was reduced to life imprisonment due to lack of sufficient evidence of premeditation.
relative to the seven strong aggravating factors presented and because a defendant’s reliance on mitigation based on intellectual disability can backfire and cause the jury to give more weight to the aggravating factor of future dangerousness. Thus, determination of intellectual disability in 1991 was not adversarial and the prosecution had little incentive to challenge the evidence presented by the defense. But in 2010, the prosecution did challenge the evidence of intellectual disability because it was now a bar to capital punishment.

Although in 1991 the prosecution did not challenge Hall’s claim of intellectual disability, the sentencing judge provided an extensive discussion of his doubts about whether Hall’s abilities and behaviors were consistent with those expected of a person with an intellectual disability. In particular, the court was skeptical of the expert testimony offered by the defense. In addition to questioning how an intellectually and adaptively disabled individual could plan to steal a car and commit a robbery, the court also questioned whether other behaviors exhibited by Hall were consistent with a diagnosis of intellectual disability. For example, Hall had driven the stolen car with the victim inside on public streets in broad daylight through the city for eighteen miles to the woods outside the city. He had also lived a relatively normal life, been employed and avoided any run-ins with the law during the prior five years he had been on parole for his first rape conviction. In the court’s opinion, the defendant’s deliberation and planning behaviors preceding and during the commission of the crimes were inconsistent with the clinical characterization of the defense experts that Hall was severely intellectually disabled. Consequently, the sentencing court had discounted the expert testimony as somewhat exaggerated.

The Majority Opinion

In a 5-4 decision, the U.S. Supreme Court held that Florida’s interpretation of its intellectual disability statute violated the Eighth Amendment and was unconstitutional. Since the Atkins court had determined that capital punishment is not permitted for individuals with intellectual disabilities but had not specified a standard for determining who is intellectually disabled, the Court had to determine the appropriate standard. The Court credited expert testimony from professional organizations, including the APA, that supported consideration of the SEM and a compensatory standard. The Court held that Florida’s rule disregards established medical practice in two interrelated ways. It takes an IQ score as final and conclusive evidence of a defendant’s intellectual capacity, when experts in the field would consider other evidence. It also relies on a purportedly scientific measurement of the defendant’s abilities, his IQ score, while refusing to recognize that the score is, on its own terms, imprecise (p. 1996).

In articulating the appropriate standard for determining intellectual disability, the Court stated

For professionals to diagnose – and for the law then to determine – whether an intellectual disability exists once the SEM applies and the individual’s IQ score is 75 or below, the inquiry would consider factors indicating whether the person had deficits in adaptive functioning [including] evidence of past performance, environment, and upbringing. … The rejection of the strict 70 cutoff in the vast majority of States and the ‘consistency in the trend’ toward recognizing the SEM provide strong evidence of consensus that our society does not regard this strict cutoff as proper or humane. … Freddie Lee Hall may or may not be intellectually disabled, but the law requires that he have the opportunity to present evidence of his intellectual disability, including deficits in adaptive functioning over his lifetime (p. 1995-96, 1998, 2001).

The Court remanded the case back to the Florida court for further proceedings to determine whether Hall is intellectually disabled using the Court’s newly stated criteria.

The Dissent

The dissenting justices took issue with the majority basing its holding “largely on the positions adopted by private professional associations” in contrast to its holding in Atkins that “the clearest and most reliable objective evidence of contemporary values [used to define appropriate standards] is the legislation enacted by the country’s legislatures” (p. 2002). They accused the majority of applying standards based on the view of professional societies rather than of American society as a whole. The dissenters also objected to the use of a compensatory model allowing defendants to use adaptive disability evidence to bolster IQ scores between 70 and 75 (approximately 2 SEMs from the cutoff). They favored Florida’s conjunctive model requiring a defendant to show both a qualifying IQ score and evidence of adaptive disability. They also questioned the majority’s assertion of a national consensus noting that of states with capital punishment, 10 (including Florida) do not require consideration of the SEM, 12 consider the SEM and 9 have taken no position on this matter. They also note that the APA position is new and that there is some precedent for professional associations to sometimes rescind changes. For example, they cited the American Association on Intellectual and Developmental Disabilities (AAIDD) which changed its definition of intellectual disability in 1992 from an IQ score of 70 or below to an IQ score of approximately 70-75 or below. In 2002, with only 4 states using the new definition, AAIDD changed back to the original definition. If the views of associations differ, the dissenters wondered how the Court would decide which should prevail.
Defending their position that multiple IQ scores above the generally accepted cutoff of 70 for intellectual disability provide more objective and relevant evidence of individuals’ abilities to assume personal responsibility for their actions, the dissenters stated

[In a capital case], intellectual functioning is important because of its correlation with the ability to understand the gravity of the crime and the purpose of the penalty, as well as the ability to resist a momentary impulse or the influence of others. … By contrast, in determining eligibility for social services, adaptive functioning may be much more important (p. 2006).

Finally, the dissenters argued that the SEM was applied inappropriately to Hall’s data. They noted that SEMs differ by test and age and that the majority had used a value more than twice that estimated for the IQ test administered to Hall. Recalling that the defendant is required to prove intellectual disability by a preponderance of the evidence (>50%), and using Hall’s average IQ score of 74, an SEM of 2, and a 95% confidence interval, the dissenters calculated that the probability of Hall’s true IQ being above 70 would be roughly 97.5%, more than enough to satisfy the evidentiary standard. They observed that using the majority’s rule, a defendant could prove subaverage intelligence when a 1 or 2 SEM confidence interval barely included a score of 70. According to their calculations, this could result in a lower bound probability of a true IQ below 70 of roughly 17% or 2.5%, respectively. They also criticized the majority for not explaining “why its criticisms of the uncertainty resulting from the use of a single IQ score apply when a defendant consistently scores above 70 on multiple tests. … [T]he well-accepted view is that multiple consistent scores establish a much higher degree of confidence” (p. 2011, emphasis in original).

The dissenters summarized their views as follows

There is therefore no excuse for mechanically imposing standards that are unhinged from legal logic and that override valid state laws establishing burdens of proof. The appropriate confidence level is ultimately a judgment best left to legislatures …” (p. 2011).

Conclusion

In addition to Florida, the majority opinion in the Hall case casts doubt on the validity of intellectual disability statutes in Arizona, Kansas, Kentucky, North Carolina, Virginia, and Washington. From the language of the opinion, it appears that states using the 3-prong test for intellectual disability must construct a confidence interval around the defendant’s IQ score and allow testimony regarding adaptive disability to establish intellectual disability when the confidence interval contains the threshold value. The purpose of this rule is to ensure that no individual who may actually be intellectually disabled will receive capital punishment. The Court considers deprivation of life very seriously and apparently believes that erroneously finding a defendant not intellectually disabled is a far more serious error than the possibility of erroneously qualifying a defendant as intellectually disabled who is not. It will be interesting to see how the various states implement the Court’s new criteria and whether greater consensus regarding IQ score standards is actually achieved. Perhaps like Ohio, Florida will decide to consider IQ scores slightly above 70 (e.g., in the 70-75 range) as rebuttable presumptions of lack of intellectual disability that can be overcome by strong evidence of adaptive behavioral disabilities manifesting prior to adulthood. On the other hand, Florida could adopt a different criterion based on multiple retest IQ scores. In any case, because the Hall decision was rendered by the U.S. Supreme Court, the statutory interpretation of intellectual disability in all states with capital punishment must conform to its mandates. Could a convicted felon with an IQ score of 76 (one point from the 2 SEM range) and a history of adaptive behavioral disabilities be the next case to challenge the criteria for barring capital punishment of offenders with intellectual disabilities?

ANNUAL MEETING COMMITTEE UPDATE

Terry Ackerman, University of North Carolina – Greensboro

We recently completed a survey the NCME membership seeking their opinions on holding our annual meeting jointly with AERA. This information has been shared with the Board of Directors and will be useful as we begin to negotiate with AERA for future contracts. Note, our current contract with AERA runs through 2016. Complete results of the survey follow this brief update.

The Annual Meeting Committee continues to meet via conference call once a month. I updated the Board of Directors about our activities at their July meeting in Madison, WI. Our current focus is the creation of an Annual Meeting Handbook that will provide information on each task/event associated with the NCME Annual Meeting. Currently we have identified 24 such events. Information that will be recorded includes timelines, individual(s) or committees responsible, budgets, history, means of evaluation, and previous recommendations for improvement. We are working closely with Leah Knopke, a Project Manager from The Rees Group (TRG) on the Handbook. This handbook will be extremely helpful for newly elected NCME officers, members of the Board of Directors, as well as
TRG. It should also be noted that Jane Shepherd, our Meeting Planner from TRG also sits on our committee and keeps us updated on the timelines and preparation for our April 15-19, 2015, meeting in Chicago.

I will continue to share the activities of this committee with the membership in the Newsletter and strongly encourage anyone who has suggestions for this committee to email me at taackerm@uncg.edu.
### When do you prefer the NCME annual meeting be held?

<table>
<thead>
<tr>
<th>Answer Options</th>
<th>Response Percent</th>
<th>Response Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>During the week (Not including weekend days)</td>
<td>63.9%</td>
<td>299</td>
</tr>
<tr>
<td>Over the weekend (Including weekend days)</td>
<td>36.1%</td>
<td>169</td>
</tr>
</tbody>
</table>

Answered question 468

Skipped question 28

### If NCME and AERA held separate annual meetings (i.e., at different times), which would you most likely attend?

<table>
<thead>
<tr>
<th>Answer Options</th>
<th>Response Percent</th>
<th>Response Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCME only</td>
<td>63.1%</td>
<td>303</td>
</tr>
<tr>
<td>AERA only</td>
<td>14.2%</td>
<td>68</td>
</tr>
<tr>
<td>Both NCME and AERA</td>
<td>20.0%</td>
<td>96</td>
</tr>
<tr>
<td>Neither</td>
<td>2.7%</td>
<td>13</td>
</tr>
</tbody>
</table>

Answered question 480

Skipped question 16

### If NCME scheduled its annual meeting independent of AERA, when would you prefer the NCME meeting be held? (Assume the AERA meeting will be...)

<table>
<thead>
<tr>
<th>Answer Options</th>
<th>Response Percent</th>
<th>Response Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall (Sept-Nov)</td>
<td>34.6%</td>
<td>168</td>
</tr>
<tr>
<td>Winter (Dec-Feb)</td>
<td>15.6%</td>
<td>76</td>
</tr>
<tr>
<td>Spring (Mar-May)</td>
<td>31.3%</td>
<td>152</td>
</tr>
<tr>
<td>Summer (Jun-Aug)</td>
<td>22.6%</td>
<td>110</td>
</tr>
<tr>
<td>No preference</td>
<td>22.4%</td>
<td>109</td>
</tr>
</tbody>
</table>

Answered question 486

Skipped question 10

### How many conferences do you receive funds to attend annually?

<table>
<thead>
<tr>
<th>Answer Options</th>
<th>Response Percent</th>
<th>Response Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>13.6%</td>
<td>66</td>
</tr>
<tr>
<td>Some funding for at least one</td>
<td>28.1%</td>
<td>137</td>
</tr>
<tr>
<td>Full funding for one</td>
<td>29.2%</td>
<td>142</td>
</tr>
<tr>
<td>Full funding for two</td>
<td>29.2%</td>
<td>142</td>
</tr>
</tbody>
</table>

Answered question 487

Skipped question 9
Note, other professional organizations or societies that members listed included:
AAMC, ACA, AEA, APS, ASA, ASCD, ATD ATP, CCSSO, CEC, CLES, CLEAR, EARLI, IACAT, ICE, IMEKO, IOMW, ITC, LTRC, MALT, NARSI, NATD, PROMSIMEKO, RSS, SIOP, SMEP, SREE, and TESOL.

NCME FITNESS WALK/RUN CORNER
Brian French & Jill van den Heuvel for the NCME Fitness Walk/Run

We hope your summer has been outstanding! We know all of you are looking forward to the beautiful fall walking and running weather. We are in the planning stages for the event in Chicago and in search for a route along the waterfront. We think the group may be in for a windy run as we have had in the past. Keep moving and enjoy the remaining fall weather!
2015 NCME ANNUAL CONFERENCE UPDATE:
EXPANDING INNOVATION & IMPACT
Ye Tong & Jennifer Randall, conference co-chairs
Caroline Wiley, training chair

We received over 700 proposals for individual presentations, coordinated sessions, GSIC, and training sessions. We are very thankful to our expert review panels currently engaged in reviewing proposals to ensure a well-balanced and engaging program this year.

Panels include 3-6 committee members with extensive expertise in one of twelve topic areas. Each proposal will be read by at least three experts who rate each submission on three criteria: (1) relevance to practitioners, researchers, and/or graduate students, (2) articulation of research problem/conceptual framework supported, and (3) technical adequacy. Members should expect to receive decision letters near the end of October.

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

Newsletter Advisory Board

LANIE BRADSHAW, University of Georgia
JOHN DENBLEYKER, National Board of Osteopathic Medical Examiners
ELLEN FORTE, edCount
JOANNA GORIN, Arizona State University
SARA HENNINGS, MetaMetrics, Inc.
JOAN HERMAN, CRESST/UCLA
ANDREW HO, Harvard Graduate School of Education

THEL KOCHER, Walden University
GERRY MELICAN, The College Board
DIANNE TALLEY, University of North Carolina, Chapel Hill
TIANOS PATELIS, The College Board
S.E. PHILLIPS, Consultant
CHRISTINA SCHNEIDER, CTB
XIAN (BO) WANG, The College Board

SUSAN DAVIS-BECKER, Editor, Alpine Testing Solutions
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

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.