NCME

national council on measurement in education

VOLUME 24, NUMBER 2 | NEWSLETTER | June 2016

FROM THE PRESIDENT

Mark Wilson, University of California, Berkeley, & University of Melbourne

It is very humbling to be elected as president by one's colleagues, to take on the role of guardian of our joint endeavors, and to seek to lead us through a new year. I am grateful for this honor, and will attempt to carry out this role as well as I can.

First, let us acknowledge the wonderful work done for the 2016 annual conference in Washington. Although I mainly spent my time in Washington in committee meetings, I did at times manage to escape from them, and people seemed to be having a good time learning and socializing. To give it a quick overview, here are some stats: There were 1,142 attendees with 594 jointly attending the AERA meeting as well. Forty-nine percent of the attendees downloaded the mobile app, and there were over 800 user contributions to the platform. The program featured five invited sessions, 40 coordinated sessions, 187 individual paper sessions, and two electronic board sessions with 25 presentations in each. Many members and staff put in long hours over the preceding year to make it such a great success—I cannot acknowledge you all in this short space, but do want to acknowledge the great work done by Past-President Rich Patz, by Program Chairs Andrew Ho and Matt Johnson, Training and Development Committee Chair Xin Li, and by the staff of our new management company, Fernley & Fernley, especially by Joe Casey, our able and devoted executive director.

We are still waiting on getting responses to the member survey about the annual meeting, so please respond if you have not done so already. The survey link was emailed to NCME members several times since the meeting.

Second, we are currently planning on an exciting annual meeting in San Antonio next year, and at the forefront of those plans are our new program chairs. Hence, I would like to introduce Lydia Liu from Educational Testing Service (lliu@ets.org), and Billy Skorupski from the University of Kansas (wps@ku.edu) as co-program chairs, and Sun-Joo Cho from Vanderbilt University (sj.cho@vanderbilt.edu) as the Training and Development Committee chair. They have been busy planning the call for proposals, investigating invited sessions, etc. If you have any suggestions, or questions, please contact me at MarkW@berkeley.edu, or the chairs at their email addresses shown above.

Third, I plan to take some space in each of my four president's messages in the newsletter to discussing an NCME initiative that we are starting up this year. We are planning on taking a special focus on classroom assessment during the coming year, and I am delighted to announce that two Board members, Kristen Huff and Dale Whittington, have kindly agreed to be Board leaders of this effort. I will be reporting on upcoming aspects of this in my next few messages. In the meantime, here are some initial thoughts, focused on why this should be an important topic for us in NCME.

The Importance of Classroom Assessment

As professionals involved in measurement in education, we mainly focus on the development and use of various forms of tests and assessments. Much of the funding that is used to develop such tests comes, ultimately, from the decisions of policy makers in education, principally leaders of state and federal bodies charged with administering education, but also administrators at other levels of education, not to mention the many boards, committees, and organizations that also seek to influence education. This leads inevitably to the major focus of our work being from a top-down perspective, at the behest, and for the purposes of, such policy makers and administrators. This is entirely appropriate, as these are the persons and bodies that are crucial large-scale decision makers in our educational endeavors.

However, there is an alternate way to perceive the educational enterprise—from the bottom up. From this perspective, educational measurement's core activity is the help in the educational progress of each student as they learn. And the agents immediately involved in that are the student and the teacher. From this perspective, the most salient moments in student learning are being orchestrated by a teacher-student pair, and that is where one might expect that the most important decisions about student learning will be made, and where the greatest impact takes place.

Before pursuing this further, it is important to define some terms. Although I have used classroom assessment above, in fact in the literature it is most often referred to now as *formative* assessment. The definition adopted here is as follows:

> An assessment activity is formative if it can help learning by providing information to be used as **feedback**, by teachers, and by their students, in assessing themselves and each other, to modify the teaching and learning activities in which they are engaged. (Black, Wilson, & Yao, 2011)

It is important to note that it is not the material itself of the assessment that is crucial, but instead it is the use to which it is put. A similar definition can then be made for *summative* assessment:

> An assessment activity is summative insofar as it is being used to **provide a summary** of what a student knows, understands, or can do, and not to help by providing feedback to modify the teaching and learning activities in which the student is engaged.

This summary could be used for decisions such as promotion, attainment of a set of standards, graduation, or admission to the next level of schooling. Again, note that it is the use of the assessment that is the distinguishing feature here—in fact, a specific assessment could be used in both ways under different circumstances.

The observation above about the salience of formative assessment in education is not just an idle observation, but has been well-established by a number of very broad research syntheses (Crooks, 1988; Natriello, 1987), reaching a peak in the seminal work by Black and Wiliam (1998a, 1998b), which detailed multiple types of *formative assessment* that have been found effective and multiple modes of feedback that affect student learning. They found effect sizes (ES) for formative assessment between 0.4 and 0.7, and the establishment of similar results continues to this day. For example, working in the area of writing assessment, Graham, Hebert and Harris (2015) found that, overall, formative assessments gave an average standardized ES of .63, with results for specific forms of feedback including an ES of .87 for adult feedback (including teacher feedback), .58 for peer feedback, .62 for self-assessment, and .38 for computer feedback.

Clearly, then, the topic of classroom assessment is indeed crucial for the entire educational enterprise and should be seen as the most likely pathway for educational measurement to make a positive and central contribution to education.

References

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7–74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–147.

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives*, *9*, 71–123.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438–481.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: meta-analysis. *The Elementary School Journal, 115*(4), 523–547.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 22*(2), 155–175.

FROM THE EDITOR

Heather Buzick, Educational Testing Service

Happy summer! Our new president, Mark Wilson, has written his first column discussing both NCME business and a new NCME initiative focused on classroom assessment. In this issue we also have several highlights and details from our recent annual meeting, as well as recurring content and committee updates. Our spotlight member is Luz Bay. I would also like to introduce you to our new legal corner columnist, Michelle Croft, who holds a Ph.D. in educational measurement and a J.D. degree. I greatly appreciate the contributions from our previous columnist, Susan Phillips, both during and prior to my tenure as newsletter editor, and wish her the best.

This year's program chairs, Lydia Liu and Billy Skorupski have selected a theme for the 2017 annual meeting in San Antonio, TX. See brief information about the annual meeting and call for proposals on p. 24 of this newsletter.

We've added a suggested citation format on the last page of this newsletter for contributors who wish to cite their columns. I encourage you to send me articles, suggestions for content, and feedback on this or previous issues.

GRADUATE STUDENT CORNER: PROFILING EARLY CAREER INTERNATIONAL PROFESSIONALS

Tiago Caliço, University of Maryland—College Park

Following my personal take contrasting the nature and practices of educational assessment in Portugal and the United States in my previous column, I am keeping with the topic of contrasts. This time, however, I am focusing on two parallel, yet opposite, careers. I interviewed two recent graduates who have in common their training in the United States and the fact that they will put that training to use outside the United States.

Dr. Diego Luna-Bazaldua graduated from Teachers College, Columbia University's Measurement and Evaluation program in December of 2015. He is currently a post-doctoral researcher at the Lynch School of Education, Boston College. Dr. Dan McNeish also graduated last December, from the University of Maryland's Measurement, Statistics, and Evaluation program. He is now a tenure-track assistant professor at Utrecht University, Netherlands.

Dr. Diego Luna-Bazaldua

1. Diego, you grew up in Mexico City and decided to work toward earning your Ph.D. in measurement and evaluation at Teachers College, Columbia University. Can you tell us a little bit about your background before graduate school in the United States?

My background is in both psychology and statistics. Unlike the liberal arts undergraduate education in the United States, in Mexico you have to enroll in a specific career at the undergraduate level. So by the time I finished high school, I decided to study psychology as my undergraduate program. During my undergraduate education, I had the chance to take several courses in statistics and test theory and to assist in research applying quantitative techniques in social psychology. After I finished my undergrad, I completed a graduate program in applied statistics in Mexico as well, which gave me a strong background in the mathematical concepts behind probability and statistical inference. By the time I finished my studies in statistics, I got a job in a university assessment department, where I had the opportunity to apply my knowledge in both psychology and statistics in the development and analysis of standardized tests.

2. So how did you decide that you wanted to pursue a Ph.D. in this field, and how did you go about choosing a program?

Educational assessment and psychological measurement are relatively new in Mexican academia compared to all the development these fields have had in the United States; therefore, there are very few universities in Mexico offering master-level programs in educational assessment, and no university has a formal doctoral track in psychometrics. Some doctoral programs offer some courses in psychological measurement, but not in the comprehensive manner and training students are exposed to in programs in the United States. Most psychometricians in Mexico have to go abroad to the United States or Europe to get trained in this field, and later they return to Mexico to work in universities, government, or testing companies. Therefore, I knew that if I wanted further specialization in psychometrics and assessment, my best choice was to come to the United States.

I chose the program at Teachers College among the programs I got accepted because they had a scholarship agreement with the Mexican National Council for Science, so there were financial resources available from the Mexican government to fund my graduate studies.

3. What are your research interests? And how do you see your career in the medium to long term? More industry or academia oriented? How did your training influence your preferences and career plans?

I started my Ph.D. interested in psychometric models for standardized testing, under both classical test theory and IRT. Throughout my doctoral education I learned a lot about very different models used in specific situations, and I have become particularly interested in very different areas: cognitive diagnostic modeling, data mining, and more recently in models for causal inference for program evaluation. I know these are somewhat different areas, and some psychometricians may not see a connection among them, but as a good friend and colleague told me once: the point of learning about all these different models is to have a big tool kit, so we psychometricians are able to use each tool to solve the specific problem at hand.

Becoming interested and familiar with several different fields has also helped me to advance in my own research. For instance, part of the work I did for my dissertation has helped me to come up with new ideas for educational data mining. At the same time, my work at Boston College has made me realize there are some gaps on research about the application of data mining models for the estimation of treatment causal effects in nonrandomized trials. In the end, we just need to keep learning more and more!

I think that my future work will involve both academia—since I have a passion to learn, to teach, and to train people—but I also see myself being a consultant for either government or industry. I know some people see them as somewhat dissociated areas, but the most expert psychometricians I have met do both things. Academia gives you the knowledge on the state of the research in the field, but consultancy provides you with experiences that are difficult to obtain or teach in the classroom.

4. What struck you as the main differences between the educational system in Mexico and the United States in what relates to assessment?
Well, there are many differences. As I said before, the state of psychometrics and educational assessment in Mexico is relatively young compared to the United States. Yet, Mexico is doing a lot of work on these areas with the relatively small number of experts there. For instance, federal reforms in the educational laws in 2012 have fostered the role of our National Institute for Educational Evaluation in the educational policies and assessment initiatives nationwide. This national institute has promoted a robust agenda that will involve more research, more data, and the training of even more people in these fields in the future years.

Still, Mexico should and can learn more about the best practices in educational assessment from the United States and other regions of the world, while keeping in mind our social and cultural context in the implementation of any policies resembling ideas from abroad.

5. Apart the academic training and hands-on experience, do you have any big lessons from your experiences in the United States that you think can have a high impact in assessment practices in Mexico?
I think an interesting case can be made about the admission to universities in Mexico, particularly at the undergraduate level. In the United States, application to universities involves a process that requires several pieces of evidence for admission: test scores, GPA, statements, and so on. Because of this admission approach, students from diverse backgrounds may have similar opportunities to get accepted into a good university no matter if their tests scores are somewhat different.

In the case of Mexico, each university has its own admission tests, which in many cases do not overlap in domains of content. So while a student in the United States knows that the SAT covers specific domains no matter what universities the student intends to apply, students in Mexico have to prepare for every test depending on the universities they anticipate applying to. In addition to that, while some universities may have a comprehensive admission system based on several pieces of evidence, others take only into account their own standardized tests scores and high school GPA as the criteria to select among applicants; the latter case is particularly true for the biggest universities, which get tens of thousands of applications every year but have a relatively small number of available spots. As a result, I believe that many students from disadvantaged backgrounds may have a smaller chance to get into the best public universities since the admission criteria does not take into account the diversity and background of the applicants. This is a situation that worries me a lot, since access to higher education is a key factor for social mobility in Mexico, and the current admissions system keeps perpetuating inequality.

6. Finally, what do you recommend to a Mexican, or any foreign student, who is contemplating pursuing training in psychometrics in the United States?
First of all, in my case it made a huge difference to have a good background in probability and statistics. I met some students in graduate school who came from the social sciences, and they encountered many of the psychometrics and statistics courses as being very challenging. On the other hand, doctoral students with a background in math found the courses easier and were able to spend a big part of their time doing research.

Second of all, once accepted, be sure to have full funding from either the university or from government agencies (CONACyT has agreements with several universities in the United States). If the university you get accepted does not provide you with good funding for your doctoral education, apply to other institutions because it is not worth it to compromise the rest of your life with debt because of 4 or 5 years of graduate education. Most graduate programs in psychometrics in the United States offer good funding for their doctoral students, and I am sure that faculty in more than one graduate program would be happy to enrich their student body with students from Mexico and the rest of Latin America.

Dr. Dan McNeish

1. Dan, you grew up in New England and Michigan. After college you decided to get your Ph.D. in psychometrics at the University of Maryland. Can you tell us a little bit about your background before grad school in the United States?
I started college at Oakland University in Michigan, where I was trying to decide between focusing on math or psychology. As with most psychology programs, there was a statistics requirement, which is where I was first exposed to psychometrics. I realized that psychometrics would be a convenient way to blend the two fields I was considering, so I transferred to Wesleyan University in Connecticut because one of the faculty members there was a psychometrician and taught a few undergraduate psychometric courses. I worked in the Psychometric Lab on campus for 2 years and the work I did there confirmed that I had chosen the right field. I think that I am a bit of an anomaly in the field in that I knew that I wanted to pursue psychometrics fairly early in my undergraduate career. Most people tend to have interesting stories about how they stumbled into the field, but my path was pretty calculated. I had asked my advisor at Wesleyan which programs he thought would be the best for my interests and that were preferably in the eastern United States. Maryland was on top of the list he gave me. Maryland had admitted me and gave me funding, so it was not a very difficult decision to make.

2. What are your research interests? And how do you see your career in the medium to long term? I understand it has been a long dream of yours to be a college professor. Anything else you see yourself doing?
My interests changed 180 degrees from when I started graduate school up until now. When I first started, I was sold on studying IRT and wanted to work in industry. Now, I work in academia, mostly on multilevel and longitudinal models with a focus on small sample problems and adapting methods from other fields in behavioral sciences. I am not quite sure where I see myself in the medium or long term. I suppose I'll just keep trying to crank out papers like it's going out of style until I run out of ideas and hope I have tenure when that happens!

3. Once you graduated, you got a job as a tenure track professor at the Department of Methodology and Statistics at Utrecht University. This is rather unusual, as almost every Ph.D. in Europe wants to land a job in the United States. Can you tell us a little bit about your job search approach as you were getting close to graduation and how you decided to work abroad?
I never had seriously considered searching outside the United States and I don't have any connections to Europe and had actually never been on the continent before I interviewed at Utrecht University. For me personally, I was looking for a department where the interests of the faculty matched closely to my own and several professors at Utrecht, like Rens van de Schoot, Mirjam Moerbeek, and Joop Hox, do a lot of small sample work. I am not very good at making decisions, so my approach was just to take the first good offer I got. I went on one other interview but the offer from Utrecht was a good one so I took it without sending out other applications (the hiring timeframe in the Netherlands precedes the United States timeframe, so I accepted the offer before most United States jobs were posted).

4. What struck you as the main differences between the educational system in the United States and the Netherlands in what relates to assessment? How about regarding working as an academic?
A major difference between the United States and the Netherlands is that Ph.D. candidates are employees of the university, rather than students. This is not the norm in the United States. This means that Ph.D. candidates in the Netherlands don't pay tuition but rather are paid a salary without having to do things like assistantships or TAs. Being a Ph.D. candidate is considered a full-time job in and of itself. At least in the Netherlands, the tenure system is also much different. Tenure and promotion are very distinct concepts and I believe that the concept of getting tenure is not as stressful or as high stakes as it is the United States system.

5. What benefits do you think graduate programs in psychometrics in the United States could get from the Dutch approach?
In the Dutch system, classes are sometimes team taught by people with various specialties rather than by a single instructor as in the United States. So if one instructor is not necessarily the most knowledge about certain topic within a particular class, another instructor who may be more knowledgeable might teach those topics instead. This doesn't happen all the time, but I don't think I have ever heard of this happening in the United States as part of a formal process. Although it can present some difficulties when multiple instructors are responsible for course grades; it seems to be generally advantageous for students because they receive instruction from those in the department who know the most about particular topics.

6. Finally, would you recommend recent graduates to expand their job search beyond the United States job market? What advice would you give them?
I think it depends a lot on the person because it is, of course, not easy to pick up and move to a different continent. Having said that, I do think that it is an option that might warrant some consideration because many universities in places like the Netherlands, Germany, and the Nordic countries work almost exclusively in English and offer many courses in English. Some universities are also making a push to internationalize, which can give some United States candidates an edge in the hiring process. I was also surprised that a flight from Washington, DC to Amsterdam is actually pretty close to the duration of a flight from Washington, DC to Los Angeles, so the physical distance is not as great as I would have guessed.

Conclusion

The two young researchers I interviewed for this article have in common the deliberate way in which they planned and executed their undergraduate and graduate career. They also benefited from the increasingly rare privilege of having funding provided by their institutions. Interestingly, both diversified and changed their interests during their graduate programs, a possible sign of how programs in this field diversify their faculty and course offerings. Another important takeaway of both Dan and Diego's experiences is how the fields of psychometrics and educational measurement are understood outside the United States. Mexico still has an inchoative, yet growing, academic and private market for psychometricians. Central and northern Europe have more of an established market, and their attractiveness to those trained in the United States is growing. Diego and Dan's experiences are a good reminder that no matter how deliberate one's academic and professional path may be, interests and opportunities are ever changing. In a globalized world, with tight integration of social, legal, and scientific values, a well-trained psychometrician has a global job market to explore, if he or she is willing and able to recognize and adapt to every market's cultural specificities.

Author note: Tiago Caliço is a Ph.D. candidate at the Department of Human Development and Quantitative Methodology at the University of Maryland—College Park. His interests include games-based assessment, simulations, psychometric modeling of process data, evidence-centered design, and Bayesian networks.

SPOTLIGHT ON THE PEOPLE WHO MAKE OUR ORGANIZATION GREAT

Luz Bay, The College Board

How did you get into the field?
Accidentally. Both my bachelor's and master's degrees are in pure math. Feeling that I did not have any employable skills after getting my master's, I started in a doctoral program—still in pure math, the only field I was familiar with. One year in the doctoral program, I decided that I'd had enough of proving theorems day in and day out. I looked into other fields like actuarial science and statistics. I was accepted in a couple of universities with good doctoral programs in statistics but neither was offering assistantships to first-year students. Then I recalled that I took a linear regression class in the educational psychology department

as part of the requirements of the master's program—when you have to take a class outside your department. I applied to the Ph.D. program in educational statistics and was accepted. Interestingly, my assistantship was teaching an MCAT review class in math. I did not realize until I was taking my first measurement class that the program I was in was educational statistics and measurement. I did not realize until then that there was a field called educational measurement. Topics like correction for guessing, cheating detection, and differential item functioning (DIF) really caught my interest. I ended up doing my dissertation in cheating detection.

There were not a lot of measurement classes offered in my program so I took advantage of available internship programs. During my internship at ETS, William H. Angoff was still running the program. I asked him to read my paper on cheating detection, and he gave me handwritten feedback. I also met with Paul Holland about a DIF idea that I was thinking was going to be my dissertation topic. His question "Why would you do that?" made me not ever look back to that topic. During my internship at ACT, my mentor, Deborah Harris, encouraged me to work on my dissertation. I was fortunate enough to get guidance from Brad Hanson, who ended up being on my dissertation committee. Mike Kolen was also kind enough to have a weekly session with me to teach me equating. This was during the time that he and Bob Brennan were writing their first equating book.

If you weren't in this field, what would you do?
Had I not stumbled into this field I would probably have ended up teaching math. That would have been okay because I wanted to be a teacher when I was 7 years old. That only lasted about a couple of days because my siblings pointed out to me that teachers don't tend to earn big salaries. Then I wanted to be a lawyer. That lasted a few years. I went to a science high school on a scholarship that required me to pursue a science or math degree in college. Then I was going to be a doctor, until my college scholarship was only allowing me to major in chemistry, marine biology, statistics, or math. I still have doctor envy.

I think I can still be a talk show host. A former colleague emailed me about The Measurement Channel (TCM). He thinks that measurement is just as important as politics so people should be talking about on it TV all the time, just like politics. Live streaming happens to be one of my hobbies, so I asked him to pitch me the idea in a meeting that I will stream live. He has so far refused. I think it points to the primary reason why people do not talk about measurement a lot—people who think it is important to talk about measurement do not want to talk about it.

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

1. Study hard. Work hard. Play hard.
2. Network, network, network.
3. Cultivate friendships.
4. Look ahead.
5. Be interested in what you're doing by either working on something that you are genuinely interested in, or developing genuine interest in whatever it is that you have to work on.

What do you like to do for fun outside of work?
I run. I have a shirt that says "I'm slower than a turtle stampeding through peanut butter, but I run." I started running about 5 years ago in a Run for God program in my church. It's very similar to a couch-to-5K program. Since then, I have run almost 70 races for a total of almost 300 miles, including the 2015 Honolulu Marathon. I am currently training for the 2016 New York City Marathon, which I will be running in November.

I also garden. I have a very large vegetable garden and a number of perennial flower gardens. I also started raising chickens this year.

What would you say has been one of the biggest innovations in psychometrics in the last decade or two?
I don't think there are any real big innovations in the last two decades. I think technology has enhanced our understanding and broadened our use of existing innovations.

When you go to conferences, how do you pick what sessions to attend?
I always go to the sessions where I'm presenting. ☺ I mostly go to invited sessions, keynote speeches, and career award winner presentations. I also go to sessions when someone I would like to see or meet is presenting. If there is a

conference paper that I'm particularly interested in, I try to get a copy of the paper instead of attend the presentation. I pick sessions based on networking opportunities.

Who has been a significant influence in your professional life?
My first job after graduate school was assistant director for the National Assessment of Educational Progress (NAEP) achievement level setting project at ACT. The client was the National Assessment Governing Board (NAGB) and the contract required a Technical Advisory Committee on Standard Setting (TACSS). I was a very impressionable youngster in the field and was very star struck with TACSS members that I am sure I was influenced by them. That august group included Bob Brennan, Mark Reckase, Michael Kane, Bob Forsyth, Ron Hambleton, Bill Mehrens, Wim Van der Linden, Barbara Dodd, Rebecca Zwick, Eugene Johnson, and Jim Carlson. Susan Loomis was the project director and my boss and is now a very good friend. She has influenced me both professionally and personally.

While at Measured Progress I had the opportunity to work closely with the company's Technical Advisory Committee (TAC), which included Paul Holland, Bob Brennan, Ed Hartel, Terry Ackerman, and Diana Pullin. I have sought their advice during my tenure at Measured Progress and even after. My boss at Measured Progress, Kevin Sweeney, influenced my views on leadership and management.

I'd be remiss if I do not mention the following individuals who have significant influence in my professional life: my husband, Michael Nering; my academic advisor, Patti Elmore; my internship mentor, Deb Harris; and the late Brad Hanson.

LEGAL CORNER: WHAT DO THE COURTS HAVE TO SAY ABOUT TEACHER EVALUATION SYSTEMS USING STUDENT GROWTH MEASURES?

Michelle Croft, ACT

A number of lawsuits on the use of student growth measures in teacher evaluation systems have emerged in the past few years (see Sawchuk, 2015, for a full list). The standard used by the courts to review these cases is a rational basis standard where the court only needs to find that a challenged policy is rationally related to a legitimate state interest. As Judge Trauger in Tennessee's *Wagner v. Haslam* (2015), which is discussed below, stated: rational basis "allows state legislators and policymakers to make both excellent decisions and terrible decisions, provided that the decisions are based on some conceivable modicum of rationality at the time of their passage or application in practice" (p. 693) Thus, a recurring theme among the cases is that although there are legitimate concerns about the appropriateness of student growth measures for teacher evaluation, it is beyond the authority of the courts. The majority of cases decided to date have sided with the state.

Challenges to the Use of Individual Scores

In Tennessee, two teachers separately challenged the use of individual TVAAS (Tennessee Value-Added Assessment System) scores (*Trout v. Knox County Board of Education*, 2014, and *Taylor v. Haslam,* 2014). The cases were consolidated due to similar claims. The claims challenged: (a) the State's reliance on TVAAS scores based on confidence intervals, (b) the precision of the TVAAS scores, and (c) the minimum number of students on which a score must be based. In *Taylor*, the plaintiff also argued that it was impermissible to base his TVAAS score on only a portion of his students instead of all of the students he taught.

The *Trout* and *Taylor* court upheld the TVAAS system. The court found no requirement that legislative decisions regarding the use of statistics require "statistically significant" results. Regarding the precision, the court noted that the concerns about precision were not unfounded, as the TVAAS was a blunt tool, but a rational policymaker could find that it was capable of measuring some impact that teachers have on their students. For *Taylor's* subset claim, the court found that it is not irrational to base the evaluation on a subset of students, highlighting that 50% of a teacher's evaluation is based on classroom observations that only account for a small portion of the teacher's classroom time.

In May 2016, a New York court released the final order in *Lederman v. King*. The *Lederman* case involved a fourth-grade teacher who taught in a small classroom of mostly high-performing students and who was rated ineffective in 2013–2014 after being rated effective the prior year. Citing briefs filed by a number of academics supporting the plaintiff, the court found that New York's evaluation system was arbitrary and capricious as applied to the plaintiff. The decision was based on the size and makeup of Lederman's class, the use of only 1 year of data, and the strict imposition of a bell curve that did not take prior performance into account. The court declined to rule on the entire system as it had been significantly changed since Lederman filed the suit, so the decision is limited to expunging Lederman's 2013–2014 evaluation score. To date, *Lederman* is the only instance where the court has found that the system—at least as applied to a particular plaintiff—does not meet the rational basis standard.

Challenges to the Use of School-Wide Scores

Two other cases challenged Florida's and Tennessee's (*Cook v. Stewart,* 2014, and *Wagner v. Haslam*, 2015, respectively) use of school-wide growth measures for teachers in nontested subjects, such as music or physical education. Both courts found that school-wide growth measures met the rational basis standard because the state has a legitimate interest in increasing student growth. For example, in *Wagner* the court identified multiple ways that school-wide growth measures may increase school performance, such as encouraging all teachers to participate in school-wide initiatives or help students become better learners in other classes. Both courts also mentioned that the policies may not be fair to teacher in nontested subjects.

Importance of State Law

A fifth case, *Stewart v. New Mexico* (2015), introduces a new wrinkle: state law. A New Mexico statute requires a "highly objective, uniform" evaluation system (School Personnel Act, 2015). In *Stewart*, plaintiffs were granted a temporary injunction prohibiting the use of the state's evaluation system to take consequential actions against a teacher (e.g., termination decisions) until there is a trial. In its decision, the court held that the state's Public Education Department had established that value-added models "have a sound policy and statistical foundation." A trial will be needed to ensure that the evaluation system met the uniformity requirement due to the multiple evaluation frameworks that depend on a teacher's teaching assignment.

Conclusion

A number of lawsuits on the use of student growth measures in teacher evaluation systems are pending. Given the pattern of prior court decisions, the decisions in these cases will likely also find the evaluation systems legal, but not ideal.

Now that the newly enacted federal education law, the Every Student Succeeds Act, does not require states to include student growth measures for all teachers as part of teacher evaluation—as was required under the Elementary and Secondary Education Act Flexibility Waivers—it is likely that many state systems will continue to change. A number of states have already introduced legislation to remove or reduce student growth requirements (Kappler Hewitt, 2016). Hopefully, the new flexibility will provide states with the opportunity to alter their evaluation systems to address many of the issues identified in the court cases discussed here while still incorporating student growth a component of the evaluation, when appropriate.

References

Cook v. Stewart, No. 1:13-cv-72 (N.D. Fla. May 6, 2014).

Every Child Succeeds Act, Pub. L. 114-95 (Dec. 10, 2015).

Kappler Hewitt, K. (2016, March 21). Reading the tea leaves: ESSA and the use of test scores in teacher evaluation. *The Brown Center Chalkboard.* Retrieved from http://www.brookings.edu/blogs/brown-center-chalkboard/posts/2016/03/21-essa-student-test-scores-teacher-evaluation-hewitt

Lederman v. King, No. 5443-14 (Sup., AL, May 28, 2015). awchuk, S. (2015, October 6). Teacher evaluation heads to the courts. *Education Week.* Retrieved from http://www.edweek.org/ew/section/multimedia/teacher-evaluation-heads-to-the-courts.html

School Personnel Act, NMSA § 22-10A-19 (2015).

Stewart v. New Mexico Public Education Department, No. D-101-CV-2015-00409 (N.M. Dec. 2, 2015).

Taylor v. Haslam, No. 3:14-cv-00113 (E.D. Tenn., filed March 19, 2014).

Trout v. Knox County Board of Education, No. 3:14-cv-00049 (E.D.Tenn., filed Mar. 19, 2014).

Wagner v. Haslam, 112 F. Supp. 3d 673 (M.D. Tenn. 2015).

HOW THE JOINT STANDARDS REFLECT CHANGES IN PERSPECTIVES ON FAIRNESS IN EDUCATIONAL TESTING

Linda Cook, Educational Testing Service, Retired

This article briefly summarizes a paper presented in a symposium organized by Maria Elena Oliveri for the 2016 annual meeting of AERA. The paper focused on how our perspectives on fairness have evolved over the past few years and how this evolution is reflected in the *Standards for Educational and Psychological Testing* (the *Standards*) that are published by American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). The scope of the paper was limited to the past 15 years and consequently focused on comparing the fairness perspectives presented in the 1999 and 2014 *Standards*.

The 1999 *Standards for Educational and Psychological Testing*

The 1999 *Standards* are a revision of the *Standards* published in 1985 and they contain a number of changes designed to improve both their clarity and interpretation. The 1999 *Standards* also reflect evolutionary changes in the field of educational and psychological testing that occurred between the publication of the 1985 *Standards* and the publication of the 1999 *Standards*. The 1999 version did not include standards that associated fairness specifically with the treatment of the test construction, administration, and documentation process. The fairness standards in the 1999 version were found in Chapter 7 (Fairness in Testing and Test Use) in Part II, which included two distinct chapters about testing specific subgroups of the population (individuals with disabilities and individuals with diverse linguistic backgrounds).

The format of the 1999 *Standards* was reflective of thinking at that time; that is, fairness was thought of more as an outcome of testing that was evaluated after the fact rather than a key concept in testing that must be addressed in the very first steps of test design and then throughout the entire testing process.

The 2014 *Standards for Educational and Psychological Testing*

The revision to the 1999 *Standards,* which would become the 2014 *Standards*, began with a specific charge to the Joint Committee.

The charge to the Joint Committee had five components. The first component was *accessibility and fairness for all examinees*. This component had three separate parts. The first part focused on the use of accommodations and modifications. A primary concern of the authors of the 2014 *Standards* was the delineation of when a test change should be considered an accommodation or a modification. The second part was to examine the appropriateness of different test changes for individuals with disabilities and those with diverse linguistic backgrounds; the third was to broaden the concept of accessibility to all test takers and to include specific groups such as young children and aging adults when interpreting the standards. The charge also included addressing issues related to flagging test scores from nonstandard administrations. Finally, the Joint Committee was specifically charged with considering the notion of universal design in the treatment of fairness and accessibility.

The overarching standard for the fairness in testing chapter is as follows:

> All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to promote valid score interpretation for the intended uses for all examinees in the intended population. (Standard 3.0, AERA, APA, & NCME, 20014, p. 63)

This particular standard represents the most significant change in thinking about fair testing since the publication of the 1999 *Standards*. The vision of fairness described in this standard is one of a foundational component of testing that is as broad and fundamental as the concept of validity. This overarching standard makes it clear that fairness must be addressed in all aspects of testing and for all individuals who are examined by the test.

The 20 standards found in the 2014 fairness chapter are clustered under four major themes that cover the entire testing process.

Fairness Themes

Theme 1, *Test design, development, administration and scoring*, contains five standards. The standards in this cluster emphasize that it is the responsibility of the test developer to design and develop tests that are free from construct-irrelevant-barriers for the widest possible range of individuals and subgroups in the test's target population.

This group of standards includes the introduction of two new and very important fairness concepts to the standards. The first one is the notion of *accessibility* and the second is the process of *universal design*. Neither of these ideas are represented in Chapter 7 (Fairness in Testing and Test Use) of the 1999 *Standards.*

The second theme used to cluster the standards in the fairness chapter of the 2014 Standards is *validity of test score interpretations for the intended use*. There are three standards in this cluster and all of them focus on the responsibility that test developers and test users have for examining the validity of score interpretations for subgroups.

Standards under this theme stress an evolutionary change in thinking about fairness as a fundamental aspect of validity. These standards echo the statement made in the Background section of the 2014 fairness chapter; "Fairness is a fundamental validity issue and requires attention throughout all stages of test development and use" (AERA, APA, & NCME, 2014, p. 49).

There are six standards found in the cluster under Theme 3, *Accommodations to remove construct-irrelevant barriers and support valid interpretations of scores for their intended uses.* Although the standards found in the 2014 fairness chapter emphasize the use of universal design; that is, the need to minimize barriers to test content and procedures by considering the widest possible range of test takers throughout the test design, development, and administration process, the standards recognize that some test takers may still need changes or adaptations to the test or testing procedures before they can demonstrate their skills or abilities on the construct(s) measured by the test.

Unlike the 1999 *Standards* that designate all test changes as modifications, the 2014 *Standards* reflect the evolution of language in the field by differentiating between two types of test changes: accommodations and modifications.

The standards found in the fairness chapter of the 2014 *Standards* reflect the evolution of the treatment of both individuals with disabilities and those with diverse linguistic and cultural backgrounds. Current practice is to differentiate between accommodations and modifications by noting that accommodations are changes to the test and testing process that do not change the construct(s) measured by the test and modifications are changes to the test and testing process that do change the underlying construct(s). It is important to note that the 2014 *Standards* specify that if a test is modified it must be treated like a newly developed test.

Theme 4, *Safeguards against inappropriate score interpretations for intended uses,* contains 6 standards. The standards in this cluster caution test developers, test publishers, and test users against the use of tests when there is no evidence to support the use with a particular group or for a particular purpose.

One standard in this cluster introduces fairness issues related to opportunity to learn. Opportunity to learn is also covered in the Background section of the fairness chapter of the 2014 *Standards.* This version of the Standards defines opportunity to learn as, the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test…." (AERA, APA, & NCME, 2014, p. 56). The text of the Background section goes on to point out that, "Opportunity to learn is a fairness issue when an authority provides differential access to opportunity to learn for some individuals and then holds those individuals who have not had the opportunity accountable" (AERA, APA, & NCME, 2014, p. 57).

Treating opportunity to learn as a fairness issue is a notion that has evolved over the past decade, particularly as a result of the growing prominence of accountability testing in public schools.

Summary

Over the past 15 years, a number of evolutionary changes have occurred in perspectives on fairness in educational testing. Among the most important changes are (a) the notion that fairness and validity are inseparable concepts in testing; that is, fairness concerns apply to everyone in the testing population and need to be considered in all aspects of testing, (b) the introduction of the concepts of accessibility and universal design as fundamental to the testing process, (c) the definition and use of the terms modification and accommodation for test changes that remove construct-irrelevant barriers from the testing process, and, (d) the introduction of opportunity to learn as a key aspect of fairness in some types of testing. In addition, the 2014 *Standards* take a different approach to flagging scores on score reports, emphasizing that flags should be attached only when there is "clear evidence that scores from regular and altered test administrations are not comparable…" (AERA, APA, & NCME, 2014, p. 61).

As Plake and Wise (2014) pointed out in a recent issue of *Educational Measurement: Issues and Practice*, "…because educational and psychological testing is a dynamic field, the *Standards* will need to be revised again, and again, and again" (p. 12). And, I believe it is safe to say, so will there be a need to revise our perspectives on what constitutes fair interpretations of test scores for their intended uses and target populations.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME *Standards for Educational and Psychological Testing*? *Educational Measurement: Issues and Practices, 33*(4), 4-12.

2016 NCME Award Winners

Career Contribution Award: Marck Reckase (see the March 2016 issue of this newsletter)

Alicia Cascallar Award: Dr Okan Bulut

Annual Award: Mark Gierl and Hollis Lai

Bradley Hanson Award: Sun-Joo Cho

Brenda H. Loyd Award: Yuanchao Emily Bo and Youn-Jeng Choi

Jason Millman Award: Katherine Furgol Castellano

The 2016 Winner of the Alicia Cascallar Award for an Outstanding Paper by an Early Career Scholar

Christina Wikström, Alicia Cascallar Award Committee, Past Chair

Dr. Okan Bulut is the 2016 winner of the Alicia Cascallar Award for an Outstanding Paper by an Early Career Scholar. He is an assistant professor of educational assessment and technology as well as a member of the Centre for Research in Applied Measurement and Evaluation (CRAME) at the University of Alberta.

Okan Bulut received the award for his paper "Examining Subscore Reliability Within the Multidimensional IRT Framework." The paper was part of the coordinated symposium session titled "Three Ways to Improve Subscores Using Multidimensional Item Response Theory" in the 2014 annual meeting of NCME.

Both recommendation letters support the quality and importance of Dr. Bulut's work, and the award committee concurs that the winning paper is excellent. The paper gives both a theoretical and a practical contribution, showing how subscore reliability can change under diverse testing conditions, with practical suggestions for researchers and practitioners.

The Alicia Cascallar Award for an Outstanding Paper by an Early Career Scholar has been established to honor Alicia Cascallar's professional commitment and accomplishments and to continue her practice of mentoring and encouraging promising new scholars in the area of educational measurement. The award will be given to an early career scholar for an outstanding paper presented at one of the two most recent NCME annual meetings. More information about the award and how to nominate can be found on the NCME website.

The 2016 Winner of the Annual Award

Jonathan Templin, Annual Award Committee, Past Chair
Karla Egan, EdMetric LLC, Current Chair

The 2016 NCME Annual Award was awarded to Drs. Mark Gierl and Hollis Lai, both of the University of Alberta. Dr. Gierl is a professor of educational psychology and director of the Centre for Research in Applied Measurement and Evaluation (CRAME). He also holds the Canada Research Chair in Educational Measurement. Dr. Lai is an assistant professor at the University of Alberta School of Dentistry.

The annual award is given for exceptional service or for outstanding *technical* or *scientific* contribution to the public or the field of educational measurement. Through their ongoing research program, Drs. Gierl and Lai have made significant technical contributions for developing automatic item generation methods. There is a growing need for large banks of assessment items in both education and licensure assessment. The work of Drs. Gierl and Lai represent an exciting means for addressing practical item development problems. Congratulations to Dr. Gierl and Dr. Lai!

The 2015 Winner of the Bradley Hansen Award

David Miller, Bradley Hansen Award Committee, Past Chair

Dr. Sun-Joo Cho, an associate professor at Peabody College of Vanderbilt University, won the 2016 Bradley Hanson Award for Contributions to Educational Measurement. Dr. Cho won for her research project, "Specification, Illustration, and Evaluation of Multilevel Reliability Measures." Her study includes (a) specifying multilevel reliability measures in various multilevel structures, (b) investigating the performance of multilevel reliability measures via simulation studies, and (c) developing Mplus syntax for parameter estimation and R code to calculate multilevel reliability measures in various multilevel designs.

The 2016 Winner of the Brenda H. Loyd Outstanding Dissertation Award

Seock-Ho Kim, Brenda H. Loyd Award Committee, Past Chair

Yuanchao Emily Bo and Youn-Jeng Choi are the corecipients of the award. Yuanchao Emily Bo is currently a postdoctoral researcher at UCLA CRESST. She received her doctorate in psychology from Fordham University. Her award winning dissertation titled *An Option-Based Partial Credit Item Response Model* was directed by David V. Budescu and Charlie Lewis. Her dissertation reflects the intellectual creativity and innovation in test scoring using a novel technique. Youn-Jeng Choi is currently an assistant professor at the University of Alabama. She received her doctorate in educational psychology from the University of Georgia. Her award winning dissertation entitled *Metric Identification in Mixture IRT Models* was directed by Allan S. Cohen. Her dissertation demonstrates the outstanding technical quality in modeling of complex assessment data. Both dissertations are excellent and make great contributions to the field of educational assessment.

The 2016 Winner of the Jason Millman Promising Measurement Scholar Award

Lihua Yao, Jason Millman Award Committee, Past Chair

Katherine Furgol Castellano was this year's Jason Millman Promising Measurement Scholar. This award serves to honor Jason Millman's work by recognizing a scholar at the early stages of their career whose research has the potential to make a major contribution to the applied measurement field. Dr. Castellano was awarded for her work and contributions on student growth percentiles (SGPs). She has grounded SGPs in a broader statistical framework, providing the first formal treatment of the topic as well as critiquing the approach through contrasts with other established methods.

Dr. Castellano is currently a psychometrician at Educational Testing Service, San Francisco, CA. She received her Ph.D. in Educational Measurement and Statistics at the University of Iowa.

Thank you to Dr. Castellano for her commitment to applied measurement and to the Jason Millman award committee for their great efforts in reviewing nomination packets and deliberating a deserving winner. I'd also like to thank all those who submitted nominations and packets. Please consider nominating someone you know next year!

PUBLICATION COMMITTEE UPDATE: MOST DOWNLOADED ARTICLES ON WILEY ONLINE LIBRARY

Rose C. McCallin, Colorado Department of Regulatory Agencies

The Publications Committee is responsible for working with publishers on behalf of NCME. We review and report metrics relevant to evaluating the success of NCME journals and books. Members include James Carlson, Educational Testing Service; Kimberlee Everson, Western Kentucky University; Jennifer Kobrin, The Graduate Center, City University of New York; Kevin Krost, Virginia Polytechnic Institute and State University; William Lorié, Pearson Center for NextGen Learning & Assessment; Rose McCallin, Colorado Department of Regulatory Agencies; and Mark Raymond, National Board of Medical Examiners. Kristen Huff is our Board liaison.

Jim Carlson

Kim Everson

Jennifer Kobrin

Kevin Krost

Will Lorié

Mark Raymond

As the publisher of the *Journal of Educational Measurement* (*JEM*) and *Educational Measurement: Issues and Practice* (*EM:IP*), Wiley provides NCME an annual report summarizing activities it undertook to publish the two journals and its work to promote discoverability, circulation, and readership of articles in each journal. The annual reports include a list of the top 10 most downloaded *JEM* and *EM:IP* articles on Wiley Online Library for the year with the number of online accesses. Using the top 10 lists from the 2014 and the recently received 2015 Wiley annual reports and supplemental data from the publisher, this article offers a 24-month snapshot of the most downloaded articles from *JEM* and *EM:IP* on Wiley Online Library from January 1, 2014, through December 31, 2015.

Sixteen articles published from 2001 (with June 15, 2006, as the version of record online) to 2015 comprised the most downloaded *JEM* articles on Wiley Online Library over the past 24 months (Figure 1). Four of the 16 articles (25%) appeared in both the 2014 and 2015 top 10 lists in the Wiley annual reports. Fourteen articles published from 1993 (with October 25, 2005, as the version of record online) to 2015 comprised the most downloaded *EM:IP* articles on Wiley Online Library (Figure 2). Six of these articles (43%) were in the top 10 lists in the 2014 and 2015 annual reports.

Please note that articles on Wiley Online Library for longer amounts of time during the 24 months summarized herein (i.e., those with online version of record dates before and in early 2014) had a better chance for more downloads, compared to those with online version of record dates in late 2014 and early 2015. The online version of record date, which indicates when the article was first available for downloads on Wiley Online Library, is included in Figures 1 and 2 to help readers keep this point in mind. For more information on these articles, please see the supplemental figures available on the NCME website.

Remember, your NCME membership includes free access to all *JEM* (1964–present) and *EM:IP* (1982–present) articles on Wiley Online Library. Simply sign in as a member on the NCME website, select the Publication tab on the home page and the journal name, and follow the prompts. While you are on the Wiley website, please consider signing up for new content alerts under Journal Tools if you have not done so. There were 991 individuals registered to receive new content alerts for *JEM* and 1,126 for *EM:IP* as of December 31, 2015.

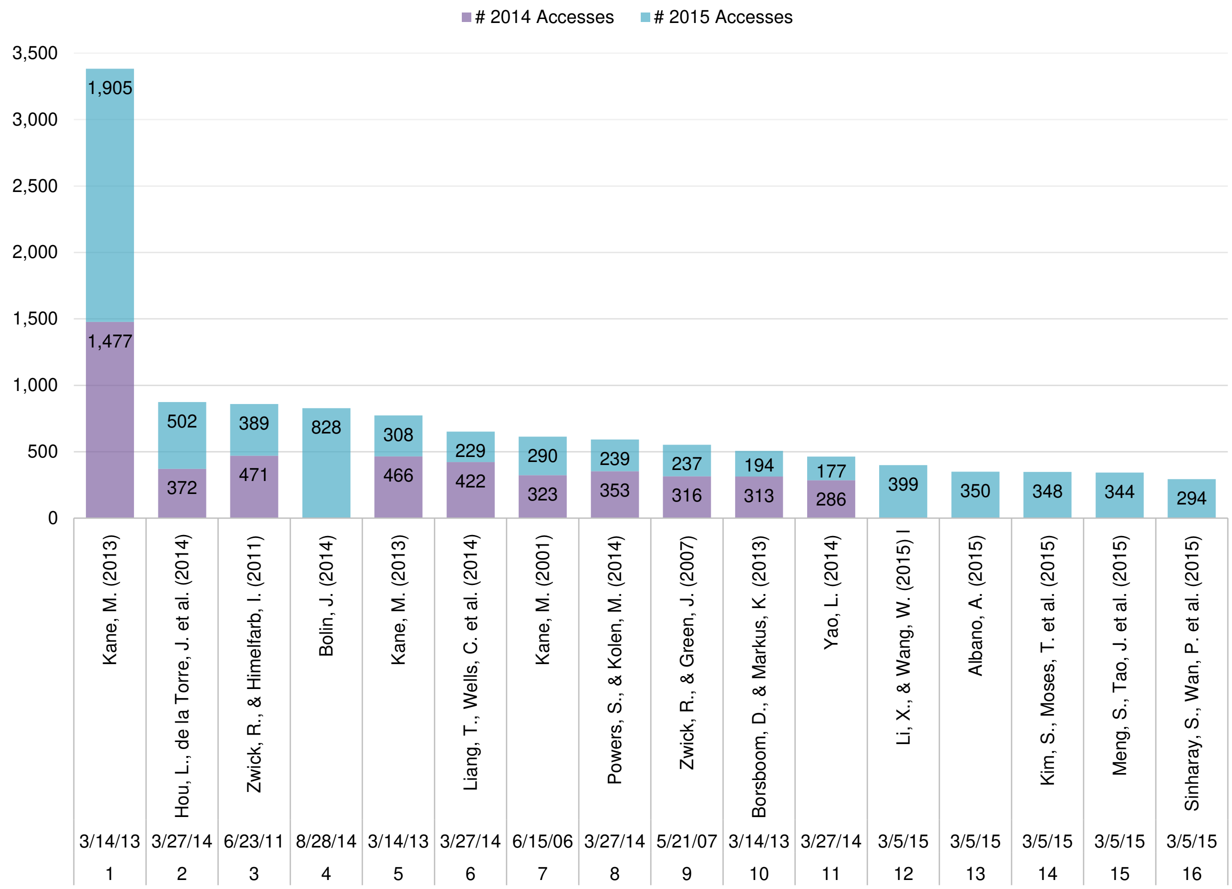


Figure 1. Most downloaded *JEM* articles on Wiley Online Library from 01/01/2014–12/31/2015: Author(s), year, and online version of record date. See below for the titles of the articles.

[1]Validating the Interpretations and Uses of Test Scores
[2]Differential Item Functioning Assessment in Cognitive Diagnostic Modeling: Application of the Wald Test to Investigate DIF in the DINA Model
[3]The Effect of High School Socioeconomic Status on the Predictive Validity of SAT Scores and High School Grade-Point Average
[4]Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach by Andrew F. Hayes
[5]Validation as a Pragmatic, Scientific Activity
[6]An Assessment of the Nonparametric Approach for Evaluating the Fit of Item Response Models
[7]Current Concerns in Validity Theory
[8]Evaluating Equating Accuracy and Assumptions for Groups That Differ in Performance
[9]New Perspectives on the Correlation of SAT Scores, High School Grades, and Socioeconomic Factors
[10]Truth and Evidence in Validity Theory
[11]Multidimensional CAT Item Selection Methods for Domain Scores and Composite Scores With Item Exposure Control and Content Constraints
[12]Assessment of Differential Item Functioning Under Cognitive Diagnosis Models: The DINA Model Examples
[13]A General Linear Method for Equating with Small Samples
[14]A Comparison of IRT Proficiency Estimation Methods Under Adaptive Multistage Testing
[15]A Conditional Joint Modeling Approach for Locally Dependent Item Responses and Response Times.
[16]Assessing Individual-Level Impact of Interruptions During Online Testing

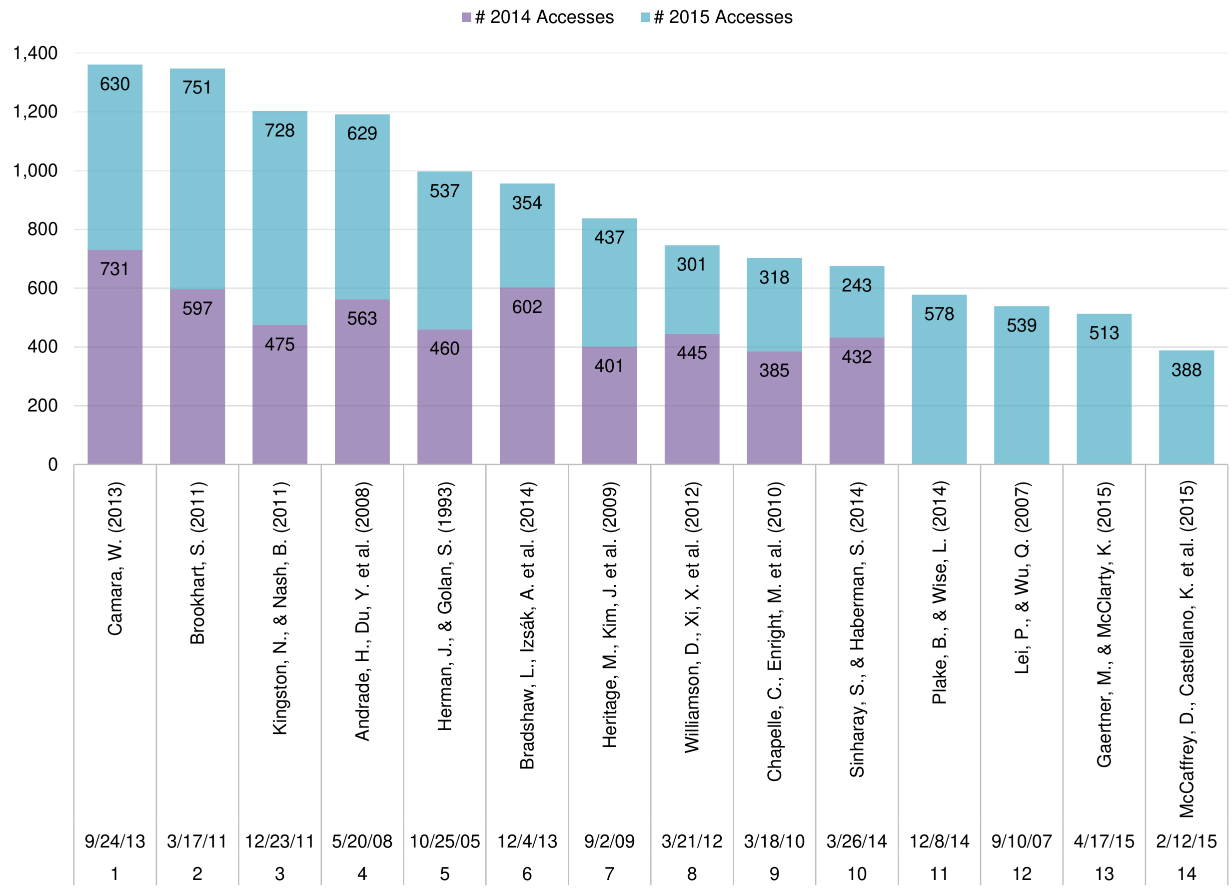


Figure 2. Most downloaded *EM:IP* articles on Wiley Online Library from 01/01/2014–12/31/2015: Author(s), year, and online version of record date. See below for the titles of the articles.

[1]Defining and Measuring College and Career Readiness: A Validation Framework
[2]Educational Assessment Knowledge and Skills for Teachers
[3]Formative Assessment: A Meta-Analysis and a Call for Research]
[4]Putting Rubrics to the Test: The Effect of a Model, Criteria Generation, and Rubric-Referenced Self-Assessment on Elementary School Students' Writing
[5]The Effects of Standardized Testing on Teaching and Schools]
[6]Diagnosing Teachers' Understandings of Rational Numbers: Building a Multidimensional Test Within the Diagnostic Classification Framework
[7] From Evidence to Action: A Seamless Process in Formative Assessment?
[8]A Framework for Evaluation and Use of Automated Scoring
[9]Does an Argument-Based Approach to Validity Make a Difference?
[10]How Often Is the Misfit of Item Response Theory Models Practically Significant?
[11]What Is the Role and Importance of the Revised AERA, APA, NCME Standards for Educational and Psychological Testing?
[12]Introduction to Structural Equation Modeling: Issues and Practical Considerations
[13]Performance, Perseverance, and the Full Picture of College Readiness
[14]The Impact of Measurement Error on the Accuracy of Individual and Aggregate SGP

NCME Outreach Committee
Peer Review of State Assessment Systems—Update

Tracey Hembry, Alpine Testing Solutions

This year, the federal peer review process for state assessment programs has resumed. Chris Domaleski provided us with an overview of the process and the updated guidelines in the March 2016 newsletter. As part of the peer review process, the U.S. Department of Education (ED) solicited applications for education professionals, especially those with measurement expertise, to participate as reviewers (see the ED website for more details). In November 2015, the Outreach Committee emailed NCME members to inform the group of this opportunity. While ED is no longer accepting peer reviewer applications, the Outreach Committee has been in contact with ED to gather information and provide our membership with a status update on the peer review process.

The first round of peer reviews occurred in April 2016. However, only a few states had materials ready for the review and therefore few reviewers were needed. Two additional rounds of peer review will occur this summer, one in June and another in August. Applicants who were selected for participation in these reviews have been contacted by ED.

Peer reviewer applications are evaluated for demonstrated knowledge of and experience with state assessment programs as well as potential conflicts of interest. The ED ethics division carefully considers applicants' current and prior work as well as relationships with vendors, states, and other professionals in the consideration of potential conflicts of interest. Both direct and indirect conflicts of interest are defined using a 12 month period, meaning that past work, beyond 12 months, will likely not impact an applicant's ability to participate in the process.

Typically, each assessment program is reviewed by a panel of three to five peer reviewers. Following training from ED on the federal peer review guidelines, the panel spends approximately 4-6 days working individually and remotely to review the materials provided by the assessment program. The panel then convenes in person for several days to work together to continue the review. At the conclusion of the review, the panel provides feedback to ED and the state or consortium about various technical details of the assessment program.

For those members who applied to serve as peer reviewers, thank you for your time and service to the broader education community.

If you become aware of additional opportunities for NCME members to collaborate with or contribute to the professional community, please feel free to share this information with the NCME Outreach Committee at tracey.hembry@alpinetesting.com.

Diversity Issues in Testing Committee Update

Priya Kannan, Educational Testing Service

The Diversity Issues in Testing Committee is entrusted with the responsibility to ensure that diversity is considered throughout NCME's various activities. The committee primarily focuses on issues surrounding diverse test takers and their experiences, but is also responsible for ensuring diversity in NCME's internal governance, membership, and directed activities. A number of such activities are undertaken in collaboration with other NCME committees. For example, in collaboration with the Outreach and Partnership Committee and the Standards and Test Use Committee, we recently reviewed storyboards submitted by different organizations in response to the NCME Board's initiative to revise assessment literacy materials developed for the general public.

One of this committee's primary responsibilities is to organize a symposium around diversity issues for NCME's annual meeting each year. During the 2016 annual meeting, we successfully organized and moderated a symposium, "Implications of Computer-Based Testing for Assessing Diverse Learners," in which representatives from all six consortia (PARCC, Smarter Balanced, DLM, NCSC, ELPA-21, and WIDA) described lessons learned from their 2014–2015 operational administration. The focus of each consortia's presentation was on the implications of designing and administering valid and inclusive computer-based tests for diverse learners.

The Diversity Issues in Testing Committee is composed of six members, one of which is a student member. Members serve a 3-year term, with the student member appointed annually. The current members of the committee are Annie Davidson (co-chair 2016–2017), Dorinda Gallant, Priya Kannan (chair 2016–2017), Meagan Karvonen (past chair), Hongling Lao (student member), and Lietta Scott. Please consider volunteering to serve on the committee.

NCME CAREER CONTRIBUTIONS AWARD COMMITTEE

Barbara Plake, University of Nebraska-Lincoln, Emeritus

The Career Contributions Award Committee membership consists of Barbara Plake University of Nebraska-Lincoln, Emeritus (Chair), Allan Cohen, University of Georgia (Co-Chair), Wayne Camara, ACT (past Chair), Michael Kane, Educational Testing Service, Suzanne Lane, University of Pittsburgh, Anita Rawls, College Board, and Stella Kim, University of Iowa (student member). Deadline for nominations for the 2017 Career Contributions Award is August 1, 2016. See the NCME website for details.

NCME FITNESS WALK/RUN CORNER

Jill van den Heuvel, Alpine Testing Solutions
Katherine Furgol Castellano, Educational Testing Service
(with Brian French, Washington State University, Pullman, advising)

The 2016 NCME Fitness Walk/Run in Washington, DC was a great success again this year! We appreciate all the NCME members who braved the chilly, windy, and (somewhat) rainy weather conditions to join us for an early, but invigorating, start to the last day of the conference. We are greatly indebted to race coordinator Matthew Eastman of the Fun Runz organization for making this event such a success. His efficiency in boarding the buses and running the event once we got to the park had us back to the hotel in record time, giving everyone plenty of time to prepare for the first morning sessions of the day. And thanks to Matthew's quick thinking, we were able to seamlessly change the start/finish line when we found we couldn't drive into the park because the park gate was down.

We are also pleased to announce the winning teams for Team Participation Competition. The University of Iowa and Pacific Metrics held their winning streaks for the university and testing company categories, respectively. We hope their enthusiasm for this event each year motivates other groups to compete in the future. In an effort to continue to increase participation in the Team Competition and NCME fitness events overall, we teamed up with the NCME Sunrise Yoga coordinators, Kristen Huff and Brian Patterson, to develop and administer a survey to Yoga and Fitness Run/Walk registrants soliciting feedback on these events. We are still tabulating the results so be on the lookout for them in the next issue of the *NCME Newsletter*. Feel free to contact us anytime throughout the year with feedback and suggestions on making these events highlights of your NCME conference experience.

Keep moving and have a great summer!

NCME Fun Run 2016 participants

Members of Team Iowa

Members of Team Pacific Metrics

PDF ACCESSIBILITY OF THE NCME NEWSLETTER

Kimberlea Fryer, Educational Testing Service

The goal of accessibility for web content is to make the content easier for people with disabilities to access. By following standards such as those advocated by the World Wide Web Consortium, authors of web content can use various tools and techniques to make the content accessible to users who rely on assistive technologies to access the web, such as screen readers or screen magnifiers. Web content encompasses such things as web pages, blog entries, videos, and files that can be downloaded or viewed on the web, such as pdfs.

We are endeavoring to create accessible pdfs for the NCME newsletter, following the Web Content Accessibility Guidelines (WCAG) 2.0 from the World Wide Web Consortium. These guidelines cover content that can be accessed or downloaded on the web. Our effort to create accessible files for the newsletter entails optimizing the pdfs to make them more accessible to people with visual disabilities such as low or no vision or who have mobility issues that restrict use of a navigation device, such as a mouse. Among the requirements that WCAG 2.0 guidelines call for in accessible pdfs are tagging, identification of certain document properties, use of alternate text (called *alt text*) and color specifications for any figures or other graphic elements, identification of rows and columns in tables, and clickable links. While there other requirements for some pdfs, depending on the content, the requirements covered in this article pertain to the content we encounter most often in this newsletter.

Tagging

One of the most important steps in creating an accessible pdf is to ensure that the pdf is properly tagged. These tags, which are not visible to the user, structure and identify the content for assistive technologies such as screen readers or text magnifiers so that the content is presented in a logical reading order that can be easily navigated. The tags, which are similar to those used in HTML, identify such content elements as headings, paragraphs, figures, lists, and tables. Many of the tags can be created when generating the pdf from the native file, such as a Word file, when the native file is properly set up. For example, in Word, when the heading styles are used to format/identify headings, a pdf that is generated using Adobe Acrobat will have the headings tagged. However, some tags need to be created or edited after the pdf is generated.

Document Properties

Among the document properties that must be identified in the pdf is the primary language in which the content has been written. The title of the document must also be specified in the properties and the document title should appear in the title bar of the pdf.

Graphical Elements

Graphical elements such as figures need to include alt text describing each image. Decorative graphical elements, such as rules, can be tagged as artifacts so that they are not read by screen readers and thus do not need alt text. For people who have visual disabilities such as color blindness or low vision, color in figures should be carefully considered so that the color alone is not used as a unique identifier for figure elements (e.g., trend lines in a graph) or so that the color contrasts sufficiently with other colors so that text or other elements are not obscured.

Tables

In accessible pdfs, tables generally should not be used as a layout tool (i.e., for positioning of text or figures). Regardless of whether this kind of table is visible to sighted users, assistive technologies such as screen readers will identify the table, which could prove confusing to users with visual disabilities. For data tables, column and row headers should be identified through tagging and linked to the data cells.

Links

Hyperlinks in pdfs should be active links and use short descriptive text instead of the url whenever possible. In addition, the links should be keyboard accessible for those users who cannot use a mouse or have difficulty using one.

Other Requirements

The requirements for accessible pdfs covered in this article aren't comprehensive, but the article does introduce the WCAG 2.0 requirements that we most often need to meet in the newsletter, as dictated by the newsletter content. Other WCAG 2.0 requirements for accessible pdfs can include such things as bookmarks for longer pdfs, summaries for large tables, and description of fields for forms.

More Information

I have compiled a list of resources that provide more information on accessible pdfs. These resources include guidelines, tutorials, and best practices.

Adobe. (n.d.). PDF accessibility overview. Retrieved from http://www.adobe.com/accessibility/pdf/pdf-accessibility-overview.html

Adobe (n.d.). What is accessibility? Retrieved from http://www.adobe.com/accessibility/gettingstarted.html

United States Department of Veteran Affairs. (n.d.). Creating accessible PDFs with Adobe Acrobat Professional. Retrieved from http://www.section508.va.gov/support/tutorials/pdf/index.asp

W3C. (2008). Web content accessibility guidelines (WCAG) 2.0. Retrieved from http://www.w3.org/TR/WCAG20/

W3C. (2012). PDF techniques for WCAG 2.0. Retrieved from http://www.w3.org/TR/WCAG20-TECHS/pdf.html

Outgoing Board members Amy Hendrickson and Cindy Walker with Past President Laurie Wise at the 2016 annual meeting.

NCME Board members and NCME presidents at the 2016 annual meeting (L to R) Won-Chan Lee (board member), Kristen Huff (board member), Cindy Walker (past board member), Laurie Wise (past president, 2014–2015), Rich Patz (past president, 2015-2016), Mark Wilson (president, 2016–2017), Luz Bay (board member), and Amy Hendrickson (past board member).

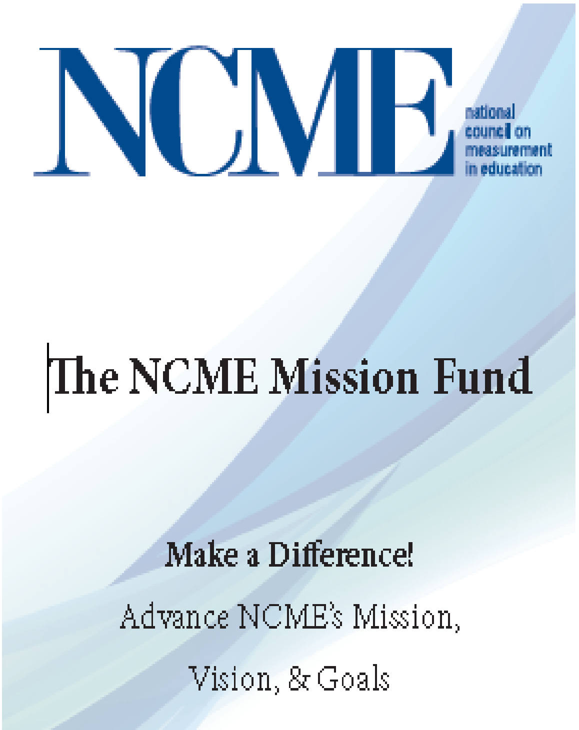
Next Annual Meeting of the National Council on Measurement in Education

April 26–30, 2017
San Antonio, TX

Theme: Advancing Large Scale and Classroom Assessment Through Research and Practice

Proposals may be submitted from: June 20 to August 1, 2016

Approximate date for notification of acceptance/rejection decisions: November 15, 2016



What is the NCME Mission Fund?

The Mission Fund allows NCME to make a difference by carrying out additional activities, beyond those supported by NCME funds, which advance our mission in the science and practice of measurement in education. The Mission Fund entered its third year as of the 2016 annual meeting.

How has the Mission Fund made a difference?

- In 2015 the Mission Fund supported a workshop at the June CCSSO National Conference on Student Assessment on the implications of the newly released AERA/APA/NCME Standards in state testing accountability. Click here to view a video of the workshop.
- In 2016 the Mission Fund provided partial fee reimbursements for graduate students who attended NCME training sessions. A total of 17 graduate students, representing 13 different institutions, received reimbursements.

What can you do to make a difference?

- *Make a donation!* More than $5,000 has been contributed this year at the time of the annual meeting, but more is needed to reach the Year 3 goal of $50,000. *No amount is too small (or too large).* Make your tax-deductible donation online at http://www.ncme.org/
- *Tell us why you support the Mission Fund!* Email any Mission Fund Development Committee member and tell us what the Mission Fund means to you and why you support it. We welcome all comments and ideas.

2016-2017 Mission Fund Development Committee Members

Cathy Wendler (Chair) cwendler@ets.org
Linda Hargrove (Past Chair) Linda.Hargrove@THECB.state.tx.us
Deborah Harris deborah.harris@act.org
Suzanne Lane sl@pitt.edu
Seohong Pak seohong-pak@uiowa.edu
John Willse jtwillse@uncg.edu

Send articles or information for this newsletter to:

Heather Buzick
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541

e-mail: hbuzick@ets.org

To cite an article in this newsletter, please use this suggested format:
Author, A. A. (Year, Month). Title of article. *NCME Newsletter, X*(X), x–x.

The newsletter is published quarterly (March, June, September, and December). Download issues at http://www.ncme.org/ncme/News