CHAPTER 25

The Times They Are A-Changing, but the Song Remains the Same

Future Issues and Practices in Test Validation

Stephen G. Sireci and Molly Faulkner-Bond

Come gather 'round people
Wherever you roam
And admit that the waters
Around you have grown
And accept it that soon
You'll be drenched to the bone
If your time to you, Is worth savin'
Then you better start swimmin'
Or you'll sink like a stone
For the times they are a-changin.'
—Bob Dylan, "The Times They Are A-Changin'"[1]

As we write this Chapter 15 years into the 21st century, the world of educational assessment is full of change. It is the first year the new Race to the Top assessment consortia (groups of states and jurisdictions in the United States that have come together to develop tests that measure new K–12 curriculum standards) administered their tests. The new standards measured by these tests focus on getting students ready for 21st-century colleges and careers, and involve "on-track" or "readiness" benchmarks for students from third through 11th grade. In addition, these consortia tests, like assessments used in licensure and certification, are incorporating technology in exciting ways. The consortia assessments are also being used, or considered for use, to hold public schools accountable in the United States, which may lead to rewards and sanctions for teachers, educational administrators, schools, districts, and

[1] "The Times They Are A-Changin'" lyrics copyright © 1963, 1964 by Warner Bros. Inc., renewed 1991, 1992 by Special Rider Music.

states. Clearly, the waters around us have grown, and it appears the times are certainly a-changin' in educational measurement.

But how new and different are these "changes"? One can only answer that question by knowing what has come before. Fortunately, in this book there are several chapters on the history of educational assessment, and others that describe current practices and projections of future practice. Thus, this book provides a unique opportunity for us to review the history of our field through the lens of the present. In this chapter, we relate the current, and somewhat frenetic, trends in educational assessment to the most pressing issues and important developments that have occurred since educational tests came into vogue over 100 years ago. Our motivation behind this inquiry is to learn from the past so that it can inform our future. A key question we ask is, "Are the times really a-changin', or is what we are experiencing essentially the same as what has occurred in the past?" We are hopeful that our field is not an example of what George Santayana envisioned when he wrote, "Those who cannot remember the past are condemned to repeat it" (1905, p. 284).

Current Trends in Educational Assessment: A Look Forward or Backward?

A review of the chapters in this book, current measurement journals, and the popular press uncovers several current "hot topics" in educational assessment. These topics include:

- Using students' performances on tests to evaluate teachers, schools, and others involved in public education.
- Measuring "growth."
- Providing more diagnostic information regarding students' proficiencies.
- Improving the assessment of students with disabilities and English learners.
- Using tests for international comparisons of educational achievement.
- Embedding assessments into instruction for more formative purposes.
- Using technology to improve educational assessments.

A theme across these topics is the desire to have educational tests do more than they have traditionally been called to do. We want to use tests not only to say something about a student, but also to infer something about that student's teacher, principal, school, and district. We want tests to tell us not only how well a student is doing overall in a subject area; we want details about his or her strengths and weaknesses. We want tests to be standardized so that they provide a level playing field for everyone, but we also want them to be flexible enough to accommodate the needs of students who are English learners or who have disabilities. We want to use tests to compare and evaluate students not only within a classroom, but also across states and countries. We

want tests not only to measure how well students learn, but also to help them to learn. We want tests to be an interactive experience for students, so we can measure skills beyond those that are measurable in a paper-based format. Clearly, the demands on 21st-century educational tests are daunting.

However, before concluding that we will get "drenched to the bone" by these "waters of change," it is important to remember there are several fundamental constants in educational testing that will not change. For example, although the way we build tests or the way we use their scores may change, the criteria of validity and fairness with which we evaluate such use remains an important constant. In the next section of this chapter, we sort these "hot topics" into five current trends to discuss both current issues and historical constancies within each topic area.

Current Trend 1: Accountability Testing

As Linn (Chapter 18, this volume), Geisinger and Usher-Tate (Chapter 1, this volume) and many others have pointed out, the No Child Left Behind (NCLB) Act of 2001 elevated the role of tests in education reform movements. NCLB originally required all states receiving federal funds for K–12 education to develop math and reading tests aligned with statewide curriculum frameworks in grades 3 through 8, and in one grade in high school. By 2007, science assessments in at least one grade in each of three grade spans (3–5, 6–9, and 10–12) had to be implemented as well. States were also required to establish at least three standards of performance on the assessments, one of which needed to signify "proficient" in the subject area for that grade. Students were classified into proficiency categories based on their test scores, and schools were evaluated for "adequate yearly progress" (AYP) based on how well they were meeting the goal of attaining 100% proficiency for all students by the year 2014. Schools could be rewarded or sanctioned based on their yearly progress. Districts had a similar accountability process.[2] Thus, the inferences from these assessments were generalized beyond the individual student level to support inferences about schools and school districts.

Subsequent federal initiatives increased the use of tests for generalizing from student inferences to inferences at a larger system level. "Flexibility waivers" allowed states to avoid the NCLB requirements if they adopted rigorous curriculum standards (e.g., the Common Core State Standards) and used students' test scores to evaluate teachers (typically by using changes in students' test performance across years as a measure of "growth"; see Keller, Colvin, & Garcia, Chapter 19, this volume). The Race-to-the-Top initiative funded two consortia of states to develop common assessments to measure whether students were "on-track" (below grade 11) or "ready" for college and career. Again, it was implicit that the performance of students on these consortium assessments would be used to evaluate teachers, schools, and school districts.

[2]The details of AYP under NCLB are beyond the scope of this chapter; interested readers are referred to Chudowsky and Chudowsky (2005) and the White House (2002).

Test-based accountability systems have been criticized for using test scores beyond the purposes for which they have been validated. The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). Most, if not all, validity evidence for NCLB assessments focuses on interpreting the proficiency of individual students. Little to no evidence has focused on using aggregations of students' performance, such as "value-added estimates" of teaching effectiveness or "median student growth percentiles" to evaluate teachers, schools, or districts.

This lack of validation of educational tests for accountability purposes is troubling and puts many statewide educational testing programs in violation of the AERA, APA, and NCME (2014) *Standards*, which state:

> An index that is constructed by manipulating and combining test scores should be subjected to the same validity, reliability, and fairness investigations that are expected for the test scores that underline the index. (p. 210)

and:

> Users of information from accountability systems might assume that the accountability indices provide valid indicators of the intended outcomes of education . . ., that the differences among indices can be attributed to differences in the effectiveness of the teacher or school, and that these differences are reasonably stable over time and across students and items. These assumptions must be supported by evidence. (p. 206)

Unfortunately, there is little research to support value-added estimates of teachers' effectiveness (Braun, 2013) or aggregations of student "growth" percentiles (Wells, Sireci, & Bahry, 2014).

The use of tests beyond the individual student level for which they have been validated has amplified traditional criticisms of educational tests. Critics rightly claim that evaluating teachers and schools based on students' test performance does not control for preexisting differences across students assigned to teachers (because the assignment of students to teachers is not random) and does not control for differences in many variables that are known to be associated with achievement, such as socioeconomic status, parental education, and school and community resources. Interestingly, however, the use of tests to evaluate teaching, and criticisms of the practice, are not new developments. The practice itself, and criticisms of it, can be traced to the earliest days of modern educational testing.

Ruch (1929), for example, remarked that the use of tests for evaluating teaching seemed like a good idea at the outset, but was quickly criticized by the National Education Association for not accounting for differences across the students assigned to teachers (p. 12). On determining the causes of differences across teachers, Pressey and Pressey (1922) wrote:

> First to be considered is the possibility that the work in one school may be poor because the children in the school are unusually dull and consequently, even with the best of instruction, do not learn readily. . . . A supervisor will do a large injustice to the teachers in a "poor" district of a city if she fails to take account of this factor and attributes to poor teaching what is really due to the poor . . . capacity of the children who are being taught. (p. 28)

The excerpt was written almost 100 years ago, which suggests the times aren't a-changin' as much as we aren't a-learnin'. Pressey and Pressey also advised, "Test results must be used along with, not to the exclusion of, other sources of information" (p. 69). This sounds like sage advice to us, and also sounds remarkably similar to the AERA (2000) "Position Statement; High-Stakes Testing," which relayed the same cautions regarding student testing. We believe, in addition to reminding policymakers and others of the limitations of a single test score, research is needed on sensible and practical methods for combining multiple measures into an accountability system in a manner that takes into account the unique characteristics of each individual, school, and classroom. If student progress based on educational tests is to be a major component of the system, more research is needed on better ways for measuring student progress. Keller, Colvin, and Garcia (Chapter 19, this volume) present some promising ideas in this area.

Current Trend 2: Making Assessments More Accessible

The most recent version of the AERA and colleagues (2014) *Standards* features a much larger chapter on "Fairness" that focuses on issues regarding testing special populations such as examinees with disabilities and linguistic minorities (e.g., English learners). The *Standards* acknowledge fairness can be defined in different ways and state:

> This chapter interprets fairness as responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended uses. . . . A test that is fair within the meaning of the *Standards* reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct. (p. 50)

An important component of fairness is *accessibility*, which the *Standards* define as "the notion that all test takers should have an unobstructed opportunity to demonstrate their standing on the construct(s) being measured" (p. 49). Accessibility is an important issue for many populations of examinees, such as those with disabilities or those who are not fully proficient in the language in which the test is administered [in the United States, this second group is referred to as *English learners* (ELs) (Forte & Faulkner-Bond, 2010)].

A positive recent trend in educational assessment is considering the unique needs of ELs and examinees with disabilities in test development and administration (Thurlow

& Quenemoen, Chapter 24, this volume). The concept of *universal test design* informs test development by minimizing any features of the test that may present barriers to certain groups of examinees (e.g., overly linguistically complex items that may hinder ELs). Universal design can also be applied to testing conditions to make them more flexible so that accommodations are not needed for certain groups. Traditionally, examinees with disabilities applied for and received accommodations to test administration conditions such as extra time, oral administration, or increased font size. Universal design in test administration would involve (1) setting time limits under which virtually all examinees would have sufficient time to complete the test and (2) making other options, such as larger print or oral administration, available for all examinees.

Even with universal design in test development and administration, some examinees may still need accommodations. However, the ubiquitous presence of universal test design and the provision of test accommodations illustrate the serious efforts undertaken by testing agencies to ensure fairness in educational testing. We believe that these *practices* are new, relative to testing in the 20th century, and we appreciate their contribution to fairness in testing. However, it is interesting to note that the *concerns* are not new. For example, Likert (1932), in his famous introduction of survey development that became known as Likert scaling, remarked, "Because a series of statements form a unit or cluster when used with one group of subjects which justifies combining the reactions to the different statements into a single score, it does not follow that they will constitute a unit on all other groups of persons with the same or different cultural backgrounds" (pp. 51–52). Likert's insight was impressive in that he was essentially concerned with measurement invariance across subpopulations of students during a time when racism was more common than fairness. Today, measurement invariance is a fundamental aspect of fairness for educational and psychological assessments (AERA et al., 2014).

Concerns regarding the construct-irrelevant effects of language proficiency have also been around a long time. For example, Pressey and Pressey (1922) cautioned, "The possibility of a language handicap should always be considered in interpreting test results" (p. 67). These excerpts from Likert, and from Pressey and Pressey, were written over 80 years ago, but if the dates beside the citation were missing, most would guess they were far more recent writings. Thus, concerns over fairness are a hallmark of these a-changin' times in educational assessment, as they have been for over 80 years. Happily, the difference now is that practitioners have begun to heed the call, and improvements in test development, test administration, and test validation have made many educational assessments fairer.

Current Trend 3: International and Cross-Lingual Assessment

As the chapters by Allalouf and Hanani (Chapter 15, this volume), van de Vijver and Poortinga (Chapter 16, this volume), and Muñiz, Elosua, Padilla, and Hambleton (Chapter 17, this volume) illustrate, testing people who operate in different languages is becoming increasingly common. The world is becoming smaller, but cultural

identity remains strong. Given that language is intimately linked with culture, a global community necessitates a multilingual community. For educational assessments to be useful in a global society, they must, in some sense, transcend language. The need for cross-lingual assessment has made test translation (adaptation) popular, but as the aforementioned chapters attest, it is difficult to ensure that we are measuring the same constructs with equal precision and utility across languages. The *Guidelines* proposed by the International Test Commission (ITC), described by Muñiz and colleagues (Chapter 17, this volume), are extremely helpful in identifying the issues in test development, administration, and validation to be considered in cross-lingual assessment. Emerging adaptation tools, such as those described by Allalouf and Hanani (Chapter 15, this volume) are also encouraging. However, as van de Vijver and Poortinga (Chapter 16, this volume) remind us, we need to be aware of what is lost in cross-lingual assessment relative to assessing examinees within a single language. In our view, the hard work comes in providing cautions with respect to interpretations of cross-lingual test results, and making policymakers and others who interpret the results aware of their limitations.

Rios and Sireci (2014) pointed out that although the ITC *Guidelines* are helpful, they are not enforced, and much of the published literature in cross-lingual assessment makes no reference to them. Thankfully, international educational assessment practices seem better, with the major assessment programs (e.g., Programme for International Student Assessment [PISA], Progress in International Reading Literacy Study [PIRLS], Third International Mathematics and Science Study [TIMSS]) devoting considerable time and resources to test adaptation. Nevertheless, comparisons across countries' performance on these assessments are often made without considering their limitations (Ercikan, Roth, & Asil, 2015). With respect to cross-lingual assessments, it appears the practices are outpacing the validation to support them. Like the use of educational tests for accountability, the lack of balance across use and validation is troubling.

Current Trend 4: Using Technology to Improve Assessment

Way and Robin (Chapter 11, this volume) provided a comprehensive review of the history of computer-based assessment, and Mills and Breithaupt (Chapter 12, this volume), von Davier and Mislevy (Chapter 14, this volume), and Zenisky and Luecht (Chapter 13, this volume) described exciting new developments in this area. It is hard to keep up with developments in computer-based testing because the technology evolves faster than we can write about it. Currently, technology is being used to determine the most appropriate sets of items to administer to individual examinees (i.e., computerized-adaptive testing), to expand what we can measure beyond what is possible in a paper-based environment (e.g., research skills; Mills & Breithaupt, Chapter 12, this volume); to embed video and other media into assessments; and to engage, motivate, and accommodate specific subpopulations of examinees (e.g., students with disabilities).

Technology has always been a big part of large-scale educational assessment. In fact, Reynold Johnson's invention of the scanner that could read and score "bubble sheets" in 1931 made large-scale assessment possible and popularized the multiple-choice item. However, of the five current trends we identified, using technology to improve assessment stands out as the one that is most a-changin'. Educational assessments may lag far behind the use of technology in other areas—such as entertainment or surveillance—but current tests that incorporate technology are very different from 20th-century tests. A key development in technology-enhanced assessments is the ability of the computer to "learn" something about an examinee. Adaptive testing is one example, whereby the test adjusts its difficulty to best match the estimated proficiency of an examinee. But the computer can also be used to provide supports such as encouragement or accommodations, if it "senses" that the examinee needs them. We are impressed with current developments in computer-based assessments (e.g., see Mills & Breithaupt, Chapter 12, this volume), and we anticipate further benefits of incorporating technology into assessments, such as integrating instruction and assessment, and making tests more "fun" for examinees by giving them more control over the assessment experience (e.g., choosing avatars to represent themselves, pausing the assessment to access tutorials; see Drasgow, 2015, for further descriptions of technology-enhanced assessment).

Current Trend 5: Improved Score Reporting and Diagnostic Assessment

As the use of testing increases, focus has also increased on how scores from assessments are interpreted and reported for consumption by various audiences. This interest has been driven by examinees and test users (e.g., institutions, professional organizations, educators) alike, all of whom have a vested interest in a clear understanding of what test scores mean about examinee ability. As Zenisky, Mazzeo, and Pitoniak (Chapter 20, this volume) discuss, the practice of score reporting matured considerably in the last quarter of the 20th century, with assessment developers attending increasingly to design and audience as they sought to report test information in ways that are clearer, more useful, and more tailored for a variety of users.

Like assessment, score reporting also stands to benefit from the transition to digital platforms, where navigation tools and dynamic presentation methods make it possible for users to call up additional information to help them understand and interact with scores in real time based on their particular interests. Zenisky and colleagues (Chapter 20, this volume) also documented how the National Assessment of Educational Progress (NAEP) program took advantage of digital and online functionalities to enhance the breadth, depth, and utility of information reported from those assessments. In this century, online score reporting is likely to become increasingly common, as more K–12 summative and formative assessments are administered and scored on digital platforms.

A related area of interest is extracting and communicating more meaning from test scores about examinees' strengths and weaknesses. Particularly for examinees who score poorly, this type of diagnostic information is often cited as an important tool to help them improve their future performance. Providing such information touches on not only score interpretation and report design, but also on assessment design itself. As Ackerman and Henson (Chapter 9, this volume) note, diagnostic classification models (DCMs) typically require an assessment that taps multiple abilities or dimensions, which could be built and scored using advanced techniques such as multidimensional item response theory (MIRT). These types of models have garnered increasing interest and research in the 21st century as psychometricians work to develop assessments that can produce reliable, valid, and usable diagnostic information that can be reported to students, instructors, parents, examinees, and other stakeholders. However, it should be noted that the "new" idea of MIRT dates back at least 40 years to the work of Mulaik (1972) and Reckase (1972), and has its origins in the work of Spearman (1904) and Thurstone (1935).

Current Trend 6: Embedded and Formative Assessment

In the first quarter of the 21st century, one of the most "futuristic" phrases one can utter is *stealth assessment*. This approach, in which assessment items are embedded within a nontest context (e.g., a game, a lesson, a simulation) allows test users to collect information about examinees in ways that purport to be authentic and noninvasive. Such assessments are often presented as a promising tool for instructors and learners alike, as they may be able to provide finer-grained details about an examinee's progress and ability in a shorter time frame, and via means that are more instructionally relevant and actionable.

As with score reporting and diagnostic assessment, technological advances have made stealth assessment seem like a plausible reality only recently. In particular, computers offer a convenient way to surreptitiously collect examinee process and response information without announcing that this is happening—thereby creating the potential to reduce testing time and anxiety, and boost examinee engagement and task authenticity.

Stealth assessment does present a number of measurement challenges, but advances in these areas have also proliferated since the turn of the 21st century, as von Davier and Mislevy discuss (Chapter 14, this volume). The vast amounts of process data produced by stealth assessment require new and more flexible measurement models that, in particular, are flexible or agnostic with respect to dimensionality and error distribution. Even prior to the use of such models, examinee actions in simulation- and game-based assessments must be meaningfully parceled and assigned with values; this process also requires advanced measurement and computing, as von Davier and Mislevy discuss.

Stealth assessment sounds, in many ways, like one area of assessment that truly is new and futuristic. Without the technological and methodological bells and whistles

of the 21st century, however, stealth assessment is really a combination of two relatively "old" ideas in assessment: formative assessment and embedded assessment.

Constancies

Aside from technology-enhanced assessment, the other trends we identified as "current" have historical roots that, once realized, make them seem long-standing. Thus, with respect to using tests for accountability purposes, for cross-lingual/cross-cultural assessment purposes, and for improving instruction, it may seem more like the assessment song has remained the same, rather than the testing times are a-changin'. One reason for these constancies is that the fields of education and psychology, from which the science of psychometrics emerged, have long histories of concern over validity.

Since the earliest days of modern testing there has been a great awareness of the limitations of tests and the need to be explicit about what test scores represent, and what they do not represent. There have been debates for sure, with some proclaiming that tests can do more than what validity evidence might suggest, and others who claim that test scores held little utility. But these debates led to progress. For example, due to debates over what *validity* means and how tests should be validated, the three major professional associations most involved in testing practices and research came together to create standards for test development, administration, and interpretation.

The first version of the joint testing standards (APA, 1954) stated, "Validity information indicates to the test user the degree to which the test is capable of achieving certain aims" (p. 13)—which is similar to the AERA and colleagues (2014) edition (which represents the sixth version of these joint *Standards*) that describes validity as the degree to which evidence and theory support the use of a test for a particular purpose. A review of the evolution of these *Standards* indicates that the educational and psychometric professions have consistently demanded evidence that a test measures what it purports to measure and that the use of test scores is justified based on evidence that the test scores are appropriate for that use (Sireci, 2009). To illustrate this constant concern regarding the need to justify test use with evidence, we present two quotes on this topic. These two quotes were written about 60 years apart. We challenge readers to identify which is the more recent writing.

> Validity is not a property of the test. Rather, it is a property of the proposed interpretations and uses of the test scores. Interpretations and uses that make sense and are supported by appropriate evidence are considered to have high validity (or for short, to be valid). . . .
>
> No [technical] manual should report that "this test is valid." . . . The manual should report the validity of each type of inference for which a test is recommended. If validity of some recommended interpretation has not been tested, that fact should be made clear.

The first quote is from Kane (2013, p. 3); the second is from APA (1954, p. 19). A review of the history of validity reveals fundamental constants (Sireci, 2009, 2015) that apply not only to traditional test score interpretations that pertain to individuals, but also to aggregations of test scores that are used for accountability purposes.

Another constancy in the field of educational testing is criticism of that very testing. It is important for the testing community to embrace criticisms because they may point to deficiencies in tests or in testing policies that could be addressed, and thereby improve educational assessments. However, like the other "current events" in educational testing, criticisms of tests are not new. What might the most common criticisms be? We could review current newspapers and education blogs, or perhaps revisit this list of criticisms put together by Odell (1928):

I. Examinations are injurious to the health of those taking them, causing overstrain, nervousness, worry, and other undesirable physical and mental results.
II. The content covered by examination questions does not agree with the recognized objectives of education, but instead encourages cramming, mere factual memorizing, and acquiring items of information rather than careful and conscientious study, reasoning, and other higher thought processes.
III. Examinations too often become objectives in themselves, the pupils believing the chief purpose of study is to pass examinations rather than to master the subject or gain mental power.
IV. Examinations encourage bluffing and cheating.
V. Examinations develop habits of careless use of English and poor handwriting.
VI. The time devoted to examinations can be more ably used otherwise, for more study, recitation, review, and so forth.
VII. The results of instruction in the field of education are intangible and cannot be measured as can products in industry.
VIII. Examinations are unnecessary. Capable instructors handling classes which are not too large are able to rate the work of their pupils without employing examinations.

Current critics may use newer terms such as *teaching to the test,* but it is remarkable how the spirit of the criticisms raised in 1928 is essentially the same as those raised in 2015.

Concluding Remarks

In this chapter, we reviewed current trends in educational testing and illustrated that many of them have roots deep into the earliest days of modern testing. Our review brings the adage "The more things change, the more they stay the same" to mind.

However, as the collection of chapters in this book illustrates, great progress has been made in the science and practice of educational measurement. On the science side, new measurement models such as developments in IRT and cognitive modeling have improved the precision and efficiency of our measures and enhanced test score interpretations. On the practice side, concerns for testing fairness are now incorporated into test development, and are manifested through statistical procedures such as differential item functioning (DIF) analysis, quality-control procedures such as sensitivity review, and through validation efforts such as differential predictive validity and measurement invariance studies. In addition, test administration conditions have become more flexible, and great strides have been made in the area of providing test accommodations to examinees who need them, while maintaining fidelity to the construct measured (Abedi & Ewers, 2013).

Although we acknowledge the great strides made in measurement theory and practices, our historical review also points to areas greatly in need of improvement. First, the use of tests for multiple purposes has far outpaced validity studies to evaluate or justify such use. Therefore, we recommend that (a) much more research be conducted on derivative measures such as value-added estimates and "growth" percentiles used for accountability purposes, and (b) these and other newer metrics *not* be used until there is a substantial research base to support such use. Although the amount of research in this area may seem daunting, Cronbach (1988) noted that if we all work together, we can make great progress. As he put it:

> Fortunately, validators are also a community. That enables members to divide up the investigative and educative burden according to their talents, motives, and political ideals. Validation will progress in proportion as we collectively do our damnedest—no holds barred—with our minds and our hearts. (p. 14)

Another area where more progress is needed is in measuring the academic progress of students. Twenty-first century educational tests should be able to quantify how much a student has learned over the course of a school year, but we seem to have great trouble doing so. Keller, Colvin, and Garcia (Chapter 19, this volume) provide one example of research needed in this area, but clearly more needs to be done.

A third area where we anticipate greater progress in the near future is using technology not only to improve educational assessments, but also to make the assessment experience more enjoyable for examinees. Although a "fun test" may presently seem like an oxymoron, technology can be used to make tests more personal for examinees (e.g., choose an avatar, scene, decide when to pause), and the opportunity for gamification to make tests more engaging, and to integrate them with instruction, is strong.

In closing, we note that the field of educational measurement has an interesting history that is founded on a concern for the legitimacy of what we are measuring, appropriate due process for examinees, and evaluation of the degree to which the goals of the assessment are met. Much progress has been made, critical aspects of validity and fairness have endured, and the field remains one with many interesting problems to tackle. Thankfully, as the contributors to this volume and those who attended the

Ronference illustrate, there are many creative and talented people in our community to move this field forward. We thank our friend and colleague, Professor Ron Hambleton, for enlightening us on the important problems in educational measurement, and encouraging us to solve them. One "song" that has remained the same over the last 40 years is Ron's commitment not only to educational measurement, but also to the mentorship of us all. For that, we remain grateful.

REFERENCES

Abedi, J., & Ewers, N. (2013). *Accommodations for English learners and students with disabilities: A research based decision algorithm*. Smarter Balanced Assessment Consortium.

American Educational Research Association. (2000, July). AERA position statement: High-stakes testing in preK–12 education. Retrieved February 19, 2015, from *www.aera.net/AboutAERA/AERARulesPolicies/AssociationPolicies/PositionStatementonHigh-StakesTesting/tabid/11083/Default.aspx*.

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

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2, Suppl.).

Braun, H. (2013). Value-added modeling and the power of magical thinking. *Applied Measurement in Education*, *21*, 115–130.

Chudowsky, N., & Chudowsky, V. (2005, March). *Identifying school districts for improvement and corrective action*. Washington, DC: Center for Education Policy. Retrieved January 28, 2006, from *www.ctredpol.org/nclb*.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Drasgow, F. (Ed.). (2015). *Technology and testing*. New York: Routledge.

Ercikan, K., Roth, W.-M., & Asil, M. (2015). Cautions about inferences from international assessments: The case of PISA 2009. *Teachers College Record, 117*(1), 1–28. Retrieved March 3, 2015, from *www.tcrecord.org*.

Forte, E., & Faulkner-Bond, M. (2010). *The administrators' guide to federal programs for English learners*. Washington, DC: Thompson.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 44–53.

Mulaik, S. A. (1972, July). *A mathematical investigation of some multidimensional Rasch models for psychological tests*. Paper presented at the annual meeting of the Psychometric Society, Princeton, NJ.

Odell, C. W. (1928). *Traditional examinations and new-type tests*. New York: Century.

Pressey, S. L., & Pressey, L. C. (1922). *Introduction to the use of standard tests: A brief manual in the use of tests of both ability and achievement in the school subjects*. Yonkers-on-Hudson, NY: World Book.

Reckase, M. D. (1972). *Development and application of a multivariate logistic latent trait model*. Unpublished doctoral dissertation, Syracuse University, Syracuse, NY.

Rios, J. A., & Sireci, S. G., (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing, 14*, 289–312.

Ruch, G. M. (1929). *The objective or new type examination*. Chicago: Scott Foresman.

Santayana, G. (1905). *The life of reason: Reason in common sense*. New York: Scribner's.

Sireci, S. G. (2009). Packing and upacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). Charlotte, NC: Information Age.

Sireci, S. G. (2015). On the validity of useless tests. *Assessment in Education: Principles, Policy, and Practice*.

Spearman, C. (1904). General intelligence: Objectively determined and measured. *American Journal of Psychology, 15*, 201–293.

Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. Chicago: University of Chicago Press.

Wells, C. S., Sireci, S. G., & Bahry, L. (2014, April). *Estimating the amount of error in student growth percentiles*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia.

White House. (2002). *No child left behind*. Retrieved January 30, 2006, from *www.whitehouse.gov/news/reports/no-child-left-behind.html*.