Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Spring 2021, Vol. 40, No. 1, pp. 7–16

NCME Presidential Address 2020: Valuing Educational Measurement

Stephen G. Sireci, *University of Massachusetts Amherst*

Abstract: *The community of educational measurement researchers and practitioners has made many positive contributions to education, but has also become complacent and lost the public trust. In this article, reasons for the lack of public trust in educational testing are described, and core values for educational measurement are proposed. Reasons for distrust of educational measurement include hypocritical practices that conflict with our professional standards, a biased and selected presentation of the history of testing, and inattention to social problems associated with educational measurement. The five core values proposed to help educational measurement serve education are: (a) everyone is capable of learning; (b) there are no differences in the capacity to learn across groups defined by race, ethnicity, or sex; (c) all educational tests are fallible to some degree; (d) educational tests can provide valuable information to improve student learning and certify competence; and (e) all uses of educational test scores must be sufficiently justified by validity evidence. The importance of these core values for improving the science and practice of educational measurement to benefit society is discussed.*

Keywords: history of testing, psychometrics, standardization, test bias, validity, values

I began my 2020 Presidential Address for the *National Council on Measurement in Education* (NCME) by stating my belief that educational measurement is an altruistic profession. Psychometricians and other educational measurement specialists are highly trained in statistical analysis, data management, research design, and evaluation. These are highly valued skills in the more lucrative business world, but we chose to work in education. Why? I believe it is because we value the contributions we make to society through improving the science and practice of educational measurement. After all, that is the explicit mission of our beloved organization.[1] Why then, I asked the audience, is there so much public outcry *against* educational testing?

In this article, I provide my answers to that question. Those answers illustrate that, as a community of measurement specialists and practitioners, we have held on too long to outdated notions of the way educational tests should be developed, administered, used, and validated. Fortunately, as in my Presidential Address, I do not end with criticisms. Rather, I propose steps we can take to restore public faith in educational testing so we can accomplish the radical goals that can be achieved when educational tests focus on promoting the success of *all* students. The first step on that path is the development of core values for our profession.

My main thesis in this article is if we want the public to *value* educational tests, and if we want educational tests to *have* value in helping students learn, then we must establish professional values to support those goals. Although there has been important work to illustrate the ubiquity of values in educational assessment (e.g., Messick, 1989a,b; Mislevy, 2018), there has not yet been a formal discussion of what values *are*, or should be, inherent in educational testing. Given that such values have not yet been articulated, I propose five values to serve as a starting point for establishing core values for the educational measurement profession. The values I propose draw from the scientific principles of measurement codified in our *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & NCME, 2014), and other seminal writings in our field. However, they also incorporate the altruistic notion that educational tests should serve education (Gordon, 2020).

Before proposing these core values, I first provide some historical context that explains how we lost the public's trust in assessment. Thus, this article has two parts. In Part 1, I describe the current lack of public trust in educational assessment and the reasons for the emergence and growth of this distrust. In Part 2, I describe what we can do about it, starting with the development of core values for educational assessment and actions to support them.

Shoot the Messenger: How Educational Tests Became the Enemy

In the United States, public protests against educational tests are vehement, frequent, and well organized. The "opt-out movement" resulted in up to 90% of public school students in some districts refusing to take the statewide tests mandated under No Child Left Behind (NCLB) and the Every Student Succeeds Act (ESSA), and these refusals were supported by parents and teachers (Bennett, 2016; Marland, Harrick, & Sireci, 2019). Examples of these grassroots and

[1]See https://www.ncme.org/about/mission.

well-coordinated efforts can be found across the states. For example, in Massachusetts, the Massachusetts Teachers Association, which happens to be the union into which I pay dues each month, has a web page devoted to this topic that includes *A Parent's Guide to Opting Out of State Standardized Tests* and a section "What Else Can I do to Support Less Testing, More Learning?[2]" Similarly, the organization, *Citizens for Public Schools* has a fact sheet for parents on how to opt out of state testing, and urges parents to do so to "protect your child."[3] This website links to the website of *Parents Across America*, which ardently claims standardized tests are "harming our children's mental health" and that the "damage to our children is overwhelming." How can educational measurement be an altruistic profession if we are damaging children?

Criticisms that tests cause great harm are not limited to the general public and teachers unions; they are prominent in academic discussions, too. Amy Stuart Wells entitled her 2019 Presidential Address to AERA, which has over 25,000 members, "An Inconvenient Truth about the New Jim Crow of Education." In her address, educational tests were depicted as tools to continue the oppression of students from historically marginalized backgrounds, just as the Jim Crow laws of the past were designed to oppress the educational, civic, and occupational opportunities of African Americans. Essentially, Stuart Wells argued educational tests perpetuate an unequal education system for students of color. Her address was widely applauded, and at the time of this writing, had 1,313 views on YouTube.

The criticism of test-based inequality in education is also evident in other aspects of American society. For example, New York City Mayor Bill Di Blasio pointed out that for a competitive public high school in the city with a freshman class of 1,000 students, only 10 African American students were admitted, based on the sole admissions criterion of performance on the admissions test. He exclaimed, "We don't believe in a single test as a way of making decisions," and "the problem, in fact, is the test on many levels" (National Public Radio, 2018). It is hard to disagree with his statements, and in fact, they are consistent with the AERA Position Statement on High Stakes Testing (AERA, 2000). We can talk about a lack of differential predictive validity and differential item functioning (DIF) ad infinitum, but if the adverse impact is so consequential it prohibits educational opportunities for a whole community of children, how can we justify use of the test for this purpose? How does such use benefit society—society as a whole, not solely the privileged portions of our society?

The situation is the same for high school admissions tests in other cities (e.g., Boston, see Vanznis, 2018), but perhaps most alarming to our field is the recent decision by the University of California System to prohibit use of the ACT and SAT as criteria for admission to their universities due to concerns over adverse impact against Black, Latino, and other underrepresented groups of students (University of California Office of the President, 2020). Before this decision, ACT and SAT scores were annually used in admissions decisions for the over 200,000 students applying to these universities. This decision to stop using these scores was made in spite of a Blue-Ribbon Standardized Testing Task Force (STTF) assembled by the University of California Board of Regents that recommended against immediate elimination of the college admissions testing requirement (STTF, 2020). Although the College Board and ACT may have been shocked by this decision, they should not have been. Complaints about the adverse impact of SAT scores against Black and Latinx applicants from the University of California began at least two decades ago (Gehring, 2001), but the response was consistently muted, focusing on defense of the current tests based on predictive validity evidence, rather than on considering how admissions testing can be changed to reduce adverse impact.

Criticisms of educational tests are not limited to the United States. In Chile, the national university admissions test, the Prueba de Seleccion Universitaria (PSU), was so heavily criticized for its unfairness to less-privileged students that not only did students refuse to take the tests, which postponed the annual test administration twice, protesters broke into the national testing office, stole test booklets, and burned them in the streets (Ramos, 2020)!

I could provide further examples of the public outcry against educational tests in the United States and abroad, but I believe the point has been made—the public (broadly defined) has lost trust in the validity, credibility and utility of educational tests. The question remains, "Why?" In the next section, which borrows the title of my Presidential Address (Sireci, 2020b), I provide my answers to that question.

Psychometricians in the Hands of an Angry Mob

As I view the situation, there are at least four reasons why the public has lost confidence in educational assessments. These are (a) measurement professionals are often hypocritical (i.e., we impose standards, but don't follow them); (b) we present a censured history of educational and psychological testing to ourselves and the public, but the public knows better; (c) we focus on what was important 100 years ago, rather than what is important today; and (d) we are entrenched in a culture of distrust. Each reason is discussed in turn.

Psychometric hypocrisy. Why do I say we are hypocritical in that we impose standards, but don't follow them? Although there are many examples, for the sake of brevity I will provide two: (a) the current widespread use of student growth percentiles (SGPs) in the absence of validity evidence to support them, and (b) nonexistent or superficial validity arguments to support test use. These examples contradict the professional standards we developed to guide our field.

Unvalidated uses of SGPs. As the AERA et al. (2014) *Standards* state, "accountability indices based on aggregates of students' test scores "should be subjected to the same validity, reliability, and fairness investigations that are expected for the test scores that underlie the index" (AERA, APA, & NCME, p. 210). I believe that is an important standard. SGPs are used in about half the states in the United States. These uses include (a) reporting children's "growth" to parents on students' score reports, (b) serving as a major criterion for teacher evaluation, and (c) developing school improvement plans (Clauser, Keller, & McDermott, 2016; Sireci & Soto, 2016; Wells & Sireci, 2020).

What do investigations of the "validity, reliability, and fairness" of SGPs tell us? Focusing only on published research involving real or simulated data, there are little empirical data to support the use of SGPs. With respect to using

[2]https://massteacher.org/current-initiatives/high-stakes-testing/opting-out-of-high-stakes-testing/a-parents-guide-to-opting-out-of-state-standardized-tests.

[3]https://www.citizensforpublicschools.org/the-facts-on-opting-out-of-mcas-or-parcc/.

aggregates of students' SGPs to evaluate teachers, Soto (2013) found that teacher classifications based on their students' median SGPs were not consistent across different samples of students. In another study, Lash, Makkonen, Tran, and Huang (2016) used generalizability theory to evaluate the stability of teacher-level SGPs and concluded more than half the variance in teacher scores was random.

Other studies have come to similar conclusions. For example, McCaffery, Castellano, and Lockwood (2015) concluded teachers' SGP indices were systematically biased such that the most effective teachers were likely to have SGP scores lower than they should, and the least effective teachers were likely to have SGP scores higher than they should—the exact opposite of the intent of teacher evaluation (a finding confirmed by Castellano & McCaffrey, 2017). At the student level, McCaffery et al. (2015) concluded, "Students with [observed SGPs] = 50, who may be classified as having typical normative growth, have about an 80% chance of having a true SGP anywhere from 20 to 80!" (p. 16). Craig Wells and I reported similar results (Wells & Sireci, 2020).

It is hard to imagine validity or fairness in the face of such unreliability. In response to a call to abandon SGPs (Sireci, Wells, & Keller, 2016), the National Center for the Improvement of Educational Assessment (NCEIA), the organization from which SGPs were created and are promoted, critically refuted our points (Betebenner, DePascale, Marion, Domaleski, & Martineau, 2016), but cited only one validity study to support the use of SGPs. That study, which focused on the use of SGPs in teacher evaluation, was an unpublished study by Briggs, Dadey, and Kizil (2014). We were unaware of this study at the time we wrote our policy brief, but I have since reviewed it and the subsequent published version (Briggs & Dadey, 2017). This research compared data on principals' ratings of teachers, teachers' scores from a classroom observation protocol, and teachers' mean SGPs (all data were from Georgia). The results suggested some congruence among the three measures, particularly for teachers rated highest and lowest by the principals. However, one study does not equate to a compelling argument for the validity of a measure used for high stakes decisions across the United States, especially in the face of multiple studies that suggest otherwise. Furthermore, there are still no studies to support the validity of reporting SGPs at the student level. Given that our *Standards* state, "A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation" (AERA et al., 2014, p. 23), *our tolerance of the widespread use of SGPs for evaluating students, teachers, and schools in the absence of a body of evidence to support such use is hypocritical*.

I invite the NCEIA to commission a comprehensive validity research agenda to justify their championing of SGPs in "more than two dozen states" (Betebenner et al., 2016, p. 3). Similarly, I encourage all testing companies using SGPs or setting "growth targets" to commission research to illustrate these metrics have a positive effect on instruction and student learning, as they claim to do. It is time for our hypocrisy to end. If we want to have standards for validation research, we must adhere to them whenever indices derived from test scores are reported and used.

Nonexistent or superficial validity arguments. A second example of psychometric hypocrisy is the lack of comprehensive programs of validation for many educational tests. Our *Standards* require that "Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided" (AERA et al., 2014, p. 23). Furthermore, they state "A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (p. 21). I am proud of these professional standards because they stipulate that tests should demonstrate their utility before being used. In a sense, these standards attempt to protect the public against faulty or inefficient tests the same way the Food and Drug Administration attempts to protect the public against dangerous or ineffective products.

Although there are outstanding examples of comprehensive validity arguments for many testing programs (e.g., ACT, 2000; College Board, 2017; Smarter Balanced, 2018), these examples tend to be the exceptions, rather than the rule. For example, it was embarrassing for me when I recently gave my "Principles of Testing" class an assignment to review the validity evidence for the *Massachusetts Comprehensive Assessment System* (MCAS, the statewide summative tests used in Massachusetts public schools). The three-page validity chapter in the *Technical Manual* had one paragraph on four of the five sources of validity evidence (Massachusetts Department of Elementary and Secondary Education [MSDE], 2019). The entire section on "validity evidence based in relationships to other variables" was two sentences, which read,

> Massachusetts has accumulated a substantial amount of evidence of the criterion-related validity of the MCAS tests. This evidence shows that MCAS test results are correlated strongly with relevant measures of academic achievement. (p. 74)

In digging deeper to find this "substantial amount of evidence," I looked to the previous version of this Technical Manual (MSDE, 2017). In that version, the same verbatim text appeared in the validity chapter, but an intriguing sentence was appended: "Specific examples may be found in the *2007 MCAS Technical Report*" (p. 81, emphasis in original). Ever the researcher, I found the 2007 Technical Report in which the substantial evidence was reported. There was no additional evidence reported, but two studies were cited, both of which were conducted in 1999 (Gong, 1999 and Thacker & Hoffman, 1999 cited in MSDE, 2007, p. 214). Thus, in the 20 years since the studies were conducted, MSDE offered no other validity evidence to justify use of the MCAS. Considering that the high school graduation requirement for the MCAS did not kick in until 2003, the relevance of these older studies is limited. Not only is the lack of a comprehensive validity argument for the MCAS contradictory to the AERA et al. (2014) *Standards*, it is troubling that the documentation uses a 2007 citation to describe studies written up in 1999. In searching for validity arguments for other testing programs, I have encountered similar disappointments (Sireci, Meng, Yoo, & Zenisky, 2009). These disappointments have convinced me we are often hypocritical in imposing standards, but not following them. This problem is particularly likely to occur when the business marketing goals of a testing program collide with the psychometric goal of having sufficient validity evidence to endorse use of a test for a particular purpose.

Psychometric censorship. As a student and Professor of educational measurement, I have long studied our history, with a focus on facilitating validity (e.g., Sireci, 1998, 2009, 2013, 2016). In my validity and scaling classes, I discuss the works of early pioneers in the field such as Weber, Fechner, Galton, Pearson, Spearman, and Thurstone. I have several prepared lectures on the contributions of these scientists and use them to trace the history of modern measurement and statistical techniques in validation. Only recently have I come across the darker side of our history. What I recently learned is I have helped perpetuate a censured version of the history of educational and psychological testing that ignores its central role in the Eugenics movement (eugenics is the science of selecting desired heritable characteristics to improve future generations, typically in reference to humans; Wilson, 2019), and its negative effects on historically marginalized populations.

An example of the history I never learned in graduate school can be found in the writings of Louis Terman. In my classes I previously depicted Terman as somewhat of a hero in that he was the architect of the Stanford-Binet, a pioneer in the large-scale testing movement, and directly responsible for the advent of the SAT. In fact, according to APA he is the 72nd-most cited psychologist in history.[4] However, I recently learned he was also a proponent of using intelligence tests to sterilize "feeble-minded" people. As he wrote,

> …in the near future intelligence tests will bring tens of thousands of these high-grade defectives under the surveillance and protection of society. This will automatically result in curtailing the reproduction of feeble-mindedness and in the elimination of an enormous amount of crime, pauperism, and industrial inefficiency (Terman, 1916, pp. 6–7).

The idea that we can administer tests for the purpose of deciding whether someone should be sterilized flies in the face of a free and democratic society. Yet through the 1990s, over 60,000 Americans were sterilized based on the results of intelligence testing (Radiolab, 2019). When considering 79% of the Italians who immigrated to America who were tested at Ellis Island using the Goddard's intelligence tests were classified as feeble-minded (Gould, 1996), I feel fortunate to have been born!

Another untold story in the history of testing is the use of citizenship tests to deny African Americans the right to vote. For example, in 1896, Louisiana had 130,334 registered black voters. Eight years later, only 1,342, 1%, could pass the state's new voting tests. Louisiana was not alone in their use of these tests, which were enormously effective in suppressing the rights of Black people in the United States (Sireci & Randall, in press).

The history of educational testing also features channeling minorities into "special education" tracks that resulted in substandard education that limited their opportunities. Culturally inappropriate intelligence tests, such as the WISC, were used so often to place African American children in California into special education classes that a class action lawsuit (*Larry P vs. Riles*, *1979*) resulted in the prohibition of intelligence testing of African American children in California public schools that exists to this day.

Thankfully, these unjust testing practices are no longer in place—at least not explicitly. However, prohibition of these practices is not due to NCME or other measurement advocates protesting against them. Rather, these practices were outlawed by the courts after civil suits were brought against testing agencies (Sireci & Parker, 2006).

One way to somewhat rectify our censured history is to call attention to the academic contributions of educational researchers and psychometricians from cultural groups that have been underrepresented in our field. For example, there are many articles I recently discovered from non-White authors that contain valuable lessons for our field, but never made it onto the reading lists in the courses I took in graduate school, or in the courses I teach. Five publications I now recommend to students and colleagues to broaden their perspectives on fairness issues in educational assessment are Dixon-Roman (2020), Dixon-Roman, Everson, and McArdle (2013), Gordon (2000), Helms (2006), and Johnson (2000). These publications are authored or coauthored by researchers from cultural groups that are underrepresented in educational measurement, and help us understand how we can improve the validity of our educational measures by revisiting, (a) our notions of bias, (b) our test development and validation practices, and (c) the role of tests used in education.

I invite all of you, especially my university colleagues, to add to this list of recommended articles to help address the dominant White culture that has permeated our field, and if continued, will prevent us from understanding and acknowledging the diversity of talent in our community. As Johnson (1980) warned, "Failure to recognize the validity of the important issues as defined by highly trained minority researchers negates the value of the extensive training which they have received and extends the traditional racist concept of the ethnocentric validation of the importance of ideas to still another arena" (p. 262).

Psychometric paralysis. I use the term "psychometric paralysis" to summarize my criticism that we focus on what was important 100 years ago, rather than what is important today. What I mean by this term is we are stuck in the 19th-century practices from which our field emerged and have not been responsive to new understandings that enhance the science and practice of educational measurement. Three examples of this paralysis are (a) blind adherence to standardization, (b) overreliance on comparability, and (c) "psychometric deafness" to calls for change.

Standardization in educational testing refers to ensuring uniform conditions with respect to test content, test administration conditions, and scoring. By ensuring such uniformity, we provide a level playing field (i.e., standard conditions) for all examinees so the testing experience does not advantage or disadvantage anyone. Thus, standardization is designed to promote fairness in testing. As I wrote in Sireci (2020c), the standardization process emerged from the psychophysical experiments that grew from Weber's experimental psychology laboratory in Leipzig Germany in the middle 19th century.

Standardization remains important today, but unlike the early psychophysicists such as Weber and Fechner, we are not using standardized procedures to justify the science of psychological measurement. Furthermore, over the past 150 years we have learned we cannot make the playing field level for all examinees through standardization because of the great variety of personal characteristics examinees bring to the testing situation. Thus, rather than being paralyzed by requests for accommodations or more flexible testing conditions, we need to adapt and allow them—particularly

[4]https://www.apa.org/monitor/julaug02/eminent.

when children taking educational tests are not competing against one another.

Let me provide one example of psychometric paralysis in this regard. I was recently involved in a court case where the plaintiff was a blind person who was denied a read-aloud accommodation on a teacher licensure test. The plaintiff wanted to be licensed so he or she could teach in a school for the blind. During test administration, the proctor denied a read-aloud accommodation request on the reading section of the exam. The proctor encouraged the plaintiff to just guess at the answers if he or she could not view the test content. In this case, standardization was seen as more important than actually measuring how well the plaintiff could comprehend written material as he or she experiences it. The psychometric paralysis exhibited by the testing agency held true to the mandates of standardization, but the costs to validity and fairness were appalling. Once the case approached litigation, common sense resulted in a settlement in favor of the plaintiff before trial. Adherence to standardization more than common sense is a reflection of psychometric paralysis.

Overreliance on test score comparability can also lead to psychometric paralysis. Comparability of test scores is certainly an important validity issue, especially when examinees are competing against one another as in admissions or employment testing. However, most educational tests today are criterion-referenced with respect to their intended purposes. In such cases, examinees are not competing against each other and strict scale score comparability becomes less important. One frustrating example is when testing agencies want strict comparability across paper-based and digitally based versions of a test. Why would we want a more technically advanced system to maintain the weaknesses of a legacy (i.e., paper) system so that scores can be strictly comparable (e.g., in the equated sense)? What needs to be comparable are the inferences derived for students across the different testing platforms. If we conclude a student is "proficient," for example, that conclusion should be valid based on whatever platform on which the test was administered. More importantly, the child should be tested on the platform on which he or she can best demonstrate that proficiency. The point is we can relax comparability if it enhances validity (see Abedi & Ewers, 2013, as one example of this point in the area of test accommodations). By considering the purposes of the test and the types of inferences that are to be drawn, we can look at comparability in a new way—one that allows flexibility, rather than paralysis. As we step further toward flexibility in test administration, we step toward empowering students as active agents in the testing process and increase their engagement.

Psychometric deafness with respect to calls for change is another manifestation of psychometric paralysis. This form of paralysis occurs when a call to action is heard, but never heeded. The previous criticisms of high school and university admissions tests are examples. There have been criticisms against admissions tests for decades, yet these tests remain largely the same—descendants of norm-referenced testing that grew out of the Eugenics movement. However, whether students have the knowledge and skills to succeed is essentially a criterion-referenced question. Rather than address this problem we have stuck with norm-referenced testing models for admissions purposes, and have confirmed they are not "biased" against underrepresented minorities using our own standards for bias (c.f., Linn, 1984). We may sleep a little better at night given that our tests do not show any differential predictive validity, and that we have screened out any large DIF items. However, how can we sleep when we are being called out for supporting a system that perpetuates the obstruction of access to higher education for so many of our children who come from Black and Brown cultural groups? Psychometric deafness may help us sleep at night in that we will not be awakened by the clamor over adverse impact in admissions or about the stress caused to children and teachers by summative assessments. However, such deafness is ignorance, and prevents us from accomplishing the NCME mission: "to advance theory and applications of educational measurement to benefit society."

An educational culture of distrust. A fourth reason we have lost public confidence is the current culture of distrust that undergirds most educational policy in the United States and elsewhere. People like to be trusted and trust builds collegiality and synergy. However, educational policies throughout the United States are rooted in a distrust of schools, teachers, and to a lesser extent, students, too. We do not trust schools and teachers to decide what to teach. We do not trust teachers' expectations, grading practices, or standards of performance. We also do not trust teachers to grade constructed-responses from students, such as essays. With respect to students, we do not trust they will not cheat. I believe statewide curriculum frameworks and protocols to prevent cheating are important. We want to ensure teachers are teaching knowledge and skills that are agreed upon to be important, and we do not want students inflating their scores by cheating or inflating others' scores by sharing their knowledge of the test. However, I think our default notion is that teachers do not know how to teach, and all students are ready to cheat. These notions make the exceptions (incompetent teachers or cheating students) the rule, and they inhibit efforts to form partnerships with teachers and students. Many of our colleagues have long argued education is most effective when curriculum, instruction, and assessment are aligned and integrated (e.g., National Research Council, 2001). For that to happen, we need to engage with students and teachers as trusted partners. The current culture of distrust in education makes us enemies, not partners. As Johnson (2000) pointed out, "The language of high standards and testing is often conveyed to the recipients of today's testing products, usually students and their teachers, in a punitive, blame-filled, and even threatening rhetoric which asserts that both have left undone what should have been done and have done what they should not have done" (p. 155).

Summary of Causes of Public Distrust

I proposed four reasons why we are experiencing well-organized and prominent protests against educational testing. These reasons fall into three general categories: (a) our hypocrisy in imposing standards, but not following them; (b) a censured history of testing; and (c) our lack of attention to what educators need, and how we are perceived by society. The cure for psychometric hypocrisy is self-evaluation. I have presented my personal evaluation of the current status of our field in the preceding section. Clearly, that is neither comprehensive nor sufficient, but hopefully, it encourages us to begin to think more critically about our assessment practices and how we interact with policy makers and the public. We need to regularly step back and evaluate

our practices and their impact on society. Noting that "value" is embedded within evaluation, I next address how we can regain public trust and elevate the status of our profession by establishing core values for educational assessment.

Core Values in Educational Measurement

Messick (1975, 1980, 1989a, 1989b, 1994) has long reminded us of the ubiquity of values in educational assessment. For example, he pointed out "[V]alidity, reliability, comparability, and fairness are not just measurement issues, but *social values* that have meaning and force outside of measurement wherever evaluative judgments and decisions are made" (Messick, 1994, p. 13; emphasis original). Values permeate all activities in educational measurement, as they do all scientific research. Values determine what we decide to measure, how we measure it, how test scores are interpreted and used, and how (whether!) decisions based on test scores are validated.

Although Messick and others (e.g., Mislevy, 2009, 2018; Shepard, 1993, 1997) have pointed out the decisive roles values play in educational testing, there has not yet been a call for the establishment of core values for our field. This lack of an agreed-upon set of values to guide our profession could be due to the very reasonable assumption that we, as a community, may not be able to agree on a set of core values. As an optimist and a believer that educational testing is a noble profession, I believe we can. Moreover, I think by establishing core values we can address the criticisms and public outcry against educational testing head on by illustrating our beliefs that educational testing can contribute to the improvement of education and student learning, and lead to a more educated and equitable society.

With these goals in mind, I propose five values I believe can serve as core values for the educational measurement profession. I do not think these values will be controversial because they are either rooted in altruism, the AERA et al. (2014) *Standards*, or both. They are,

1. Everyone is capable of learning.
2. There are no differences in the capacity to learn across groups defined by race, ethnicity, or sex.
3. All educational tests are fallible to some degree.
4. Educational tests can provide valuable information to (a) improve student learning, and (b) certify competence.
5. All uses of educational test scores must be sufficiently justified by validity evidence.

Discussions of each value follow.

Core Value 1: Everyone Is Capable of Learning

The act of educating presumes learning can and will occur. How can we work in education without believing all students can learn? Although this value does not specifically address assessment, as assessment professionals working in education, we need to affirm our belief that educational assessments are for everyone, and the first step in that affirmation is asserting education is for everyone. Essentially, this core value implies all students have the right to learn. Given that educational assessments measure learning, it is important we start with this core value.

I believe this first core value does not need to be buttressed by theory or evidence because it speaks directly to the human condition of compassion. However, I will add one supporting observation: one of the positive consequences of alternate assessments (i.e., assessments for the severely cognitively or physically disabled) is the enlightenment it gave to teachers and caretakers of these students regarding how much they could learn (Browder, Wakeman, & Flowers, 2006).

Core Value 2: There Are No Differences in the Capacity to Learn Across Groups Defined by Race, Ethnicity, or Sex

For far too long—decades upon decades—we have wasted time studying differences across racial and ethnic groups as if skin color could somehow tell us something about intelligence or capacity to learn. Entire books have been written on the topic of statistical comparisons of group test scores (e.g., Herrnstein & Murray, 1994), without anyone bothering to point out the test scores were a culturally laden, rather than an infallible, criterion for comparison. We too often get lost in the statistics without thinking about where the numbers come from. Contrariwise, Fons van de Vivjer, one of the greatest cross-cultural psychologists of our time, and his colleagues experimentally pointed out that the rank-order of cultural groups on cognitive tests could be reversed simply by using tests created to be more culturally relevant for each group (e.g., Malda, van de Vijver, & Tamane, 2010). Just as today we chuckle at the phrenologists who measured intelligence by the size of a person's skull, future researchers will chuckle that one culture pointed to lower intellectual capacity of another culture based on a test they created.

Although other evidence for the vacuousness of sex-based and race-based theories of differential intelligence are not needed, it important to point out there is far more variation in test scores *within* groups than *between* them (Gordon, Boykin, & Gunaseharan, 2019). If people differ more within groups than between them, how is information regarding group membership helpful to us for interpreting and using test scores? Clearly, we can wholeheartedly ignore research on group differences that ascribes observed differences to race, or culture. Rather, I suggest we heed the call of Helms (2006), who pleaded for us to search for the actual causes of mean group differences in test scores, rather than throwing in the towel after finding no systematic sources of bias.

It is not worthwhile to study racial/ethnic group differences in the capacity to learn. Any observed group differences in test scores are likely to reflect (a) confounding variables such as poverty, quality of education, or acculturation; (b) assessments that are differentially relevant across racial/ethnic groups; and (c) other measurement or statistical artifacts (e.g., differential restriction of range). I believe we will make more progress in improving educational assessment by assuming there are no racial, ethnic, or sex group differences in capacity to learn and spend our time in developing measures that are valid for all individuals.

Core Value 3: All Educational Tests Are Fallible to Some Degree

Although I can envision some philosophical arguments against the first two proposed core values, this one seems irrefutable and is consistent with the history of the *Standards for Educational and Psychological Testing* and virtually every other respected publication in our field. From the classical test theory perspective of an observed score being the sum of a "true score" and an "error score," to the more modern

notions of model fit, it is clear we as a field have long acknowledged educational tests are imperfect approximations of the constructs we attempt to measure. Publicly acknowledging all educational tests are fallible to some degree by instilling it as a core value grounds our work in an appropriate cloak of humility. Such humility is needed to improve assessment practices and helps inoculate us against hypocrisy.

Core Value 4: Educational Tests Can Provide Valuable Information to (a) Improve Student Learning and (b) Certify Competence

In this core value we justify the value of our profession to society. This value acknowledges the significance and power of educational assessments and gives justification to our profession. If we did not believe in this core value, why bother working in educational measurement at all? This core value also equips us with a challenging research agenda. Can we provide empirical evidence to support this claim? I believe we can—and what a noble and productive program of research it would be. By emphasizing and working on this core value we can forestall public criticism of testing by providing strong evidence of its benefits. It is encouraging to see emerging work in this area such as the work of the NCME Classroom Assessment Task Force (e.g., Wilson, Ruiz-Primero, & Paek, 2019) and other work in classroom assessment (e.g., Brookhart & McMillan,2020), as well as emerging work using diagnostic classification modeling (e.g., Madison & Bradshaw, 2018).

Core Value 5: All Uses of Educational Test Scores Must Be Sufficiently Justified by Validity Evidence

This core value should also be noncontroversial because it summarizes the essence of the *Standards* that have long guided our profession. For example, the AERA et al. (2014) *Standards* state, "Evidence of the validity of a given interpretation of test scores for a specified use is a *necessary* condition for the justifiable use of the test" (p. 11, emphasis added). Without sufficient justification for test use, how can we justify the process of testing at all? Thus, like the fourth core value, this core value demonstrates the importance of our profession. This importance is reflected in the U.S. Department of Education's (2018) peer review requirements for approving statewide testing programs.

This core value can be thought of as the battle cry for our field. It reminds us of our responsibility to evaluate and study test validity, and to signal when we have arrived at a sufficient aggregation of research results that confirms (a) a test measures what it purports to measure, (b) test scores are useful for their intended purposes, and (c) any negative effects are minimal and do not rise to a level recommending against use of the test. It is this core value that summarizes the hundreds of research presentations that fill the NCME annual meeting program each year as well as the publications that fill our journals. In short, this core value is our moral imperative, and our failure to adhere to it permeates the aforementioned reasons we have lost the public's trust. It is also important to point out that without evidence for validity to support test use, we have no evidence for the fairness of test use.

Summary of Core Values

I have proposed five core values that root our profession in the service of the public good by focusing on the NCME mission "to advance theory and applications of educational measurement to benefit society."[5] These values are rooted in the Socratic philosophy that true intelligence is being aware of how little we know, and in the moral principle of altruism. It is easy for me to state these core values should be non-controversial, and in the preceding section, I attempted to defend them. However, I know others may have different opinions on them. If we cannot agree on these five values per se, I hope they will at least have heuristic value so we can together establish core values that demonstrate our commitment to our mission.

Discussion

> "…the words 'valid' and 'value' derive from the same Latin root 'valere,' meaning 'to be strong'" (Messick, 1989b, p. 59).

I have learned a lot from the writings of Samuel Messick and from the decades of work that has gone into the seven iterations of the *Standards for Educational and Psychological Testing* (Sireci, 2020a). The preceding quote from Messick is one example. We describe validity as "the most fundamental consideration in developing tests and evaluating tests" (AERA et al., 2014, p. 11). It is interesting that the root word for validity also relates to values and to strength. In this article, I am calling for us to be strong by determining a course for our future. That course must be directed by core values—values that root our science and practices in a commitment to using educational tests for the betterment of society.

When I was elected President-elect of NCME, one of my first complaints to the Board of Directors was that the NCME mission statement was vapid. At the time our mission statement was "To advance the science and practice of measurement in education." My complaint was it sounded self-serving and did not explain why we should work to advance our field or what "advance" signified. Through working with the Board, we revised our mission statement to "The National Council on Measurement in Education is a community of measurement scientists and practitioners who work together to advance theory and applications of educational measurement to benefit society." The addition of "to benefit society" may seem like a subtle change, but it illustrates our belief that improvements in educational measurement will serve the common good. That subtle change may be the most significant improvement of our organization during my most recent tenure on the Board.

In this article, I challenged us to continue this on this trajectory—to boldly step into public debates on education and show how we can help. By equally valuing all people as learners, by acknowledging educational tests have limitations—but also strengths, and by demonstrating we are engaged in a process of research to evaluate and improve testing, we will not only regain public trust, we will be seen as important leaders in efforts to improve education for all. Is this too noble a call for our field? I think not. I think educational measurement is a noble profession; or at least it should be. It has the potential to contribute to the improvement of education—and hence society—if we focus on doing so.

A Path Forward

To restore public faith in educational testing and enable educational tests to promote the success of *all* students will

[5] https://www.ncme.org/about/mission.

involve several steps. The first is establishing core values for our field. By the time this article is in print, I will be rotating off the NCME Board of Directors. However, I will suggest the establishment of such values be an immediate priority. I hope the five values I suggested in this article represent a helpful starting point for that process.

I believe there are also other steps we should take that will be helpful for improving our public image and helping us accomplish our mission. They are: (a) ensure we enforce adherence to our AERA et al. (2014) *Standards*, particularly as they relate to the provision of validity evidence to defend test use; (b) de-emphasize norm-referenced competitiveness in educational testing except in those rare instances where examinees actually are competing for a benefit; (c) reorient our practices so that we value students more than the score scale; (d) engage with teachers and other educators to collaboratively develop tests and interpret test scores; (e) reconceptualize our notions of standardization to make tests more flexible to students' needs and funds of knowledge; (f) design test score reports for students that emphasize their strengths, rather than their weaknesses; and (g) take full advantage of technology to allow assessments to tailor themselves to the needs of each specific examinee, foster engagement in the testing process, and to be fully aligned with and integrated into instruction.

It is beyond the scope of this article to discuss each of these steps in depth, but some are described elsewhere (e.g., Sireci, 2020b, 2020c). These steps reflect what I hope is my personal research agenda for the foreseeable future. I will try not to be a victim of my own criticism and instead practice what I preach. In particular, I intend to focus on step (f) as part of my suggestion that we develop the field of *positive assessment*. This idea comes from positive psychology, which is "…a scientific approach to studying human thoughts, feelings, and behavior, with a focus on strengths instead of weaknesses, building the good in life instead of repairing the bad, and taking the lives of average people up to 'great' instead of focusing solely on moving those who are struggling up to 'normal'" (Ackerman, 2020). I also plan to incorporate the concept of "understandardization" (Sireci, 2020c) in my future test development activities.

Summary

In this article, I pointed out how we have lost public trust and have sometimes fallen off the path of serving the public good. I am not the first to point out some of our professional shortcomings (e.g., Popham, 2003). However, I am hopeful we can learn from these mistakes of the past to improve our future. I also proposed how we can improve our public image and accomplish our NCME mission by grounding our practices in core values that serve education, the science of educational measurement, and the public good. Moving us forward in the establishment of core values may be difficult in that NCME represents a great wealth of intellectual capital that includes a diversity of opinions. However, I believe it can be done, and should be done; and it is the right time to do so. In that belief, I am encouraged by the words of a much more important President who, when confronted with the frustrating obstacles toward social progress in our country, remarked,

> "what's troubling is the gap between the magnitude of our challenges and the smallness of our politics—the ease with which we are distracted by the petty and the trivial, our chronic avoidance of tough decisions, our seeming inability to build a working consensus to tackle any big problem" (Obama, 2006, p. 22)

These words resonate with me because we so often focus on solving statistical problems rather than validity problems. That focus stems from psychometric blindness. It is far easier to solve a mathematical problem than a social one, but the most useful mathematical solutions make positive contributions to society. We need to make the tough decision to build a working consensus on core values for our field for us to bring our statistical talents to the solution of the most consequential educational problems.

One of the honors of being President of NCME is publication of the Presidential Address. I am humbled to have that honor, but it is pale in comparison to the honor of working with the other members of the NCME Board of Directors over the past 3 years, and to representing this organization of approximately 1,800 members who are an amazing community of brilliant professionals who care about measurement and about students. This community has allowed me to listen to, and work with, some of the smartest people in the world. Thank you for the opportunity to briefly lead this organization. I believe we are further along the path to improve the science and practice of educational measurement to benefit society, and I hope the thoughts I have shared in this article help us to move further down that path.

Acknowledgments

This article is based on my Presidential Address for the National Council on Measurement in Education, "Psychometricians in the Hands of an Angry Mob," delivered via the Internet on September 10, 2020. I would like to thank Susan Brookhart, Joseph Rios, and the *EM:IP* editors for feedback on an earlier version of this article.

References

Abedi, J. & Ewers, N. (2013). *Accommodations for English learners and students with disabilities: A research based decision algorithm*. Smarter Balanced Assessment Consortium. Downloaded January 22, 2021 from https://portal.smarterbalanced.org/library/en/accommodations-for-english-language-learners-and-students-with-disabilities-a-research-based-decision-algorithm.pdf

Ackerman, C. E. (2020). What is positive psychology and why is it important? Downloaded January 3, 2020 from https://positivepsychology.com/what-is-positive-psychology-definition/.

ACT (2020, August). *ACT technical manual*. Iowa City: Author.

American Educational Research Association (2000, July). AERA Position Statement on High-Stakes Testing in Pre-K – 12 Education. Downloaded January 2, 2021 from https://www.aera.net/About-AERA/AERA-Rules-Policies/Association-Policies/Position-Statement-on-High-Stakes-Testing.

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

Bennett, R. (2016). Opt out: An examination of issues *(ETS Research Report Series)*. Princeton, NJ: Educational Testing Services. https://doi.org/10.1002/ets2.12101.

Betebenner, D. W., DePascale, C., Marion, S., Domaleski, C., & Martineau, J. (2016, July). *Precision, interpretability & utility of SGPs: A response to why we should abandon student growth percentiles by Sireci, Wells, and Keller*. Dover, NH: National Center for the Improvement of Educational Assessment.

Briggs, D. C., & Dadey, N. (2017). Principal holistic judgments and high-stakes evaluations of teachers. *Educational Assessment,*

Evaluation and Accountability, *29*, 155–178. https://doi.org/10.1007/s11092-016-9256-7.

Briggs, D. C., Dadey, N., & Kizil, R. C. (2014, October). *Comparing student growth and teacher observation to principal judgments in the evaluation of teacher effectiveness* [Research Report]. Boulder, CO: University of Colorado, Center for Assessment, Design, Research and Evaluation.

Brookhart, S. M., & McMillan, J. H. (2020). *Classroom assessment and educational measurement*. New York: Routledge.

Browder, D. M., Wakeman, S., & Flowers, C. P. (2006). Assessment of progress in the general curriculum for students with disabilities. *Theory Into Practice*, *45*, 249–259.

Castellano, K. E., & McCaffrey, D. F. (2017). The accuracy of aggregate student growth percentiles as indicators of educator performance. *Educational Measurement: Issues and Practice*, *36*(1), 14–27. https://doi.org/10.1111/emip.12144

Clauser, A. L., Keller, L. A., & McDermott, K. A. (2016). Principals' uses and interpretations of student growth percentile data. *Journal of School Leadership*, *26*(1), 6–33. https://doi.org/10.1177/105268461602600101.

College Board (2017). SAT® suite of assessments technical manual: Characteristics of the SAT. New York: Author.

Dixon-Román, E. (2020). A haunting logic of psychometrics: Toward the speculative and indeterminacy of blackness in measurement. *Educational Measurement: Issues and Practice*, *39*(3), 94–96.

Dixon-Román, E., Everson, H. T., & McArdle, J. J. (2013). Race, poverty and SAT scores: Modeling the influences of family income on black and white high school students' SAT performance. *Teachers College Record*, *115*, 1–33.

Gehring, J. (2001, February). UC president pitches plan to end use of SAT in admissions. *Education Week*, February 28, 2001.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice*, *39*(3), 72–78.

Gordon, E. W., Boykin, A. W., & Gunaseharan, P. A. (2019). Human diversity and its implications for pedagogy and assessment. In E. Armour-Thomas, C. McCallister, A. W. Boykin, & E. W. Gordon (Eds.), *Human variance and assessment for learning* (pp. 3–38). Chicago: Third World Press Foundation.

Gould, S. J. (1996). *The mismeasure of man*. New York: W. W. Norton & Company.

Helms, J. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist*, *61*(8), 845–859.

Herrnstein, R. J., & Murray, C. A. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

Johnson, S. T. (1980). Major issues in measurement today: Their implications for Black Americans. *Journal of Negro Education*, *49*, 253–262.

Johnson, S. T. (2000). The live creature and its expectations for the future. *Journal of Negro Education*, *69*, 150–158.

Larry P. V. Riles (1979). No. C-71-2270 RFP California.

Lash, A., Makkonen, R., Tran, L., & Huang, M. (2016). *Analysis of the stability of teacher-level growth scores from the student growth percentile model (REL 2016–104)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West.

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement*, *21*, 33–47.

Madison, M. J., & Bradshaw, L. P. (2018). Evaluating intervention effects in a diagnostic classification model framework. *Journal of Educational Measurement*, *55*, 32–51.

Malda, M., van de Vijver, F. J. R., & Tamane, Q. (2010). Rugby versus Soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence*, *38*, 582–595.

Marland, J., Harrick, M., & Sireci, S. G. (2019). Student assessment opt out and the impact on value-added measures of teacher quality. *Educational and Psychological Measurement*, *80*, 365–388. DOI: https://doi.org/10.1177/0013164419860574.

Massachusetts Department of Elementary and Secondary Education (2007). *2007 MCAS technical report*. Malden, MA: Author.

Massachusetts Department of Elementary and Secondary Education (2017). *2016 MCAS and MCAS-alt technical report*. Malden, MA: Author.

Massachusetts Department of Elementary and Secondary Education (2019). *2018 legacy MCAS technical report*. Malden, MA: Author.

McCaffery, D. F., Castellano, K. E., & Lockwood, J. R. (2015). The impact of measurement error of individual and aggregate SGP. *Educational Measurement: Issues and Practice*, *34*, 15–21. https://doi.org/10.1111/emip.12062.

Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955966.

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

Messick, S. (1989a). Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher*, *18*, 511.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 83–108). Charlotte, NC: Information Age Publishing Inc.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. New York: Routledge.

National Public Radio (2018, June). NYC mayor on diversity problems with city's elite public high schools. Downloaded from https://www.npr.org/2018/06/18/620939157/nyc-mayor-on-diversity-problems-with-citys-elite-public-high-sch ools, (accessed January 1, 2021).

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.

Obama, B. (2006). *The audacity of hope: Thoughts on reclaiming the American dream*. New York: Crown Publishers.

Popham, W. J. (2003). Seeking redemption for our psychometric sins. *Educational Measurement: Issues and Practice*, *22*(1), 45–48.

Radiolab (2019, July). *G: Unfit*. Podcast available at https://www.wnycstudios.org/podcasts/radiolab/articles/g-unfit.

Ramos, N. A. (2020, January). Chilean university admissions tests hit by fresh protests. *US News and World Report*, January 6, 2020.

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

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–8, 13.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, *45*, 83–117.

Sireci, S. G. (2009). Packing and unpacking 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 Publishing Inc.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, *50*, 99–104.

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice*, *23*, 226–235.

Sireci, S. G. (2020a) De-"constructing" test validation. *Chinese/English Journal of Educational Measurement and Evaluation*, *1*. 教育测量与评估双语季刊. Downloaded from https://www.ce-jeme.org/journal/vol1/iss1/3, (accessed January 1, 2021).

Sireci, S. G. (2020b, September). Psychometricians in the hands of an angry mob. Presidential addressed delivered at the annual meeting of the National Council on Measurement in Education [virtual]. Downloaded from https://www.youtube.com/watch?v=5X2uYYax9yw.

Sireci, S. G. (2020c). Standardization and understandardization in educational assessment. *Educational Measurement: Issues and Practice*, *39*(3) 100–105.

Sireci, S. G., Meng, Y., Yoo, H., & Zenisky, A. L. (2009). *Building validity arguments for educational testing programs*. *Center for Educational Assessment Research Report No. 729*. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Sireci, S. G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice*, *25*(3), 27–34.

Sireci, S. G., & Randall, J. (in press). Evolving notions of fairness in testing in the United States. In M. Bunch & B. Clauser (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice*. New York: Routledge

Sireci, S. G., & Soto, A. (2016). Validity and accountability: Test validation for 21st-century educational assessments. In H. Braun (Ed.). *Meeting the challenges to measurement in an era of accountability* (pp. 149–167). New York: Routledge.

Sireci, S. G., Wells, C. S., & Keller, L. A. (2016, June). *Why we should abandon student growth percentiles*. *Center for Educational Assessment Research Brief 16-1*. Amherst, MA: Center for Educational Assessment, University of Massachusetts. Downloaded from http://www.umass.edu/remp/news_SGPsResearchBrief.html.

Smarter Balanced Assessment Consortium (2018). *2017-2018 summative technical report*. Downloaded January 4, 2021 from https://portal.smarterbalanced.org/library/en/2017-18-summative-assessment-technical-report.pdf.

Soto, A. (2013). *Measuring teacher effectiveness using students' test scores*. Unpublished Dissertation, University of Massachusetts Amherst.

Standardized Testing Task Force (2020, January). *Report of the UC academic senate standardized testing task force*. Oakland: University of California Academic Senate.

Terman, L. M. (1916). *The measurement of intelligence*. Boston: Houghton Mifflin.

University of California Office of the President (2020, May). University of California Board of Regents unanimously approved changes to standardized testing requirement for undergraduates [press release]. Downloaded January 2, 2021 from https://www.universityofcalifornia.edu/press-room/university-california-board-regents-approves-changes-standardized-testing-requirement.

U.S. Department of Education (2018, September). *A state's guide to the U.S. Department of Education's assessment peer review process*. Washington, DC: Author.

Vanznis, J. (2018, October). Highly skilled black, Latino students face admissions barriers to exam schools, study finds. Boston Globe, October 1, 2018. Downloaded from https://www.bostonglobe.com/metro/2018/10/01/highly-skilled-black-latino-students-face-admission-barriers-exam-schools-study-finds/LOKwnprnVnL6XAUffuJ8yK/story.html.

Wells, C. S., & Sireci, S. G. (2020). Evaluating random and systematic error in student growth percentiles. *Applied Measurement in Education*, *33*, 349–361, https://doi.org/10.1080/08957347.2020.1789139

Wilson, P. K. (2019). Eugenics. In *Encyclopedia Britannica*. Downloaded December 12, 2020 from https://www.britannica.com/science/eugenics-genetics.

Wilson, M., Ruiz-Primero, M. A., & Paek, P. (2019). Introduction to the special issue on classroom assessment. *Journal of Educational Measurement*, *56*, 667–669.