Toward More Equitable Measurement and Assessment:
Implications for the Test Standards

American Educational Research Association Annual Meeting, San Diego, CA
April 22, 2022

https://conservancy.umn.edu/handle/11299/194887

Abstract

The Standards for Educational and Psychological Testing is the gold standard guiding the "sound and ethical use of tests" (Plake & Wise, 2014, p. 2). While fairness is a core principle, is the Standards' current conception sufficient to meet today's needs? The COVID pandemic and Black Lives Matter have underscored longstanding disparities and opportunity gaps. Demands for equity are pervasive; growing research documents a range of cognitive, psychological, social and cultural factors that mediate performance; and backlash against testing and its role in maintaining the status quo is growing.

This interactive symposium explores how the Standards may need to change to address current demands for equity and fairness. Feedback from the symposium will contribute to the Standards' upcoming revision.

Summary

According to the 2014 Standards, a fair test "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 of disadvantage some individuals because of characteristics irrelevant to the intended construct" (AERA, APA, NCME, 2014, p.50). This idea of equivalence of construct meaning– what is being measured – permeates the overall fairness standard and the 20 specific standards that comprise the chapter. The overall standard presents the guiding principle of the chapter and infuses all operational chapters.

Standard 3.0: All steps in the testing process, including test design, validation development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population.

The specific standards are organized into four clusters dealing with test design, development, administration and scoring; validity of test score interpretations; accommodations to support valid score interpretations; and safeguards again inappropriate score interpretations for intended uses.

The Standards make clear that fairness cannot be an afterthought, but rather is a fundamental principle that must guide all stages of testing, beginning with test design and ending with score interpretations and use. Still, the Standards define fairness in terms for which measurement

professionals can reasonably be accountable. They focus on measurement bias as a central theme, but also include attention to accessibility, the concept that "…all test takers should have an unobstructed opportunity to demonstrate their standing on the construct(s) being measured." (AERA, APA, NCME, 2014, p. 49); universal design, meaning design that takes into account the characteristics of all members of the intended population (see Thompson, Johnstone, & Thurlow, 2002), and the reality that test adaptations may be needed to assure fairness for some individuals.

The measurement orientation of the fairness standards has been criticized as too narrow, as has the Standards' indeterminate stance on the role of consequences in test development and use. Recent research on socio-cultural learning and the role of cultural relevance also give pause. What, if any, changes might be needed for the upcoming revision of the Standards to strengthen its approach to equity and fairness? This is the central question addressed by the symposium.

The symposium will start with an overview of the Standards revision plan. Then three prominent researchers specializing in equity will present brief opening remarks (6-8 minutes) on what changes are needed in the Standards approach to fairness and why. Symposium participants will then have the opportunity to raise questions and provide feedback. The session chairperson will also serve as discussant to conclude the session

Rodriguez, M.C. (2022, April 22). *Toward more equitable measurement and assessment: Implications for the Test Standards* [Invited session facilitated by F. Worrell]. American Educational Research Association Annual Meeting, San Diego, CA.

We have standards – that is a good thing. We have a full chapter on Fairness in the *Standards for Educational and Psychological Testing* – and that is a good thing.

The *Standards* state that a fair test "reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population" (p. 50). What it doesn't do is address the justice of such outcomes – is the test ready to provide the kind of interpretation to support uses that will meet the needs of diverse communities? This is what we are hearing from scholars to meet the interests and goals of diverse communities.

The word justice never appears in the *Standards*.

The *Standards* could require us to be more explicit about addressing two early questions:
(a) whether a test should be used for the intended purpose and
(b) whether the qualities of the test as a measure of the intended constructs or domains are sufficient for that purpose

The *Standards* state: "For some test takers, factors related to individual characteristics such as age, race, ethnicity, socioeconomic status, cultural background, disability, and/or English language proficiency may restrict accessibility and thus interfere with the measurement of the construct(s) of interest" (AERA et al., 2014, p. 52). More about the challenges diverse test takers bring than the qualities of the measure itself

The authors of the *Standards* offered general examples of how tests may limit access to the construct for some by including idiomatic phrases and regional vocabulary unrelated to the target construct or stimulus contexts unfamiliar to test takers given their cultural background (pp. 52-53). These test characteristics were not further explained and specific examples were not offered.

A threat to fairness is in test content, vis-à-vis test taker culture and linguistic histories. As an example, the authors of the *Standards* argued that a critical reading test "should not include words and expressions especially associated with particular occupations, disciplines, cultural backgrounds, socioeconomic status, racial/ethnic groups, or geographical locations, so as to maximize the measurement of the construct (the ability to read critically) and to minimize confounding of this measurement with prior knowledge and experience that are likely to advantage, or disadvantage, test takers from particular subgroups" (p. 54).

The authors continued:
"Differential engagement and motivational value may also be factors in exacerbating construct-irrelevant components of content. Material that is likely to be differentially interesting should be balanced to appeal broadly to the full range of the targeted testing population (except where the interest level is part of the construct being measured). In testing, such balance extends to representation of individuals from a variety of subgroups within the test content itself. For

example, applied problems can feature children and families from different racial/ethnic, socioeconomic and language groups" (p. 55).

I contend, as do the scholars and measurement specialists exploring CLR, that such a balance is contradictory with the earlier advice of what to avoid and destroys the value of measures of critical reading, when the readings themselves are based on narrowly defined content domains void of the cultural and linguistic realities of students. How is it possible to "feature children and families from different racial/ethnic, socioeconomic and language groups" (p. 55) while avoiding "words and expressions especially associated with particular occupations, disciplines, cultural backgrounds, socioeconomic status, racial/ethnic groups," (p. 54) that make it possible to represent the world?

Finally, the authors of the *Standards* presented opportunity to learn as a relevant context factor, which is "the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test" (p. 56). Disparities in school resources and availability of diverse teaching force affect the quality and content of teaching and learning. The concepts of CLR are reiterated through UDL, test content can be made more accessible "by avoiding item contexts that would likely be unfamiliar to individuals because of their cultural background" (p. 58). Opportunity to learn is a restricted concept, as it presumes the target content domain is appropriate and [exhaustive] comprehensively taught and covered on the test.

Test and item developers must contend with the somewhat conflicting guidance in the *Standards*, by revising long-standing test development procedures to not just focus on what to avoid in item and test content and contexts, but what to include. Perhaps we can attend to construct underrepresentation as much as construct-irrelevant variance.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. AERA. https://www.testingstandards.net/open-access-files.html

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

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). National Center on Educational Outcomes, University of Minnesota. http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html