The 2014 Standards for Educational and Psychological Testing: What Teachers Initially Need to Know

Robin Thomas Pitts

University of North Carolina at Greensboro

Oksana Naumenko

University of North Carolina at Greensboro

The Standards for Educational and Psychological Testing is a joint publication of the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) that outlines professional expectations for the design, implementation, scoring, and reporting of educational and psychological assessments (2014; hereafter *Standards*). As did previous versions, the present iteration of the *Standards* argues that "those who are participants in the testing process" should have adequate knowledge of tests and assessments (APA, AERA, NCME, 2014, p. 3), suggesting a wide readership. Although the primary audience for the new *Standards* appears to be the test development community, a secondary audience – teachers – would benefit from familiarity with the current guidelines. Teachers comprise an essential group of stakeholders in educational assessment. Prior to testing, teachers develop curriculum, design instruction, and facilitate learning in alignment with state and national standards and may be required to do so for multiple courses, content areas, and/or grade levels during a given academic year or across multiple years. Teachers later administer tests, interpret outcomes in context, and operationalize findings by making decisions and taking action in response to test scores. In addition to using test scores, teachers are also affected by test score-based decisions made by other educational professionals at the school, district, state, and national levels. By way of example, end of grade (EOG) test scores are sometimes used to make personnel decisions in regard to teacher promotion, dismissal, and salary changes (Anderson, 2005; Briggs, 2011).

Despite the significant impact test scores have on teachers, the *Standards* do not include teachers as primary intended audience and do not facilitate teacher engagement with those guidelines within the *Standards* that are most directly applicable to teaching professionals. The intent of the *Standards* is to guide the practices of those professionals who "specify, develop, or select tests, and for those who interpret, or evaluate the technical quality of, test results" (APA, AERA, NCME; 2014, p. 1). Part of the rationale for the exclusion of teachers from the primary audience is the understanding that teachers should not be expected to follow the *Standards* when they serve as test developers for classroom assessments (Plake & Wise, 2014). Indeed, the developers of the *Standards* "felt that classroom teachers would benefit from reading the *Standards* and that promoting assessment literacy for teachers was another important goal" (Plake & Wise, 2014, p. 5). Despite these intentions, because teachers are not included in the primary audience, they may not be able to engage fully with the content of the *Standards*: the technicality of terminology, focus on operational practices, and lack of guidance on how to redress the unintended consequences of test score use represent a significant omission of teachers' issues as related to educational testing.

A case can be made that teachers who can interpret, articulate, and cite the *Standards* are equipped to ensure that test scores are being used accurately and appropriately. Familiarity with "the intended use of the assessment, evidence supporting the validity of inferences concerning the use of assessment results, and test content and test characteristics of the test taking population" (Camara, 2010, p. 4) can be expected to reduce the likelihood of test score misuse. Fluency in the expectations for test score use might also permit teachers to identify and respond to score misuse, helping to deter possible negative impacts on both students and teachers and possibly helping to inform test developers of areas for improvement in test design, implementation, and reporting. As such, we argue that familiarity with the *Standards* can empower teachers to advocate for the appropriate use of assessments within their schools and school systems. To facilitate this vision, and to address teachers' issues as related to the *Standards,* we present an elucidation of teacher-relevant guidelines.

The goal of this paper is to provide teachers with a brief overview of three key issues in current educational practices that relate to the use of test scores in high-stakes decision-making and to contextualize each issue using the *Standards*. Given that the target audience for the *Standards* excludes classroom teachers, this paper focuses on presenting sections of the *Standards* that have implications for teacher practice. (Note that issues related to classroom teaching practices will be mentioned briefly but will not comprise a substantive focus in this paper). The three key issues are (a) avoiding the use of a single test score for high-stakes decision-making; (b) ensuring the opportunity to learn prior to testing for high-stakes decision-making; and (c) considering validity evidence for high-stakes test score use for teacher personnel decision-making. Overall, we suggest that teachers would benefit from understanding more about these issues so as to better advocate for best practices regarding assessment score use, especially given the limited nature of the inferences that may be made using achievement test scores.

Following an initial review of the *Standards*, we identified three interrelated chapter sections as particularly relevant to teacher experiences with achievement tests: (a) Chapter 3, Cluster 4 – Safeguards Against Inappropriate Score Interpretations for Intended Users; (b) Chapter 12, Cluster 2 – Use and Interpretation of Educational Assessments; and (c) Chapter 13, Cluster 2 – Interpretation and Uses of Information from Tests Used in . . . and Accountability Systems. Themes inherent in the chosen sections that may resonate with teachers include (a) interpretations of student performance, (b) design of classroom assessments, and (c) responses to the use of test scores as it impacts teachers. For each theme, we identify the context, cite relevant sections of the *Standards*, and articulate a take-away point with teachers as the audience in mind.

Single Test Score Use

Context. Many current testing systems in the United States focus on a single measure of student performance to make high-stakes decisions. This is evident in primary and secondary educational systems that use end-of-course (EOC) and end-of-grade (EOG) assessments to make grade-level promotion decisions. Decisions made using scores from high-stakes testing impact users of those scores on several levels: school funding and accreditation status can be determined (in part or in whole) by test scores; teacher employment status and merit pay decisions can be rooted in student performance on assessments; and individual students' progress through their studies from one year to the next and are accepted into differentiated programs based on test scores (Au, 2013). While the *Standards* identified the use of a single test score as a detrimental practice for high-stakes decisions, the reality of accountability testing mandates does not always reflect this stance (Thomas, 2005).

As an example, according to the "Read to Achieve" law, all 3^{rd} grade students in the state of North Carolina must show proficiency in reading in order to continue onto fourth grade (Superintendents,

N.C. State Board of Education, Department of Public Instruction, 2014a). If promotion/retention decisions were made solely using EOG scores, such a system would violate the best practices discussed in the *Standards*. For this particular testing system, alternative indicators for proficiency have been permitted (such as proficiency on alternative assessments, through portfolios, etc.) and "good cause exemptions" have been generated that allow for exceptions to the law to be made on the basis of status indicators (e.g., previous retention, SPED/504 status, ELL status) (Superintendents NCSBE & DPI, 2014b). Such a system appropriately avoids the use of a single test score to make high-stakes decisions around promotion/retention.

Standards. Single test score use is addressed in several sections of the *Standards*. Standard 12.10 states that decisions that will significantly impact students "should take into consideration not just scores from a single test but other relevant information" (p. 198). In diagnostic decisions and for special program placement, Standard 3.18 asserts that "multiple sources of information should be used, alternative explanations for test performance should be considered, and the professional judgment of someone familiar with the student should be brought to bear on the decision" (p. 71). In accordance with this notion, Standard 12.13 states that "empirical evidence documenting the relationship among particular test scores, the instructional programs, and desired student outcomes should be provided" (p. 199) and that, when unavailable, other relevant information about students should be considered in tandem with test results. Standard 13.9 furthers this assertion, stating "test results should be used in conjunction with information from other sources when the use of additional information contributes to the validity of the overall interpretation" (p. 213).

Takeaway. Single test score-based decisions are inherently inappropriate as they are based on an insufficient summary of test taker achievement. When such decisions have the potential for significant impact on the test-taker, additional sources of information should be considered. This is especially true if scores are used to make decisions regarding diagnoses or placement. If teachers and students are required to participate in testing systems that use a single test score (such as the EOG score) to make high-stakes decisions, teachers should feel empowered to raise concerns about the use of a single test score and to cite the *Standards* as evidence for the need to seek additional information in making high-impact decisions about examinees. It might be reasonable to suggest additional measures that could serve to supplement the test score by providing additional information about student performance on the same academic standards, across a wider diversity of standards, or through different modes of assessment (performance assessment, portfolios, etc).

Opportunity to Learn

Context. In an educational setting that values the use of testing scores to make high-stakes decisions, inherent tensions exist within instructional practices that prepare students for those assessments: to what extent should course time be spent preparing students for achievement assessments? What is the opportunity cost of allocating time to test preparation that would otherwise be spent on other, perhaps higher-yielding, learning experiences? At what point do teaching practices venture into "teaching to the test" in such a way that student learning is negatively impacted? While students should receive the opportunity to learn prior to being assessed, it is necessary to distinguish such test preparation practices from practices that focus on teaching methods for "gaming" assessments, as such practices reduce the overall validity of test scores and their interpretations. Further, students should be given meaningful opportunities to learn between assessments and multiple opportunities to show proficiency prior to the making of high-stakes decisions. Such terms as "opportunity to learn" can be vague, allowing for a variety of interpretations and thus a variety of implementation practices.

Standards. Standards 12.7 and 12.8 describe the tension between "teaching to the test" and "opportunity to learn." Specifically, Standard 12.7 states that excessive teaching of practice items equivalent to those used on tests "may adversely affect the validity of test score inferences" (p. 197). Such activities negatively impact score interpretation. Further, Standard 12.8 warns that in contexts where high-stakes decisions employ the use of scores, "evidence should be provided that students have had an opportunity to learn the content and skills measured by the test" (p. 197). Finally, in Standard 3.19, we see that "examinees should not suffer permanent negative consequences if evidence indicates that they have not had the opportunity to learn the test content" (p. 72).

Takeaway. It is thus left to a teacher, as a professional within his/her content area and grade level, to determine what constitutes "opportunity to learn" without artificially inflating student scores through test preparation activities like "drill and kill" item exercises. Specific guidelines are not available to direct decisions about instructional practices so as to ensure that students are aware of the test domain, nature of items, mastery criteria and modes of test administration while avoiding the artificial inflation of test scores through inappropriate test preparation activities. This gap in addressing teacher experiences structuring time for meaningful learning is not insignificant. Chapter 12 of the *Standards* could be substantially enhanced by an explication of practical guidelines for increasing opportunities to learn while guarding against negative washback (i.e., the extent to which a test influences teachers' practice and student learning are negatively influenced by a testing program) of teaching students an inappropriately limited curriculum. Teachers could engage with reflection on opportunity to learn through a number of reflective activities: comparing course syllabi with test specifications available in testing manuals, analyzing lesson plans and daily objectives for areas in which teaching methods and assessment methods could be aligned with test expectations (within reason), and collaborating with other educators on how to ensure students are adequately prepared for assessments without detriment to other learning goals.

Test Purpose(s), Interpretation(s) and Use(s)

Context. With an increasing reliance on testing for high-stakes decision-making, the uses of test scores have become more diverse. Teachers may not be aware of the need for test designers to provide validity evidence for each specific use of test scores. Some teachers are currently being evaluated on students' EOC/EOG test scores, which claim to represent student proficiency on the content and performance standards prescribed for a given course. Using test scores as a measure of teaching quality is problematic for many reasons, but such use of test scores is unacceptable most importantly because the explicit intended uses of student achievement tests are focused on student-oriented decisions. In many schools, students enter courses with proficiency levels that are below standard for previous coursework, creating learning challenges unique to each student (Emery, Kramer, and Tian, 2003). State testing programs attempt to adjust for previous year proficiency and other factors in figuring teachers' impact on yearly test score change (Baker, et. al., 2010); however, this practice has proved to be controversial (Briggs, 2011; Koedel & Betts, 2011; Rothstein, 2009). States continue to resolve to use *value added models* (VAMs) for determining teacher impact, in spite of formal rejection by the educational and statistical academic communities (American Educational Research Association, 2015; American Statistical Association, 2014; National Association of Secondary School Principals, 2015). Facilitating teacher exposure to the *Standards* may increase awareness of the inappropriateness of some score uses by clarifying the purposes, evidences, and uses that are intended. Such an awareness can only improve the rate of accurate and appropriate test score interpretation, resulting in better data-driven decision-making.

Standards. Although a large number of the *Standards* address this issue, a sample of standards

that adequately represent the importance of providing validity evidence for each use of test scores is presented. For our purposes, *validity* is defined here as the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test (APA, AERA, NCME, 2014). Standard 12.2 highlights the responsibility of test designers to provide "evidence of validity, reliability/precision, and fairness" for *each* [emphasis added] intended use (p. 195). Standard 13.4 expands this notion by emphasizing the need for information that will "minimize possible misinterpretations or misuse of data" (p. 210). Standard 12.11 extends this to include the use of test scores to calculate difference or growth scores for individual students and also highlights the need for test designers to share with test users the technical and interpretive information inherent in the statistical models used (p. 198). Standards 12.14 and 12.15 outline the need for personnel within a given school to be able to articulate and train others on "the relationships among the tests used, the purposes served by the tests, and the interpretations of the test scores for the intended uses" (p. 199). Further, school personnel must ensure that professionals using the scores are qualified for such work through professional development and ongoing training.

Takeaway. Test scores are increasingly being used for purposes other than those for which the test was designed. Teachers are not always provided access to the evidence that supports the use of test scores to make specific types of decisions. By becoming more aware of the evidence that supports the use of a specific test, teachers will be more capable of identifying uses of test scores that are and are not supported, improving the quality of score interpretations in a way that could significantly impact individual teachers or students. Teachers should request information about the validity evidence available for a given test by researching the appropriate test purposes and claims. When score use in a specific context seems out of alignment with the intended score uses, such as in the case of VAMs, teachers can bring attention to these discrepancies by facilitating dialogue with other educators, with school and system administrators, and with test publishers (specifically, research and development departments). Open dialogue about score uses should be sought so as to clarify discrepancies and to ensure accurate and appropriate score use.

Conclusions

The *Standards* discussed in this paper speak to the importance of assessment literacy in the teaching profession. We identified three major themes from *the Standards* that we felt were relevant to teachers regarding the use of test scores for high-stakes decisions. We found that the *Standards* support testing systems that focus on making decisions about students (a) using more than a single test score, which should empower teachers to advocate for their students by providing or demanding other relevant information prior to decision-making, (b) when evidence is available to show that students have received adequate opportunity to learn, which should empower teachers to argue for students when such opportunities are lacking, and (c) when test users can evaluate the alignment of score use with the intended use(s) of scores, which can serve to empower teachers to advocate for themselves and for students in situations where discrepancies may exist. It may be useful for measurement professionals to develop materials that directly address and meet the information needs of teachers as professionals who utilize testing information for decision-making. Increasing teachers' awareness of such key guidelines would improve assessment literacy and increase the appropriateness of teacher-based practices involving the use of test scores.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

American Educational Research Association, (2015). AERA statement on use of Value-Added Models (VAM) for the evaluation of educators and educator preparation programs. Washington, DC: AERA.

American Statistical Association, (2014). ASA statement on using Value-Added Models for educational assessment. Alexandria, VA: ASA.

Au, W. (2013). Hiding behind high-stakes testing: Meritocracy, objectivity and inequality in U.S. education. *The International Education Journal: Comparative Perspectives*, 12(2), 7-19.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., ... & Shepard, L. A. (2010). Problems with the Use of Student Test Scores to Evaluate Teachers. EPI Briefing Paper# 278. *Economic Policy Institute*.

Briggs, D. C., & Weeks, J. P. (2011). The persistence of school-level value-added. *Journal of Educational and Behavioral Statistics*, *36*(5), 616-637.

Camara, W.J. (2010). From the president: A primer on standards, guidelines, and principles relevant to educational assessment. *NCME Newsletter*, 18(3), 1-6.

Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education*, *11*(1), 37-46.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education*, *6*(1), 18-42.

Koenig, J.A. (2006). Introduction and overview: Considering compliance, enforcement, and revisions. *Educational Measurement: Issues and Practice,* 25, 18-21.

National Association of Secondary School Principals, (2014). NASSP position statement on Value Added Measures in teacher evaluation. Reston, VA: NASSP.

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.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education*, 4(4), 537-571.

Superintendents, North Carolina State Board of Education, Department of Public Instruction. (2014a). NC GENERAL ASSEMBLY'S READ TO ACHIEVE LEGISLATION. *LEA*.

Superintendents, North Carolina State Board of Education, Department of Public Instruction. (2014b). ALTERNATIVE ASSESSMENT PROPOSALS UNDER THE GENERAL ASSEMBLY'S READ TO ACHIEVE LAW. *LEA*.

Thomas, R.M. (2005). *High-stakes testing: Coping with collateral damage.* Mahwah, N.J.: L. Erlbaum Associates.