Examining Standard Practice: How Published Reports on Test Characteristics Align with Professional Guidelines

Jerusha J. Henderek & Jonathan D. Rubright

National Board of Medical Examiners

Abstract

The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) outlines the expectations for reporting. It is essential to consider whether assessment organizations are meeting the standards of transparency that are necessitated by the professional community. We reviewed websites from seven licensure examinations and found that none reported all information required. When information was reported, it often was dispersed across many websites or documents. Organizations should consider what the most effective way is to communicate the depth and breadth of information demanded, and the *Standards* can help by more clearly defining what information should be available to the variety of stakeholders touched by high-stakes examinations.

Examining Standard Practice: How Published Reports on Test Characteristics Align with Professional Guidelines

The *Standards for Educational and Psychological Testing* (AERA et al., 2014; hereafter referred to as the *Standards*) outlines the professional expectations for developing tests and provides guidance on matters both legal and technical for test developers to follow. Encompassing more than 200 pages, an entire chapter (Chapter 7) of this guide is exclusively devoted to outlining the documentation required for test publishers to generate and provide to test users so that users are equipped to decide whether the assessment meets their needs of rigor. Although thorough documentation should be available regardless of the importance of decisions to be made on the basis of test scores, test developers of higher-stakes exams arguably have an even greater obligation to ensure their documentation is complete to allow test users to "assess the nature and quality of the test" (AERA et al., 2014, p. 123). This documentation may take various forms, such as test manuals or research reports, and may be written to the level of particular audiences, such as other test developers, proctors, or examinees themselves.

Tests with arguably the highest of stakes are those that make pass/fail decisions about examinees which either grant or deny a test-taker the privilege of obtaining a license or credential in their chosen professional field. In these situations, test publishers typically spend much energy to ensure their tests are up to the high task of appropriately classifying examinees in a high-stakes context. And, test takers in these settings are motivated to know that the test is indeed making appropriate decisions about their level of knowledge, skill, and ability in the given domain. In an effort to understand the extent to which test publishers are providing the information required by the *Standards* to those taking their tests, we sought to consider the

alignment between the expectations in the *Standards* and what is actually reported by major testing programs in the licensure and certification testing space.

The *Standards* provides the gold standard for educational and psychological assessments to which all test publishers should strive; it "provide[s] criteria for the development and evaluation of tests and testing practices" (AERA et al., 2014, p. 1). It is common to assess whether a particular program meets guidelines for, say, validity evidence. Even though the *Standards* outlines, in detail, the level of transparency on test characteristics expected from test publishers, the test development process may still be viewed as a black box by test users. It is essential to consider whether assessment organizations in practice are meeting the standards of transparency necessitated by the professional community.

Method

We considered criteria the *Standards* states as necessary for reporting as a framework for comparing across identified exams. We primarily focused on Chapter 7 of the *Standards*: "Supporting Documentation for Tests." Two psychometricians reviewed publicly available materials from seven high-stakes, pass/fail certification/licensure examinations and made note as to whether information was provided for each of the categories identified in the *Standards*. Elements the *Standards* declares should be reported were recorded as either present or not present across the various examinations in question.

Data were collected by first reviewing and listing the information outlined as necessary to report by the *Standards*. Seven certificate/licensure exams were compared to the *Standards* and to each other by reviewing publically available information. Information was taken from various organizations' websites, along with publically available technical reports, final reports, and research bulletins. Results are presented here anonymously.

Results

After exhaustive review of materials publically available and published on each testing organization's own website, we compiled results across the exams to compare what commonly was or was not reported on the various assessments (see Table 1). Based on this review, we found that none of the certification examinations reviewed here provided easily available information on the following standards:

- The process of reviewing items at key validation (Standards 4.10 and 7.5),
- The reliability of the assessment (Standard 7.4),
- Support for the recommended uses of the test (Standard 7.1),
- Evidence for predictions of future behavior made by the test (Standard 7.12), or
- The necessary qualifications to administer and score the test (Standard 7.7).

For certification and licensure exams in the professions, a lack of information on Standard 7.12 is not surprising given that the *Standards* notes that predictions of future behavior are of "limited applicability" (p. 175) in this testing context due to a focus on mastering the appropriate content to be certified in the area (as opposed to being confident one will perform well in the career) and a restricted range of criterion data: information on the performance of those not granted a credential are unavailable.

Concerning Standard 7.7, it may be that necessary qualifications for scoring were not reported because it is considered obvious that a more detailed scoring process takes place than can be computed directly without oversight by a psychometrician. However, this should still be stated somewhere easily accessible to the public. Of the five standards not reported by the test publishers reviewed here, it seems most concerning that information was not readily available on reliability (Standard 7.4) or information on the key validation process (7.5). Two certification

examinations that did not have reliability reported did provide some information on the statistical properties of the scores by reporting the standard error of measurement; this arguably fulfills the need to provide "evidence of the reliability/precision of scores" (p. 126). The remaining four certification examinations reported neither a reliability, the standard error of measurement, nor the standard error of estimate. It is also concerning that we did not find information supporting the recommended uses of the test for any of the examinations in this review (Standard 7.1).

Additionally, programs did not generally provide information regarding the details of the standard setting process. Only one certification examination provided information on the number and expertise of the subject matter experts that participated in their standard setting panel (Standard 7.5) and the process for reviewing items (Standard 7.4) used. Both pieces of information are crucial to ensuring the public has a transparent view on the thoroughness and appropriateness of the process that decides the passing standard for a high-stakes examination.

On the other hand, there were criteria that the majority of programs did include in their publically available information. For instance, all but one certification exam provided information on the test administration details (Standard 7.5). In addition, five out of the seven examinations provided readily accessible information on item development details (Standards 7.4 and 7.5), incident report review process (Standard 7.9), and evidence for validity (Standard 7.4).

Discussion

It is noteworthy that, although the *Standards* is widely known, read, cited, and explicit in its recommendations, there is such large variance in the specific recommendations met by each examination program reviewed here. A glance at Table 1 shows the inconsistency in how each standard is addressed by each testing program. There may be a number of reasons for this. First,

the information may be provided by each test publisher, but in a format or location that is not easily accessible to test users. The information tracked here was identified by two psychometricians working in the certification and licensure field after an exhaustive search of publically available reports. It is possible that information was missed in the review, and it is also possible that additional detail has been reported in a peer-reviewed journal, which is less accessible to the general public, or in reports or webpages less clearly marked or identifiable. Second, the *Standards* itself is very clear on what is to be reported, yet is less clear about *to whom* these data should be reported. The *Standards* outlines that these data should be available to "communicate with test users" (p. 123), and yet "test users" is a very general term. It very well may be that these data are available to internal staff or to Board Members under confidentiality agreements – yet not to the general public.

In order to improve transparency to test users, two general recommendations can be made. First, it would be helpful if the information to be shared with test users was in a single, easily identifiable and reachable location for each examination program. Many testing programs have data distributed across test manuals, practice materials, idiosyncratic search engines of research repositories, and other locations. Providing data in a single location geared to a particular audience would reduce confusion and make clear what is, and what is not, being reported and why. Second, the *Standards* could provide more specific guidance about what data must be available and to whom: what should be shared with administrators, to test takers, to the general public? This level of specificity would clarify whether testing programs are meeting their professional obligations.

It is important that test developers and publishers are transparent about their work, especially given the impact that the products and services they provide have on individual test

takers and on society more generally. The *Standards* has published guidelines on the test characteristics that should be accessible so that external stakeholders can judge the quality of the products and services provided. It is reasonable to hold test publishers to these *Standards*. As the *Standards* points out, "failure to formally document such evidence in advance does not automatically render the corresponding test use or interpretation invalid" (AERA et al., 2014, p. 123); however, all certification examinations considered in this study have been in place for years and have high-stakes consequences attached to test scores. Therefore, it is of upmost importance that the public be provided access to detailed documentation supporting the inferences made on the bases of the scores from these tests. This study helps highlight areas where high-stakes examinations are currently excelling in transparency along with areas where they could improve in providing adequate information regarding their examinations with the intention of drawing awareness to the issue and improving what is reported about examinations in practice.

This review of current practice reveals that even when information is reported for an examination, it often was dispersed across many websites or documents, as opposed to providing a single piece of documentation supporting the inferences made based on the test scores. Organizations that develop assessments should consider what is the most effective way to communicate the depth and breadth of information demanded by the *Standards*; while the *Standards* can help organizations meet these requirements by more clearly defining what information should be available to the variety of stakeholders touched by high-stakes examinations.

Reference

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

Table 1. *Alignment between the* Standards *and Practice.*

	Testing Organizations						
Criteria	A	B	C	D	E	F	G
Test development procedures							
Item development details	X			X	X	X	X
Form assembly & review details	X				X		
Test Administration							
Test administration details		X	X	X	X	X	X
Incident report review process		X	X	X	X		X
Examinee Groups							
Total group description		X	X	X			X
Reference group description	X						
Base reference group description				X			
Key Validation							
Process of reviewing items							
Scoring/Equating							
Details on scoring & scores reported			X	X			X
Description of adjustment made by equating				X		X	
Statistical Properties of Scores							
Reliability							
Overall Standard Error of Measurement (SEM)	X					X	
Standard Setting (SS)							
Type of SS conducted	X	X			X		X
Number of SMEs & expertise	X						
Process for reviewing items	X						
Decision made for cut score	X				X		X
Additional Information Supporting Test Use							
Rationale for test		X	X	X			X
Recommended uses for the test		X	X				
Supports of the recommended uses for the test							
Procedures for gathering norm data			X				
Studies about general and specific uses		X	X				
Evidence for validity	X	X	X	X			X
Evidence for predictions of future behavior made by the test							
Necessary user qualifications to administer & score a test							
Materials to assist takers with interpreting scores			X				X

Note. "X" indicates some information was found for that criteria. Letters A – G represent the testing organizations that we reviewed.