A Systematic Review of Assessment Literacy Measures

Chad M. Gotch and Brian F. French, Washington State University

This work systematically reviews teacher assessment literacy measures within the context of contemporary teacher evaluation policy. In this study, the researchers collected objective tests of assessment knowledge, teacher self-reports, and rubrics to evaluate teachers’ work in assessment literacy studies from 1991 to 2012. They then evaluated the psychometric work from these measures against a set of claims related to score interpretation and use. Across the 36 measures reviewed, they found support for these claims was weak. This outcome highlights the need for increased work on assessment literacy measures in the educational measurement field. The authors conclude with recommendations and a resource to inform a research agenda focused on assessment literacy measurement to inform policy and practice.

Keywords: assessment literacy, educational assessment, student assessment, teacher competence, teacher preparation

Assessment literacy (Stiggins, 1991) has long been viewed as an important characteristic of effective teachers. Recent reform efforts centered on teacher evaluation reflect a new context in which we can examine the assessment competencies of today’s teachers. In this article, we review available assessment literacy measures, and evaluate the psychometric evidence to support their adequacy for use in teacher evaluation. We conclude the psychometric evidence in support of assessment literacy measures is weak, highlighting the need for a research agenda focused on this topic.

Assessment Literacy as a Teacher Characteristic

Interest in teacher assessment competencies can be traced back several decades to investigations of educational measurement coursework required for state certification in teaching, uses and opinions of tests, and student evaluation techniques. Schafer and Lissitz (1987) provided a thorough review of this work. By the early 1990s, interest in the topic brought about the Standards for Teacher Competence in Educational Assessment of Students (American Federation of Teachers, National Council on Measurement in Education, & National Education Association, 1990) and a formal term, assessment literacy (Stiggins, 1991). Broadly, assessment literacy can be defined as “an individual’s understandings of the fundamental assessment concepts and procedures deemed likely to influence educational decisions” (Popham, 2011, p. 267). Specific behaviors reflecting assessment literacy include using multiple high-quality assessments aligned with precisely defined achievement targets, interpreting student performance in light of the particular form of assessment and potential impact of extraneous factors, administering and scoring assessments appropriately, accurately communicating results to interested parties, and carrying out all assessment responsibilities legally and ethically (e.g., American Federation of Teachers, National Council on Measurement in Education, & National Education Association, 1990; Brookhart, 2011; Popham, 2009; Stiggins, 1991). In empirical study, assessment literacy acts as a stable but malleable characteristic of teachers that is reflected in measures targeted toward the attribute.

Assessment Literacy as a Policy Consideration

The federal Race to the Top Fund placed clear emphasis on teacher effectiveness and evaluation systems (U.S. Department of Education, 2009). Common teacher evaluation frameworks feature elements related to effective and appropriate use of student assessment. Danielson’s framework for teaching (2013), for example, expects proficient teachers to demonstrate skill in designing student assessment, using assessment in instruction, and identifying high-quality sources of data for monitoring student growth. Marzano’s teacher evaluation model (2013) examines how teachers use assessment to track student progress and document the effectiveness of a particular lesson. These frameworks are aligned with the general definition of assessment literacy, and signal a need for quality assessment literacy measures to inform policy and practice.

A goal of the National Council on Measurement in Education (NCME) is to “influence public policy and practice concerning educational measurement.” Furthermore, educational measurement professionals are bound by responsibilities to document the technical adequacy of assessments to support appropriate interpretations and uses of scores (NCME, 1995). Moreover, government agencies and nonprofit research organizations look to NCME to act as an unbiased resource on assessment and accountability issues (Ackerman, 2009). Therefore, the measurement of teacher assessment literacy in the service of large-scale educational reform efforts is a timely and relevant issue for NCME members to address.
Evaluation of Assessment Literacy Measures

In this article, we advance a set of claims that lay a foundation for measuring teachers’ assessment competencies, especially in the context of teacher evaluation. These claims address five fundamental aspects of measurement—test content, internal consistency reliability, score stability, internal structure, and association with other variables—and build an interpretation/use argument (Kane, 2013) for assessment literacy measures. The claims can be addressed through basic psychometric analyses, such as those typically found in a test manual. We leave aside claims that may be considered difficult to frame in a validity study or for which there is active debate regarding their role in the validation process, such as those on the consequences of test use. For each claim, we evaluate the psychometric evidence provided for each assessment literacy measure against stated criteria. We conclude published support for the advanced set of claims is weak.

Method

Literature Search

We conducted a systematic search of research published from 1991 to 2012 by exploring the Educational Resources Information Center, PsychInfo, JSTOR, Education Full Text, and ProQuest Dissertations and Theses databases using the terms assessment literacy, measurement literacy (Barry, 2009; Lambert, 1991), and the combination of educational assessment or student assessment with teacher competence, teacher understanding, teacher training, or teacher preparation. We referenced Google Scholar for additional published work citing Stiggins’s (1991) article on assessment literacy. This process returned 1,205 articles, books, book chapters, reports, conference papers, dissertations, theses, reviews, unpublished manuscripts, and errati. Any work employing an assessment literacy instrument in the P-12 United States educational setting or an international equivalent was retained for review. We also examined reference sections for additional published material suitable for inclusion.

Through this process, we identified a final set of 36 instruments appearing in 50 studies. These instruments took three forms—objective tests of assessment knowledge (n = 15), teacher self-reports of assessment competence (n = 14), and rubrics evaluating teachers’ work (n = 7). The objective tests primarily included multiple-choice items. Teacher self-reports resembled tests of assessment knowledge in terms of content, yet asked teachers to judge their skills using rating scales ranging from 3 to 10 points. Rubrics provided a structure for assigning scores to teacher work such as assessment plans (Campbell & Evans, 2000), use of formative assessment in the classroom (Sato, Chung, & Darling-Hammond, 2008), and teachers’ own grading rubrics (Schmitt, 2007). Eight measures were used in works published between 1991 and 1995, 14 first appeared between 1996 and 2007, and 14 first appeared between 2008 and 2012, the last five years of the review time period. The most widely used measure was the Teacher Assessment Literacy Questionnaire (Plake, Impara, & Fager, 1993)—later revised as the Classroom Assessment Literacy Inventory (Mertler, 2003)—which appeared in 11 studies and spawned an Arabic translation (Alkharusi, 2011). This instrument and others that could be obtained by the authors can be viewed in Appendix A (available as a separate link in the electronic version of this article or by request from the first author).

Evaluation of Psychometric Evidence

Test content. To examine support for test content, we evaluated the claim that items on the instrument are relevant to and representative of the constitutive definition of assessment literacy. We required evaluation of the content of assessment literacy measures beyond literature review and solicitation of feedback to support this claim. We considered both item-level (content validity index [CVI] values of 1.00 with three to five expert raters or > .78 with six or more raters; Polit & Beck, 2006) and scale-level (CVI, proportion of agreement > .90; Polit & Beck, 2006) forms of support.

Internal consistency reliability. We evaluated the extent to which evidence supported the claim that scores demonstrate enough consistency to support decisions about individual teachers. Individual-level decisions may not have been explicit in the development of any one measure; however, given current political and practical interests in teacher evaluation, we deemed it necessary to target the claim at the individual level. We judged internal consistency estimates > .95 as acceptable for decisions about individuals (Nunnally & Bernstein, 1994).

Score stability. We examined evidence to evaluate the claim that scores were stable over time or stable across observers. We considered estimates of stability across time (e.g., test–retest correlations) or across observers (e.g., intraclass correlation coefficients) that exceeded .90 as support for this claim (Cicchetti, 1994; Floyd & Bose, 2003).

Internal structure. Several measures suggested a composition of related yet distinct assessment skills, either in the scoring routines or in the development blueprints. Thus, we examined evidence that items and scores from subscales in the assessment literacy instrument can be used to characterize teacher’s relative strengths and weaknesses across distinct yet related skillsets. We noted use of exploratory factor analysis procedures, but required well-fitting confirmatory factor analysis models to support this claim, given the theory-driven, versus data-driven, nature of the analysis. We evaluated model fit indices based on standard criteria (Hu & Bentler, 1999; e.g., CFI > .95; SRMR < .08; RMSEA < .05; WRMR = 1.0; significant parameter estimates) and interpretability of factors in relation to theory.

Association with student outcomes. Inherent in the importance placed on teacher assessment literacy and other attributes in teacher evaluation frameworks is an assumption that such skills should be associated with positive student learning outcomes. However, no clear guidance on the magnitude of that relationship exists. Thus, we relied on general criteria in the job performance literature (correlations > .40) to evaluate the claim that teacher scores on assessment literacy measures align with the academic performance of their students (Ones, Viswesvaran, & Schmidt, 1993). We also counted instances where there was a positive effect on students who were exposed to teachers undergoing professional development in assessment.
Table 1. The Number of Assessment Literacy Measures With Psychometric Evidence to Support Score Interpretation/Use Claims

<table>
<thead>
<tr>
<th>Test Content</th>
<th>Internal Consistency</th>
<th>Score Stability</th>
<th>Internal Structure</th>
<th>Association with Student Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tests of assessment knowledge (N = 15)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Self-reports (N = 14)</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rubrics (N = 7)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Results

Our review of the published psychometric evidence found little support for basic claims about score use and interpretation from assessment literacy measures. In Table 1, there are few instances of support across all forms of measurement and all types of evidence. In this section, we review the psychometric work addressing each of the claims.

Test Content

Every assessment literacy measure reviewed was developed with at least some grounding in literature on fundamental aspects of assessment or codified professional expectations (e.g., Standards for Teacher Competence in Educational Assessment of Students). We found nine instruments were reviewed by experts with qualifications in classroom assessment or measurement who were external to the authorship of the measures. We found no instance, however, of a content study producing quantifiable outcomes to evaluate against our criteria.

Internal Consistency Reliability

Most of the psychometric work associated with the measures was focused on estimating internal consistency of responses. Estimates were provided for 12 tests of assessment knowledge, but only the assessment knowledge test (Wang, Wang, & Huang, 2008) met the criterion to support decisions about an individual teacher. Four self-report measures met the criterion to support individual-level decisions (Alkharusi, Kazem, & Al-Musawai, 2010; Koloi-Keaikitse, 2012; O'Sullivan & Johnson, 1993; Zhang & Burry-Stock, 2003), as did one rubric (McMorris & Boothroyd, 1993).

Score Stability

Schmitt's (2007) meta-rubric demonstrated adequate inter-rater and intra-rater percent agreements to support the claim of score stability for overall scores. Koh (2011) reported inter-rater agreement estimates of 65%–99% across individual rubric dimensions, indicating adequate score stability for a portion of the instrument. Thus, only two of seven rubrics—no tests or self-reports—met the stability criterion.

Internal Structure

Authors of the assessment literacy measures advanced a multifactor structure for 8 of 15 tests of assessment knowledge, 7 of 14 self-reports, and 5 of 7 rubrics. Only two confirmatory factor analyses were reported (Alkharusi, 2011; Gotch & French, 2011), and neither resulted in model-data fit estimates that met criteria to support evaluation of teachers' relative strengths and weaknesses in distinct yet related assessment skill sets. Exploratory factor analyses were conducted with data from five other instruments (one test, four self-reports). Across analyses, we found no evidence for the identified factor structure that matched structure implied by test development.

Association with Student Outcomes

We found two instances where teacher scores on assessment literacy measures were investigated in conjunction with student learning outcomes. Koh (2011) reported positive effects for students whose teachers also showed improvement in assessment literacy scores after a professional development intervention targeting assessment skills. This was the sole piece of evidence for this claim across 36 measures that could be located and evaluated against the criteria. Lukin, Bandalos, Eckhout, and Mickelson (2004) observed gains in student achievement in classrooms where teachers received professional development and demonstrated improvements on the classroom assessment questionnaire (Arter & Busick, 2001). However, technical details were insufficient to evaluate support for the claim.

Discussion

In this study, we reviewed available measures of assessment literacy against claims and criteria to judge their adequacy for contributing within the teacher evaluation context. We conclude that the psychometric evidence available to support assessment literacy measures is weak. Although the evaluation of teaching effectiveness, including effective assessment practice, is at the top of the national agenda, our findings call into question how ready measures are to be used in responding to such interest.

We recognize that some psychometric studies on this topic have not yet been published or cited. For example, a “content validity study” was mentioned in the development of the teacher assessment literacy questionnaire (Plake et al., 1993), but details did not accompany the major conclusion from the results. Similarly, Lingard, Mills, and Hayes (2006) noted a confirmatory factor analysis to support the dimensions of their rubric but without clear documentation of the work, so we could not evaluate model fit in this review. We also acknowledge the context dependency of support for claims (e.g., sample dependency of internal consistency estimates). Nevertheless, the overall body of evidence suggests our conclusion from this review is robust and not dependent upon fine shifts in accounting. Thus, there is opportunity to build strength in the psychometric body of work supporting assessment literacy measures.

Concerted efforts are required to document quality of teacher assessment literacy measures, especially as new teacher evaluation systems undergo large-scale
implementation. To begin the process, we have compiled available assessment literacy measures reviewed in this study, as shown in Appendix A (available as a separate link in the electronic version of this article or by request from the first author). We suggest initial efforts consider the representativeness and relevance of content in light of transformations in the assessment landscape (e.g., accountability systems, conceptions of formative assessment). Such transformations are not reflected in the Standards for Teacher Competence in Educational Assessment of Students (Brookhart, 2011) and therefore not emphasized in measures constructed from tables of specifications that relied on those standards.

Progressing from these initial efforts, we suggest new lines of research to investigate a range of critical questions on the validity and reliability of these measures. As the present review demonstrates, psychometric investigation is needed to support even basic claims. We highlight the need for studies to connect teacher assessment literacy to student outcomes. Not only is this work needed to build support for assessment literacy measures, but also to build support for assessment literacy as a necessary attribute of effective teachers. We also highlight the potential for additional investigations of internal structure to identify subcomponents of assessment literacy. Such identification would improve understanding of the distinct ways in which a teacher may be literate in assessment, thus allowing for more nuanced and better targeted evaluation and professional development. The educational measurement field excels at the investigations we suggest. The acceptance of the challenge to improve the understanding of assessment literacy and associated measures will fulfill professional duties in educational measurement policy and practice (Ackerman, 2009; NCME, 1995).

Acknowledgments

We thank Derek Briggs for his input in shaping the manuscript and the authors who contributed their measures to the online appendix.

References

Ackerman, T. A. (2009). From the President: Update on NCME activities. NCME Newsletter, 17(4), 1–3.


