PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2015, 52(2), 34-54

Pontificia Universidad Católica de Chile
Facultad de Educación

Valid Writing Assessment from the Perspectives of the Writing and Measurement Communities

Evaluación válida de la escritura desde el punto de vista de las comunidades de investigación en escritura y medición

[1]Nadia Behizadeh and [2]George Engelhard, Jr.

[1] College of Education, Georgia State University, USA
[2] Department of Educational Psychology, The University of Georgia, USA

Abstract

This study examines the concept of validity in two distinct communities of practice: the writing research and educational measurement communities. Conceptualizations of validity have evolved differentially within each community. Three questions guide our research: (a) What is a valid writing assessment from the perspective of the writing community? (b) What is a valid writing assessment from the perspective of the measurement community? (c) What are some points of consensus and disagreement over the concept of validity in the two communities? This study aims to foster communication between these two different scholarly communities regarding validity issues in writing assessment. We also highlight the contributions that Rasch measurement theory (Rasch, 1960/1980) brings to understanding and evaluating validity. Our goals are to enhance the conceptualization of validity in writing assessment and to identify areas of consensus and disagreement regarding definitions of validity. These analyses extend earlier work by Engelhard and Behizadeh (2012) that explored consensus definitions of validity. This research has implications for improving research, theory, and practice in writing assessment.

Keywords: validity, consequential validity, writing assessment, communities of practice, Rasch measurement theory

Post to:
Nadia Behizadeh
Department of Middle and Secondary Education, College of Education
Georgia State University
P.O. Box 3978, Atlanta, GA 30302, USA.
Email: nbehizadeh@gsu.com
An earlier version of this manuscript was presented at the International Objective Measurement Workshop in Philadelphia, PA, April 2014.

ISSN: 0719-0409 DDI: 203.262, Santiago, Chile
doi: 10.7764/PEL.52.2.2015.3

Resumen

Este estudio examina el concepto de validez en dos comunidades de práctica distintas: la de investigación en escritura y la de medición educacional. Las conceptualizaciones de validez han evolucionado diferencialmente dentro de cada una de ellas. Tres preguntas guían nuestro estudio: (a) ¿En qué consiste una evaluación válida de escritura según la comunidad de investigación en escritura? (b) ¿En qué consiste una evaluación válida de escritura según la comunidad de investigación en medición? (c) ¿Cuáles son los puntos de consenso y desacuerdo sobre el concepto de validez en ambas comunidades? El presente estudio busca fomentar la comunicación entre estas dos comunidades académicas con respecto a los problemas asociados con la validez en la evaluación de la escritura. También destacamos las contribuciones de la teoría de medición Rasch (Rasch, 1960/1980) a la comprensión y evaluación de la validez. Nuestras metas son fortalecer la conceptualización de la validez en la evaluación de la escritura e identificar áreas de consenso y disenso en las definiciones de validez existentes. Estos análisis expanden el trabajo previo de Engelhard y Behizadeh (2012), el cual exploró definiciones consensuadas de validez. El presente estudio tiene implicaciones para mejorar la investigación, la teoría y la práctica en la evaluación de la escritura.

Palabras clave: validez, validez consecuencial, evaluación de la escritura, comunidades de práctica, teoría de medición Rasch

> Validation was once a priestly mystery, a ritual behind the scenes, with the professional elite as witness and judge. Today it is a public spectacle combining the attractions of chess and mud wrestling (Cronbach, 1988, p. 3).

Validity was a contested idea throughout the 20th century, and this debate has continued into the 21st century. From Thorndike's (1919) position that a valid scale is one «in respect to whose meaning all competent thinkers agree» (p. 11) to Messick's (1995) unitary conception of validity that encompasses the intended use and unintended consequences of the use of test scores, the measurement community has sought consensus on the meaning of validity (Engelhard & Behizadeh, 2012; Newton, 2012). Complicating matters further, various content area communities (e.g., writing researchers) often have different explicit and implicit conceptions of validity. As validation becomes less of a «priestly mystery» (Cronbach, 1988, p. 3) and more of a public debate, other voices are entering the arena to contribute to the next iteration of a definition of validity in assessment. In this article, we focus on large-scale writing assessment for primary and secondary schools in order to contextualize our perspective on validity.

Underlying our study is the idea of *communities of practice* (Wenger, 1998, 2010, 2015). Wenger (2015) offered three key dimensions of a community of practice: domain of interest, joint activities, and shared practices. Communities of practice share a *domain of interest* with a commitment and shared competence in the domain. Community members engage in *joint activities* and discussions that build relationships that enable them to learn from each other. Finally, members of a community of practice have *shared practices* that include ways of viewing and addressing problems. The writing and measurement communities are two communities of practice that share a domain of interest in the validity of writing assessment processes, yet hold distinct conceptualizations of validity in writing assessment. Historically, these communities have not engaged in shared activities and practices related to validation (Behizadeh & Engelhard, 2011). It is important to note that writing research, also referred to as composition research, is a broad, interdisciplinary field, including scholars in multiple disciplines. In this study, the writing community refers primarily to researchers in curriculum and instruction who are interested in literacy practices, particularly instruction in and assessment of writing.

The purpose of this study is to explore definitions of the concept of validity within two scholarly communities of practice: writing and educational measurement. We discuss how validity has been presented within and across the writing and measurement communities. Three key questions guide our study:

1. What is a valid writing assessment from the perspective of the writing community?
2. What is a valid writing assessment from the perspective of the measurement community?
3. What are some points of consensus and disagreement over the concept of validity in the two communities?

Our goals are to foster communication and clarification between the two communities and to promote advances in research, theory, and practice related to the assessment of writing. Rasch measurement theory is briefly described as a way to explore validity within the context of writing assessment.

Methodology

Our methodology is guided by the concept of epistemic iterations, which uses historical and philosophical analyses to trace how concepts, such as validity, are conceptualized and evolve over time (Engelhard & Behizadeh, 2012). Our review of previous research is selective rather than exhaustive with the goal of presenting what we consider highly influential and promising research about the concept of validity.

First, we focus on characterizing the writing and measurement communities' conceptions of validity in general. Our first sources in both the writing and measurement communities are the professional organizations that provide standards for assessment practices. In the writing community, we selected the National Council of Teachers of English (NCTE), and in the measurement community, we selected the National Council on Measurement in Education (NCME). These two organizations were chosen because of our intended focus on U.S. primary and secondary large-scale writing assessment. NCTE is a leading professional organization focused on pre-collegiate English language arts instruction while NCME is a leading U.S. organization focused on measurement. We examined the NCTE standards for assessment (International Reading Association [IRA]/National Council of Teachers of English [NCTE], 2010), as well as the recently revised *Test Standards* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). These standards reflect consensus definitions for each community. After exploring these standards, we chose one top tier journal published by NCTE and one by NCME to explore at a more granular level the definition and use of the term *validity* connected to writing assessment, focusing on the last 15 years.

In the writing community, the specific term *validity* is not used to the same degree as within the measurement community. In exploring the use of the concept of validity in writing research, we chose *Research in the Teaching of English* (*RTE*), the NCTE journal with the highest impact factor, to represent the writing community. First, we searched for the terms «validity» or «valid*» in titles of articles from 1999 to 2014, which returned zero results. In the same time period, we searched for the term «assess*» in titles and the search returned seven results, of which three were empirical studies relevant to writing assessment: Broad (2000), Bauer and Garcia (2002), Ketter and Pool (2001). Another search for «valid*» anywhere in the text of articles during this time period returned five results, yielding Ketter and Pool (2001) again and an additional four articles, of which two were relevant to writing assessment: Elliot, Deess, Rudniy, and Joshi (2012) and Murphy (2007). We gave extra attention to a special issue of *RTE* on writing assessment that was published in 2014. In particular, we selected articles by Poe (2014) and Slomp, Corrigan, and Sugimoto (2014) for careful review.

In the measurement community, the term *validity* is ubiquitous, and so our methodology was slightly different. We conducted a search of key terms, and due to the plethora of articles, we chose measurement researchers who represent key emergent positions in the ongoing debate on validity. As was the case in the writing community, there was a recent special issue of a major journal (*Journal of Educational Measurement*) on validity that provided guidance about current perspectives. In particular, we examined the views of Kane (2013), Borsboom and his colleagues (Borsboom, 2005; Borsboom, Mellenbergh, & Van Heerden, 2004; Borsboom & Markus, 2013), and Sireci (2013).

In sum, we identified influential work for each community of practice related to the discussion of validity. Next, we engaged in close readings of these selected source materials to provide the basis for

our detailed discussion of the concept of validity in writing assessment. Our findings highlight areas of agreement and disagreement and have important implications for research, theory, and practice.

What is a valid writing assessment from the perspective of the writing community?

In previous research, we have detailed how the construct of writing within the U.S. writing community evolved from a mechanical conception in the early 1900s, to a content-oriented definition by mid-century, and most recently to sociocultural definitions (Behizadeh & Engelhard, 2011) that position context as an important and necessary factor to consider in literacy teaching and learning. In the following analysis of the writing community's consensus definition of validity as represented by *Standards for the Assessment of Reading and Writing* (IRA/NCTE, 2010), one can see the dominance of sociocultural theory.

Writing community: What is validity?

The International Reading Association and National Council of Teachers of English's (IRA/NCTE, 2010) *Standards for the Assessment of Reading and Writing* state,

> Historically, a common definition of a valid measure is that it measures the construct it purports to measure. This is called *construct validity*. For example, if we claim that an assessment measures reading fluency, but it only measures speed and accuracy and does not include aspects such as intonation, the test would have poor construct validity (pp. 52-53).

This conception of construct validity recalls earlier measurement conceptions of construct validity that employed a simple definition (Kelley, 1927). However, these standards continue by noting:

> More recent conceptions of validity include an examination of the consequences of assessment practices — *consequential validity*. For instance, a test might have excellent construct validity as a measure of decoding ability. However, if it were used as the basis for adjusting teachers' salaries, resulting in an overemphasis on decoding in the curriculum, it would not be a valid assessment process. In other words, one cannot have a valid assessment procedure that has negative or misguided consequences for children. Consequently, a productive definition of a valid assessment practice would be one that reflects and supports the valued curriculum (IRA/NCTE, 2010, p. 53).

According to these definitions, construct validity and consequential validity are blended. There appears to be great emphasis placed on evaluating appropriate use of the test and the actual consequences of that use. Importantly for the writing community of practice, a shared purpose for assessment is improved instruction for English language arts students, and this purpose shapes their conception of validity.

In addition to definitions, specific elements of the IRA/NCTE (2010) assessment standards are important to consider. In particular, Standard 7 states, «The consequences of an assessment procedure are the first and most important consideration in establishing the validity of the assessment» (p. 22). The authors continue, noting, «This standard rejects the unfortunately common argument that a given test is valid in spite of the fact that its use has problematic consequences (e.g., placing a student in a program that does not serve her well)» (p. 23). In addition to very clearly elevating consequential validity, Standard 7 is referenced multiple times throughout the rest of the standards, including the explanation following Standard 1. Standard 1 states, «The interests of the student are paramount in assessment» and the line following adds, «Assessment experiences at all levels, whether formative or summative, have consequences for students (see Standard 7)» (p. 11).

Figure 1 represents the four major themes from the IRA/NCTE assessment standards for reading and writing. Standards 3 and 6 focus on the purpose of assessment, which is to inform and improve instruction for linguistically and culturally diverse students. Another theme is the construct of literacy, represented by Standard 5, which articulates a sociocultural understanding of literacy, and Standard 4, which emphasizes the complexity of both reading and writing. Process is a third theme; Standards 2, 8,

9, 10, and 11 all call for multiple perspectives to be included in the writing assessment process from development to reporting, with teachers, students, families, administrators, policymakers, and the public actively participating. The final theme is consequences, represented by Standards 1 and 7, which center on the social implications of assessment for students.

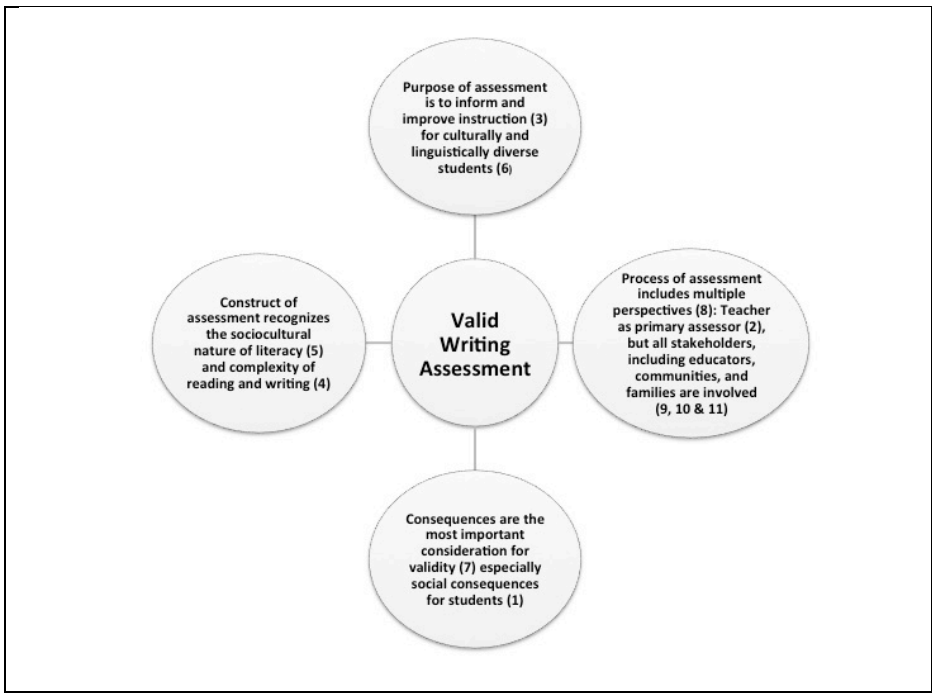

Figure 1. Writing community: Four thematic clusters (IRA/NCTE, 2010) for a valid writing assessment (Numbers in parenthesis refer to the standard).

A question arising from these assessment standards is: Why the focus on instruction for diverse students and the involvement of and effects on students, teachers, and community members? One explanation for these particular foci is that sociocultural theory is the dominant paradigm in writing research (Behizadeh & Engelhard, 2011; Perry, 2012; Prior, 2006). Sociocultural theory emphasizes the complex and culturally connected contexts in which writing takes place and how writing differs based on context. If an assessment fails to honor the culture of a student and the context in which the student is learning, then this assessment has poor construct validity, which may lead to unintended negative consequences. For example, to understand a student's linguistic achievements, a valid assessment system should consider context and allow for multiple forms of writing with multiple purposes. This suggests that portfolio assessments may increase validity as compared to on-demand, direct writing assessment (Behizadeh, 2014). The IRA/NCTE assessment standards stress the theorized high validity of using portfolios for summative assessment due to the potential of this method to foster sound instructional practices.

Analysis of Research in the Teaching of English (RTE), 1999-2014

Based on the consensus definition of validity embodied in the IRA/NCTE (2010) standards, we now turn to how writing researchers employ this term. We chose a case study approach for examining how validity is used in practice within this community in order to support and/or complicate the notion of validity embodied in the standards. First, we analyze five key articles: Broad (2000), Ketter and Pool (2001), Bauer and Garcia (2002), Murphy, (2007) and Elliot el al. (2012).

Broad's (2000) piece described the scoring practices of faculty involved with a First-Year English college composition program. A valid writing assessment, according to Broad (2000), honors the rhetorical function of writing, a function that cannot be standardized. The key question of validity,

according to Broad, is: «How can we bring our assessments in line with our theories and pedagogies of rhetoric?» (p. 215). Thus we see the primacy of construct validity where the construct of writing is context-dependent. Regarding consequences, Broad (2000) stated:

> Cronbach (1990) and Messick (1989) argued that *consequential validity*—that is, analysis of the institutional and societal impact of a given assessment program—is as important as construct validity or predictive validity. Since evaluation unavoidably drives teaching and learning, authentic assessments such as… writing portfolios better account for the complexity and context within which learning took place and offer consequential validity by supporting best practices in composition pedagogy (pp. 250-251).

Similar to the IRA/NCTE standards, Broad (2000) emphasized both construct validity and consequential validity.

In Bauer and Garcia's (2002) study, they evaluated content validity (which is used interchangeably with construct validity to some extent within the writing research community), and presented the positive washback of a classroom literacy portfolio. Regarding content validity and alternative assessments, Bauer and Garcia stated: «These types of assessments finesse the issue of content validity by evaluating students' actual performance on domain related tasks» (p. 464). The authors contrast the hypothesized authenticity of alternative assessments, such as portfolios, with standardized tests, which «at best, serve as a proxy for performance» (p. 464). In the next section, Bauer and Garcia (2002) immediately address consequential validity, noting,

> Given that teachers in high stakes settings often teach to the test (Center for the Study of Testing, Evaluation, & Educational Policy, 1992), several researchers have proposed that alternative assessments, backed by public standards may result in greater student equity, especially for low-income and low-performing students (p. 464).

The authors continued, explaining that alternative assessments may increase equity by providing students access to high-quality instruction and assessment, as well as allowing for greater student voice in the educational process.

According to the two articles reviewed so far, the consequences of writing assessments are paramount, with Broad (2000) and Bauer and Garcia (2002) advocating for literacy portfolio assessment methods because of this method's positive impact on instruction. Although there are other ways of conceptualizing consequential validity, the primary focus of the writing research community appears to be on the washback from assessment to instruction.

Complementing Bauer and Garcia's (2002) argument that alternative student-centered assessments increase the equity of instruction, Ketter and Pool (2001) explored the negative impact of high-stakes writing tests on instruction. In particular, they examined two high school classrooms for struggling writers and how the high-stakes writing test affected instruction. In addition to citing Messick (1989) and Cronbach (1990), Ketter and Pool (2001) also cite Moss' (1994) work, specifically noting her expansion of validity to include consequential validity. Throughout the piece, the authors emphasize the consequences of high-stakes standardized writing tests —specifically, the negative effects of these tests on instruction. Again, consequential validity is emphasized, as in the other articles analyzed here and the IRA/NCTE (2010) standards.

Although Ketter and Pool (2001) highlighted consequential validity in their work, like other writing researchers, they noted the connection between construct and consequential validity, stating:

> Due to the specialized nature of many conversations about writing assessment in the measurement community, composition scholars find the rationales for direct writing assessment to be unnecessarily technical and finally irrelevant because they ignore a large body of scholarly literature that theorizes how people learn to read and write and that challenges the notion of writing ability as a fixed trait unaffected by context (p. 347).

Even when focus is on consequential validity, as in Ketter and Pool's piece, a key issue is how writing is being defined by these tests.

Compared to the above three pieces, Elliot et al. (2012) and Murphy (2007) exhibit significant differences. First, Murphy's (2007) article is a position piece on linguistic minority students and assessment while Elliot et al.'s (2012) research is a quantitative validation study using predictive and concurrent validity evidence, namely correlations between scores on a placement test and (a) final course grades and (b) other writing tests. Also, Elliot et al. (2012) and Murphy (2007) both cited the AERA, APA and NCME (1999) *Test Standards* rather than standards from a professional literacy organization. Elliot et al. (2012) drew extensively on the work of Kane (2006), a measurement theorist, situating both of these studies as bridges between the writing and measurement communities. However, like all of the research from *RTE* reviewed here, the authors articulated that a concern with construct validity is at the heart of their research. When explaining why they chose to compare the placement test scores with end of course portfolio scores, Elliot et al. (2012) stated, «the limited representation of the construct of writing in purchased tests and the manifestation of that construct in student portfolios was of great significance to us» (p. 290). Similarly, Murphy (2007) articulated that the language and literacy field has advanced its understanding of literacy as sociocultural and situated, while some measurement researchers continue to define reading and writing as discrete skills that can exist outside of considerations of language or culture. In terms of consequential validity, Elliot et al. (2012) discussed how their attention to predictive validity was based on concerns with the social consequences of using this placement test, while Murphy (2007) dedicated multiple pages to detailing the disparate impact of language tests on culturally and linguistically diverse students, citing the idea of «cultural validity» as a necessary source of validity evidence.

In sum, Broad (2000), Ketter and Pool (2001), Bauer and Garcia (2002), and Murphy (2007) reflect writing researchers' view regarding assessment with a major focus on consequential validity above all other concerns. Similarly, Elliot et al. (2012) highlighted consequential validity as a key concern, although they employed a different methodology. Across all pieces, a lack of content validity is positioned as the root of unintended consequences. These key articles are well-aligned with IRA/NCTE (2010) Standard 1: «The consequences of an assessment procedure are the first and most important consideration in establishing the validity of the assessment» (p. 22). Before turning to the special issue of *RTE,* it is important to note that these articles are outliers to some extent because historically the writing community has been more focused on instructional practices than assessment practices. This connects to the Cronbach (1988) quote from the beginning of this article: assessment may have been perceived by writing researchers as primarily in the purview of the measurement community. It is not until recently that the writing community has taken more ownership of exploring validity issues.

Special issue of Research in the Teaching of English (RTE)

An example of the heightened attention to validity in the writing research community is a recent special issue on writing assessment in *RTE*. Establishing the clear focus of this set of articles, Poe (2014) entitled her introduction to the volume (2014) «The Consequences of Writing Assessment.» In this introduction, Poe (2014) highlighted the historical development of new journals designed to bridge writing research and measurement research, noting two journals in particular: *Language Testing,* first published in 1984, and *Assessing Writing,* launched in 1994. Poe (2014) articulated the goals of these journals, stating:

> The desire to recognize multiple perspectives, including the perspectives of teachers, propelled theoretical advancement. Yet, despite new venues for recognizing multiple perspectives in assessing writing, other advances in the field, such as sociocognitive models of writing, were overlooked, and the local effects of writing assessment on diverse learners were often ignored (p. 272).

As in earlier *RTE* articles, Poe (2014) established a focus on consequential validity. Also, Poe (2014) cited Kane (2006), continuing the trend from Ketter and Pool (2001), Broad (2000), and Elliot et al. (2012). After referencing Kane, Poe (2014) continued, stating that «researchers today are focused on how the construct of writing must be reinterpreted for the demands of localized writing assessment» (p. 273). In this statement, the importance of construct validity is implied, although she does not provide specific details on how educational researchers should define writing.

All of the pieces in this special issue of *RTE* elevate consequential validity to some extent. In particular, Slomp et al.'s (2014) article, «A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study», highlights the consistent focus on consequences. The authors cited Messick (1989) and Cronbach (1988) to support their focus on consequential validity, and then argued that models of validity that center on construct validity are not sufficient for examining consequential validity, «because the consequences of tests and interpretation often extend beyond the scope of the intended interpretation and use of the test» (p. 278). The authors draw on Kane's (2006, 2013) validity model to develop their own framework for examining consequential validity in large-scale assessment, using writing assessment in Canada as a case study. Slomp et al. (2014) concluded their work in a very similar manner to Ketter and Pool (2001), stating:

> If the true goal of government-mandated, large-scale assessment programs is to improve systems of education, then it is only logical that the designers and users of these tests should examine and publicly report on the consequences that accrue as a result of the use of these tests (p. 298).

Throughout this section, writing researchers have stressed the importance of evaluating actual uses and consequences, an emphasis necessitating empirical research in classrooms to examine the effects of writing assessments on instruction and students.

What is a valid writing assessment from the perspective of the measurement community?

Early measurement theorists used an apparently straightforward definition of a valid scale as «one in respect to whose meaning all competent thinkers agree» (Thorndike, 1919, p. 11). Although it is beyond the scope of this study to provide a detailed history regarding how the concept of validity evolved within the measurement community, there are several key perspectives that should be highlighted. At the start of the 20th century, Thorndike (1914) defined valid scales essentially as consensus on meaning of scores among a community of practitioners. Later, Kelley (1927) defined valid tests in terms of agreement over what the scales *purport* to measure. Both Thurstone (1931) and Gulliksen (1950) suggested that the validity of tests be evaluated using criterion-related evidence based on empirical data. Cronbach (1971) redirected the narrative about validity from valid scales to a focus on *test validation,* which is defined as an evaluative process based on the intended purpose and recommended uses of test scores. In his groundbreaking work, Messick (1989) proposed a broad perspective on construct validity that included adding a consideration of consequential validity evidence. More recently, Kane (2013) has advocated for an argument-based approach to validity (see Engelhard & Behizadeh, 2012 for more detail about the historical and philosophical iterations of the concept of validity).

Even though this is a widely accepted view of the evolution of the concept of validity, the measurement community continues to disagree about certain aspects of the term validity (Engelhard & Behizadeh, 2012; Newton, 2012). In the following section, we describe the consensus definition of validity embodied in the *Test Standards* (AERA et al., 2014). Then, we use a recent special issue of the *Journal of Educational Measurement* to represent the ongoing discussion of validity within the measurement community before comparing perspectives on validity across communities. Finally, we illustrate how advances in measurement represented by Rasch measurement theory (Engelhard, 2013; Rasch, 1960/1980) can be used to provide a coherent consideration of validity evidence in general, as well as specifically within the context of writing assessment.

Measurement community: What is validity?

> Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests (AERA et al., 2014, p. 11).

The current consensus definition of validity as outlined in the newly revised *Test Standards* (AERA et al., 2014) is as follows: «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). This guiding principle is divided into three thematic clusters:

I. Establishing intended uses and interpretations,
II. Issues regarding samples and settings used in validation, and
III. Specific forms of validity evidence.

These three clusters are shown in Figure 2. This figure highlights the idea that validity cannot be evaluated in a vacuum without specifically considering the intended uses and interpretations (Cluster I), as well as the context of the assessment system related to the persons and setting involved in the validation process (Cluster II). However, many discussions of validity primarily (often exclusively) focus on the last cluster (Cluster III). To some extent, both the measurement and writing communities tend to focus on specific forms of validity evidence, such as content, construct, criterion-related and consequential without an explicit discussion of intended uses and interpretations. Readers should consult the *Test Standards* for a more complete description of each cluster.

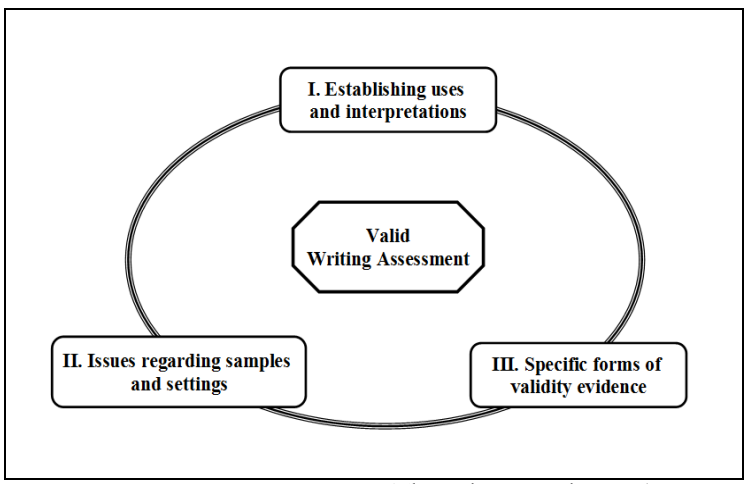

Figure 2. Measurement community: Three thematic clusters (AERA et al., 2014).

Establishing the intended uses and interpretations of an assessment system is essential for evaluating the validity of the interpretation and use of test scores from the perspective of the *Test Standards*. Educational testing, including writing assessments, is designed to serve a variety of purposes. In spite of the importance of establishing the purpose of the assessment system, it is surprising how blurry the purposes of many educational tests, including writing assessments, appear to various stakeholders.

The second thematic cluster stresses that issues regarding validation are context-dependent, and that as much detail as possible should be provided about the participants in the assessment system. This includes the demographic characteristics of the samples and settings. This focus on context is to some degree congruent with a sociocultural perspective on writing, although many members of the measurement community may not be aware of the full implications of including this thematic cluster in presenting validity evidence. IRA/NCTE (2010) standards indicate that a truly sociocultural assessment would involve all stakeholders, including students and families, in the full validation process from development to reporting. This is an area where collaboration across communities has the potential to yield significant improvements in our conceptual understanding of the intended and unintended consequences of writing assessments.

The final thematic cluster includes six forms of validity evidence. These six forms are:

- Content-oriented evidence,
- Evidence regarding cognitive processes,
- Evidence regarding internal structure,
- Evidence regarding relationships with conceptually related constructs,
- Evidence regarding relationships with criteria, and
- Evidence based on consequences of tests.

The *Test Standards* (AERA et al., 2014) highlight the importance of purpose in any assessment system, and the collection of validity evidence to support the intended uses and interpretations of test scores. It is also important to consider intended and unintended consequences of assessment systems (Engelhard & Wind, 2013).

Special issue on validity: Journal of Educational Measurement (JEM)

One of the major professional organizations in educational measurement is the National Council on Measurement in Education in the United States, and their premier journal is the *Journal of Educational Measurement* (*JEM)*. A suite of articles recently published in a special issue on validity in *JEM* featured an article by Kane (2013) reflecting an update on his views of an argument-based approach to validation (Kane, 1992, 2006), and included pieces by Sireci (2013) and Borsboom and Markus (2013). Kane's (2013) argument-based perspective is congruent with earlier views of Cronbach (1988):

> Ideally, validators will prepare as debaters do … preparing arguments, pro or con so well that he or she could speak for either side. Or, shifting the metaphor to legal counselors, so well that they could tell either party what is strong and weak in its position (p. 3).

In contrast to this argument-based perspective on validity, Kane (2013) refers to Borsboom and his colleagues, who have proposed an attribute-, trait-, and construct-based view of validation (Borsboom, 2005; Borsboom et al., 2004; Borsboom & Markus, 2013). They stated that «a *test is valid* for measuring an attribute if and only if (a) the attribute exists, and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure» (Borsboom et al., 2004, p. 150), an idea echoed in Borsboom and Markus (2013) in the special issue. This attribute-based perspective on validity has different implications for defining the validation process used to evaluate an assessment system. For example:

> What needs to be tested is not a theory about the relation between the attribute measured and other attributes but a theory of response behavior. Somewhere in the chain of events that occurs between item administration and item response the measured attribute must play a causal role in determining what value the measurements outcomes will take; otherwise the test cannot be valid for measuring the attribute (Borsboom et al., 2004, p. 1062).

This attribute-based approach to validity is aligned with earlier views of validity regarding constructs as an underlying latent variable that an assessment is designed to measure. Borsboom, Mellenbergh, and Van Heerden (2004) repositioned the test, rather than the interpretation or use, as valid, contrasting with the argument-based approach currently dominating the revised *Test Standards* (AERA et al., 2014) that appears to have moved the measurement community away from an attribute and construct driven view of validity.

On the other hand, Sireci (2013) has offered what we consider a promising reconceptualization of the validation process that connects closely to cluster three from the *Test Standards*. In his comments on Kane (2013), Sireci (2013) states that:

> [A]n explicit statement of testing purposes is the logical beginning of validation, but I go one step further —it is the logical beginning of developing a test. That is, tests are developed to fulfill one or more intended purposes. It is incumbent upon us as psychometricians to help those who commission these tests to articulate the intended purposes. Once these purposes are articulated, we know what we need to validate. We also know what it is we need to measure! (p. 100).

Sireci (2013) identified three steps in what he terms a «validation plan» that include the (a) clear articulation of purposes, (b) consideration of potential test misuses, and (c) the crossing of purposes with the collection of validity evidence. These steps are based directly on Kane's (2013) more complicated validation process, yet Sireci collapses interpretation and use into one term: purpose. Sireci's (2013) recommendations can be applied to the specific context of writing assessment in order to determine which forms of validity evidence are required.

What are some points of consensus and debate regarding the concept of validity?

> Validation is the joint responsibility of the test developer and the test user (AERA et al., 2014, p. 22).

There appears to be a general consensus in the writing community regarding the potential for negative unintended consequences of standardized writing assessments on instruction. Consequential validity is a clearly a key concern in the writing community, particularly for K-12 assessment as represented by theory, standards, and empirical research. On the other hand, the recent publication of the *Test Standards* (AERA et al., 2014) provides a consensus framework for defining validity, but this publication does not fully capture the nuances of the continuing debate regarding this foundational concept as evidenced by a special issue on validity in the *Journal of Educational Measurement* (2013). Furthermore, the *Test Standards* do not fully develop how these guidelines should be implemented within a variety of assessment areas, such as writing assessment. It is quite common to see much of the psychometric research on writing assessment focusing on only rater agreement indices without a broader perspective of validity as espoused by the measurement community.

Reflecting back on the standards of both communities, we believe that the following perspective on validity from the *Test Standards* may be attractive to members of both communities:

> Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. Different components of validity evidence are described in subsequent chapters of the *Standards*, and include evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; and careful attention to fairness for all test takers, as appropriate to the test interpretation in question (AERA et al., 2014, p. 22).

This conception of validity aligns with the IRA/NCTE (2010) standards to some extent with a clear focus on the test taker as a key stakeholder in the assessment process. However, there may exist large gaps between how measurement and writing researchers conceptualize what is meant by *careful* test development, how reliability is defined, what is *appropriate* administration, and perhaps most important, what is meant by *fairness*. For example, if fairness is equated with standardization in the measurement community (Madaus, 1994), this conception conflicts with a view of fairness prevalent in the writing community as allowing for multiple perspectives on the entire testing process to be voiced (IRA/NCTE, 2010).

Writing and measurement researchers offer unique affordances based on their theoretical frameworks and their positioning in regards to the validation process. We see great potential for more collaboration across these communities and already detect a fair amount of theoretical borrowing from the measurement community by writing researchers. Yet we hope to see this borrowing become more of a reciprocal process, especially in regard to ensuring the equity of writing assessments for culturally and linguistically diverse students (Murphy, 2007). In particular, writing researchers are in the best position to evaluate consequential validity, especially unintended negative consequences, through empirical research in collaboration with classroom teachers. Measurement researchers can offer technical advice and suggestions for research on writing assessment, including examining measurement invariance, such as differential item functioning, a point explored in the next section. The revised *Test Standards* (AERA et al., 2014) make it very clear that validity is a joint responsibility for test developers and users and the IRA/NCTE (2010) standards emphasize the importance of all stakeholders being involved, including the local communities in which assessments take place. Both research communities should view this as an opportunity to improve the theory and practice of writing assessment.

Our analyses suggest that writing researchers are most concerned with *actual* uses of validity evidence, while measurement researchers are more concerned with establishing the *intended* use as a source of validity evidence. We believe this key difference in conceptualizing consequential validity derives from differences in the perceived purposes of large-scale writing assessments. The Sireci (2013) framework has potential for informing the emergent narratives about validity from the communities of writing and measurement. Additionally, Slomp et al. (2014) proposed a promising framework for examining consequential validity in large-scale writing assessments that can serve both communities. These two frameworks are explored below.

Three of Slomp and colleagues' (2014) recommendations are particularly compelling for our discussion. The first is based on the observation that stakeholders farthest from the classroom viewed large-scale assessment positively, while teachers and students tended to have negative attitudes toward large-scale testing. It seems that these disagreements are primarily a function of perceived difference in the intended purposes of writing assessments. Second, Slomp et al. (2014) called for increased conversation across writing and measurement communities, a goal we are trying to enact in this article. Finally, they have suggested that evidence be collected related to consequences using their framework.

There is a nice correspondence between Sireci (2013), with his focus on purpose and validity evidence, and Slomp et al.'s (2014) proposal for examining consequential validity evidence. Sireci's (2013) crossing of the six types of validity evidence with intended and unintended consequences of an assessment offers a tool that can be used across the measurement and writing communities for a joint process of validation. Our rationale for choosing Sireci's validation plan is that his plan includes the thorough theoretical basis outlined by Kane (2013), yet simplifies the validation process. This simplification may allow this tool to be more portable across communities and into the writing community.

In sum, both research communities have developed valuable conceptual tools for improving valid assessment processes —tools that may be even more potent when used together. Sireci's (2013) framework can easily work in tandem with Slomp et al.'s (2014) consequential validity question matrix. Table 1 illustrates the crossing of validity evidence and consequences for a hypothetical writing assessment using the specific forms of validity evidence recommended by the updated *Test Standards* (AERA et al., 2014) and illustrated by Sireci (2013), with the addition of questions adapted from Slomp and colleagues' (2014) question matrix. If the purpose of a writing assessment is to determine levels of writing proficiency for summative assessment, then we have suggested several questions in Table 1 that should be examined relative to intended and unintended consequences. We strongly recommend that unintended consequences be explicitly explored along with each type of validity evidence, and that both hypothesized and actual consequences be investigated and documented. This table is meant to be illustrative and it is by no means exhaustive regarding types of evidence, consequences, or underlying validity questions. We believe that Table 1 provides a starting point for developing a division of labor between these two communities regarding areas of expertise in theory and methodologies. For example, the writing community is likely to be involved with teachers of writing and can provide useful insight into how writing assessments are actually being used in classrooms, including the evaluation of both intended and unintended consequences.

Table 1
Validity evidence crossed with intended and unintended consequences

Purpose of the writing assessment: Determine levels of writing proficiency with a written essay		
	Consequences	
Specific forms of validity evidence	Intended	Unintended
Content-oriented evidence	Do the test scores represent the construct of writing?	Do the test scores represent the construct of writing for students from varying sociocultural contexts?
Evidence regarding cognitive processes	Do students report engaging in the expected thoughts and behaviors related to the construct of writing?	Do students develop misconceptions about writing?
Evidence regarding internal structure	Do the components of the writing assessment system support the inferences made regarding the construct of writing?	Do the components of the writing assessment system replicate across different sociocultural contexts?
Evidence regarding relationships with conceptually related constructs	Are the writing scores associated with other related constructs?	Are the writing scores associated with other construct-irrelevant variables?
Evidence regarding relationships with criteria	Are writing scores associated with other indicators of literacy?	Are writing scores associated with other indicators of literacy for students from various sociocultural contexts?
Evidence based on consequences of tests	Do students perform as expected on other tasks requiring writing proficiency?	Do teachers teach to the test and limit the curriculum related to writing?

Note: Table based on Sireci (2013) and Slomp et al. (2014).

The purposes of an assessment system define both direct and indirect uses with potentially unintended consequences whenever a writing assessment is put in place within complex settings. As both the measurement and writing community engage in conversations about validity, they should pay more attention to explicitly expressing the purposes of writing assessment and examining the potentially contradictory implications of multiple purposes. As noted earlier, a key theme that emerged was the contrast between intended use and actual use, with Kane (2013), Sireci (2013), and other measurement scholars stressing intended uses and hypothesized misuses or unintended consequences, while writing scholars often explored actual uses and documented realized unintended negative consequences (e.g. Slomp, Corrigan, & Sugimoto, 2014). In addition to the integration of Sireci's (2013) and Slomp et al.'s (2014) work in Table 1 as a tool for exploring intended and actual consequences, the tools of Rasch measurement theory, detailed in the next section, are particularly well-suited for these explorations.

Validity evidence from the perspective of Rasch measurement theory

This section briefly focuses on modern measurement theory as described by Embretson (1996) using Rasch measurement theory (Engelhard, 2013; Rasch, 1960/1980). The *Test Standards* tend to be neutral in terms of which measurement theories are used to provide validity evidence for the intended uses of scores obtained from an assessment. Rasch measurement theory provides the opportunity to explicitly examine the responses of each student to each item or task in an assessment system, yielding validity evidence that can be used to evaluate purpose (Sireci, 2013) and the actual consequences (Slomp et al., 2014) of assessments. The overall goal of assessment systems is to create invariant measurement with student scores independent of specific raters and tasks. In our work, we have found it useful to use Rasch measurement theory as the underlying measurement model that serves as a building block in Wilson's (2005) constructing measures framework (Engelhard, 2013). Wilson's

framework can be combined with Rasch measurement theory to systematically collect validity evidence based on the six forms of validity evidence described in the *Test Standards* (Duckor, Draney, & Wilson, 2009) and emphasized in Sireci's (2013) work (see Table 1), and qualitative data can be employed to support quantitative analyses.

It is essential to collect content-oriented evidence to support the use of test scores from a writing assessment to determine levels of writing proficiency (emergent, basic, proficient, and advanced). Content-related evidence reflects the alignment between the content of the assessment and the construct intended to be represented by the test scores. Based on Rasch measurement theory and the constructing measures framework, there are several steps, including (a) definition of the construct of writing, (b) description of the prompts and writing tasks, (c) rules for rating student essays (rubrics), and (d) the creation of a variable map. Wilson (2005), Duckor, Draney, and Wilson (2009), and Engelhard (2013) describe in detail how these steps can be systematically combined to provide content-related evidence. We envision these steps being undertaken as a joint enterprise (Wenger, 2010, 2015) between the two communities, with experts from writing and measurement fields engaging in interdisciplinary conversations that can represent the needs of different stakeholders associated with large-scale writing assessment. Additionally, Wind, and Engelhard (2012) include variable maps that illustrate how content-oriented information can be represented visually to aid in supporting the intended meaning and use of scores from a rater-mediated writing assessment. Variable maps are an important visual tools that can be analyzed by both communities to help develop a common vocabulary and shared understanding of writing assessment.

Another source of validity evidence based on the internal structure of the assessment includes examining the mode of assessment being used. If writing is defined in terms of selected-response items (e.g., multiple-choice items), then there are a variety of well-established psychometric methods that can be used. However, the use of constructed-response items such as essays and portfolios —assessments that have more support in the writing research community— requires additional data about raters and the rating process. Rasch measurement theory also provides a clear set of criteria that can be used to evaluate the quality of the ratings (Engelhard, 2002). As is well known, other item response theory models do not provide the simultaneous mapping of persons, items and raters on a single underlying continuum (Engelhard, 2013).

Evidence regarding relationships with conceptually related constructs can be collected in a variety of ways. Evidence within this category includes research studies using the scores obtained from the writing assessment as variables in a broader framework, such as a nomothetic network (Whitely, 1983). For example, Behizadeh and Engelhard (2014) developed an instrument for examining student perceptions of the authenticity of writing that may be related to the scores on a writing assessment. Additionally, Differential Item Functioning (DIF) (Engelhard, 2009) can illuminate to some extent possible an actual adverse impact (Kane, 2013) on students of particular genders, ethnicities, or achievement levels. Rasch measurement theory thus offers analytic tools that can serve to examine intended and actual uses related to the consequences of testing, and we envision the quantitative and visual data generated from Rasch analyses being analyzed simultaneously with qualitative data such as interview and observation data from empirical classroom research.

Evidence-based consequences of tests can be obtained in a variety of ways as well. Slomp et al. (2014) provide a framework that should be carefully considered as a prototype for conducting consequential validity studies. The measurement community approaches consequential validity by analyzing DIF and Differential Person Functioning (DPF). These two approaches can be used to examine the fairness of assessments, including the actual consequences for individuals and subgroups of students. For example, Engelhard (2009) has used Rasch measurement theory to describe how DIF and DPF can be used to examine and improve the validity of score interpretations for individuals and subgroups of students.

In conclusion, we believe that the measurement community can offer excellent tools, such as a validation plan and variable maps, for collecting and analyzing the broad array of data necessary for validating writing assessments. However, we envision this validation process incomplete unless the writing research community, as well as teachers, students, and other stakeholders, are part of the joint enterprise of developing a valid assessment system, and this collaboration includes contributing important qualitative data on the impact of large-scale writing assessment on classroom literacy instruction.

Discussion and conclusions

> Validators are also a community. That enables members to divide up the investigative and educative burden according to their talents, motives, and political ideals. Validation will progress in proportion as we collectively do our damnedest —no holds barred— with our minds and our hearts (Cronbach, 1988, p. 14).

This study contributes to clarifying the concept of validity and includes suggestions for future collaboration between two communities (measurement and writing) that focus on issues related to writing assessment. We considered the following questions:

1. What is a valid writing assessment from the perspective of the writing community?
2. What is a valid writing assessment from the perspective of the measurement community?
3. What are some points of consensus and disagreement over the concept of validity in the two communities?

The writing community would answer the first question by focusing on construct and consequential validity. Specifically, they would (a) articulate a sociocultural view of writing that conceives of the writing construct as a set of practices that are context-dependent and (b) elevate consequential validity evidence derived from investigation of actual consequences as the most important source of validity evidence. The measurement community would answer the second question by referring to the consensus definition of validity included in the *Test Standards*. Some members of the measurement community, as demonstrated in this article, conceive of the validation process differently from the consensus definition. Kane's (2013) argument-based approach is congruent with the consensus definition, while Borsboom et al. (2004) stress an attribute-based approach based on examining the effects of attributes, latent variables, and constructs on variation in person responses.

The answer to the third question is more nuanced. There is a consensus regarding the importance of content and consequential validity, but each community has different areas of expertise that contribute to the validation process. When we conceptualized this study, we expected more debate about validity between the two communities; instead, we found a moderate degree of consensus with the potential for a more systematic division of labor regarding how to collect and evaluate the validity of writing assessments. However, a critical discussion in which research communities should engage would consider how to define the construct of writing. The writing community consistently employs a sociocultural understanding of writing, and its critique of standardized writing assessments is that these assessments are not aligned with this understanding. Additionally, another critical discussion for both communities revolves around the purpose of writing assessment for primary and secondary students. As outlined in this article, writing research emphasizes assessment for improving teaching and learning, while the measurement community has frequently been asked to develop assessments for placement and summative evaluation. We believe that by examining multiple tools offered by both communities, including Slomp et al.'s (2014) consequential validity framework and Sireci's (2013) validation framework, we can collaboratively establish a consensus definition on writing as well as the purposes of writing assessment.

Finally, unexplored in this article are the emergent validation issues related to automated essay scoring (AES), such as the research represented in Landauer, McNamara, Dennis, and Kintsch (2013) and Shermis and Burstein (2003), and critiques of AES including Deane (2013) and Perelman (2014). Future work examining the validity of automating scoring practices can use as a starting point the suggestions for an integrated framework represented by Table 1 to engage all stakeholders in the validation process.

There is much to be gained from the disciplinary expertise of each community related to writing assessment. There are unique affordances from each community, and the field of writing assessment will benefit from contributions made by both the writing and measurement communities. We look forward to the next epistemic iteration regarding the definition of validity with strong voices from these separate yet overlapping communities of practice.

The original article was received on December 19th, 2014
The revised article was received on April 14th, 2015
The article was accepted on May 13th, 2015

References

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

Bauer, E. B., & Garcia, G. E. (2002). Lessons from a classroom teacher's use of alternative literacy assessment. *Research in the Teaching of English 36*(4), 462-494.

Behizadeh, N. (2014). Mitigating the dangers of a single story: Creating large-scale writing assessments aligned with sociocultural theory. *Educational Researcher*, *43*(3), 125-136. doi: 10.3102/0013189X14529604

Behizadeh, N., & Engelhard, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment. *Assessing Writing, 16*(3), 189-211. doi: 10.1016/j.asw.2011.03.001

Behizadeh, N., & Engelhard, G. (2014). Development and validation of a scale to measure perceived authenticity in writing. *Assessing Writing, 21*, 18-36. doi:10.1016/j.asw.2014.02.001

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge: Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071. doi: 10.1037/0033-295x.111.4.1061

Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement*, *50*(1), 110-114. doi: 10.1111/jedm.12006

Broad, B. (2000). Pulling your hair out: Crises of standardization in communal writing assessment. *Research in the Teaching of English, 35*(2), 213-260.

Center for the Study of Testing, Evaluation, & Educational Policy (1992). *The influence of testing math and science in grades 4-12* (Vols. 1-5). Boston: Author.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cronbach, L. J. (1990). *Essentials of psychological testing*. (5th ed.). New York, NY: Harper.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7-24. doi: 10.1016/j.asw.2012.10.002

Duckor, B. M., Draney, K., & Wilson, M. (2009). Measuring measuring: Toward a theory of proficiency with the constructing measures framework. *Journal of Applied Measurement, 10*(3), 296-319.

Elliot, N., Deess, P., Rudniy, A., & Joshi, K. (2012). Placement of students into first-year writing courses. *Research in the Teaching of English, 46*(3), 285-313.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349. doi: 10.1037/1040-3590.8.4.341

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal, & T. Haladyna (Eds.), *Large-scale Assessment Programs for ALL Students: Development, implementation, and analysis* (pp. 261-287). Mahwah, NJ: Erlbaum.

Engelhard, G. (2009). Using IRT and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement, 69*(4), 585-602. doi: 10.1177/0013164408323240

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Engelhard, G., & Behizadeh, N. (2012). Epistemic iterations and consensus definitions of validity. *Measurement: Interdisciplinary Research and Perspectives*, *10*(1), 55-58. doi: 10.1080/15366367.2012.681974

Engelhard, G., & Wind, S. A. (2013). Educational testing and schooling: Unanticipated consequences of purposive social action. *Measurement: Interdisciplinary Research and Perspectives, 11*(1-2), 30-35. doi: 10.1080/15366367.2013.784156

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

International Reading Association (IRA), & National Council of Teachers of English (NCTE) (2010). *Standards for the assessment of reading and writing*. (IRA Stock No. 776; NCTE Stock No. 46864). USA: Authors.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527-535. doi: 10.1037//0033-2909.112.3.527

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Kelley, T. (1927). *Interpretation of educational measurements*. Yonkers, NY: World Book Co.

Ketter, J., & Pool, J. (2001). Exploring the impact of a high-stakes direct writing assessment in two high school classrooms. *Research in the Teaching of English, 35*(5), 344-393.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2013). *Handbook of latent semantic analysis.* New York, NY: Psychology Press.

Madaus, G. F. (1994). A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy. *Harvard Educational Review, 64*(1), 76-95. doi: 10.17763/haer.64.1.4q87663 r0j76rwv1

Messick, S. (1989). Meaning and values in test validation. The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11. doi: 10.3102/0013189x018002005

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi: 10.1037//0003-066x.50.9.741

Moss, P. A. (1994). Can there be validity with reliability? *Educational Researcher, 23*(2), 229-258. http://dx.doi.org/10.3102/0013189x023002005

Murphy, S. (2007). Culture and consequence: The canaries in the coal mine. *Research in the Teaching of English, 42*(2), 228-244.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, *10*(1), 1-29. doi: 10.1080/15366367.2012.669666

Perelman, L. (2014). When the «state of the art» is counting words. *Assessing Writing, 21*, 104-111. doi: 10.1016/j.asw.2014.04.001

Perry, K. (2012). What is literacy? A critical overview of sociocultural perspectives. *Journal of Language and Literacy Education*, *8*(1), 50-71. Retrieved from http://jolle.coe.uga.edu/wp-content/uploads/2012/06/What-is-Literacy_KPerry.pdf

Poe, M. (2014). The consequences of writing assessment. *Research in the Teaching of English, 48*(3), 271-275.

Prior, P. (2006). A sociocultural theory of writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 54-66). New York, NY: Guilford Press.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A crossdisciplinary perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, *50*(1), 99-104 http://dx.doi.org/10.1111/jedm.12005

Slomp, D. H., Corrigan, J. A., & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study. *Research in the Teaching of English*, 48 (3), 276-302.

Thorndike, E. L. (1914). The measurement of ability in reading. *Teachers College Record, 15*(4), 207–277.

Thorndike, E. L. (1919). *An introduction to the theory of mental and social measurements. Revised and enlarged edition*. New York, NY: Teachers College, Columbia University.

Thurstone, L. L. (1931). *The reliability and validity of tests*. Ann Arbor, MI: Edwards.

Wenger, E. (1998). *Communities of practice: Learning, meaning and identity*. Cambridge, UK: Cambridge University Press.

Wenger, E. (2010). Communities of practice and social learning systems: The career of a concept. In C. Blackmore (Ed.), *Social learning systems and communities of practice* (pp. 179-198). London: Springer Verlag and the Open University.

Wenger, E. (2015). *Communities of practice: A brief introduction*. Retrieved from http://wenger-trayner.com/introduction-to-communities-of-practice/

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197. doi: 10.1037/0033-2909.93.1.179

Wilson, M. R. (2005). *Constructing measures: An item response theory approach*. Mahwah, NJ: Lawrence Erlbaum.
Wind, S. A., & Engelhard, G. (2012). Evaluating the quality of ratings in writing assessment: Rater agreement, precision, and accuracy. *Journal of Applied Measurement*, *13*(4), 321-335.