Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 18, Number 9, June 2013 ISSN 1531-7714

Validity Semantics in Educational and Psychological Assessment

John D. Hathcoat
James Madison University

The semantics, or meaning, of validity is a fluid concept in educational and psychological testing. Contemporary controversies surrounding this concept appear to stem from the proper location of validity. Under one view, validity is a property of score-based inferences and entailed uses of test scores. This view is challenged by the instrument-based approach, which contends that tests themselves are either valid or invalid. These perspectives are contrasted by their ontological and epistemological emphases, as well as their breadth of validation focus. Ontologically, these positions diverge in their alliance with psychometric realism, or the position that attributes characterizing the aim of psychological and educational measurement exist in the actual world and that claims about their existence can be justified. Epistemologically, these positions deviate in the function of truth when accepting validity claims and inform distinct lines of inquiry in the validation process. Finally, validity under the instrument-based approach is restricted to a single proposition –namely, that observed score variation is caused by an underlying attribute. Though seemingly arbitrary, these distinct validity semantics may have a range of implications on assessment practices.

A test is valid if it measures what it intends to measure. Various textbooks repeat this statement despite a number of theorists who argue that this position oversimplifies the concept of validity as well as the validation process (see Lissitz, 2009). However, the semantics, or meaning, of validity is controversial in academic discourse. This controversy has a long history, though the contemporary debate appears to stem from disagreements about the proper location of validity. One view locates validity as a property of score-based interpretations and entailed uses of test scores (Messick, 1989; Kane, 1992). This position has come to dominate validity theory, as evidenced by the most recent standards for educational and psychological testing resembling this perspective (AERA, APA, NCME, 1999). However, the "instrument-based" approach challenges this view by locating validity as a property of tests themselves (Borsboom, 2005; Borsboom, Mellenbergh, & van Heerden, 2004). Formal testing is but one aspect of educational and psychological assessment. Nevertheless, these two perspectives have important ramifications for assessment practitioners.

The concept of validity is central to assessment processes, data-driven decisions, and reporting procedures (Moss, Girard, & Haniford, 2006). Validity semantics dictate which inferences one may legitimately label "valid" or "invalid". Moreover, semantic differences have consequences on the aims of score validation – namely, what evidence *should* one seek in the validation process. The present paper delineates validity semantics by contrasting their ontological and epistemological emphases, as well as their breadth of validation focus. Ontology is a branch of metaphysics aiming to ascertain the underlying structure of reality (Poli, 2010). Epistemology on the other hand, studies the nature, limitations, and justification of knowledge (Williams, 2001). Ontological questions pertain to "what exists," whereas epistemological questions tend to focus on the possibility and process of obtaining knowledge. For example, do observed scores reflect differences in a "real" attribute? Is such knowledge possible, and if so, how are such claims justified or warranted? These philosophical questions are fundamental to validity theory, which encompasses both ontological (e.g. reality of attributes) and epistemological (e.g. evidential standards) aspects of score-based interpretations.

Central to this paper is the concept of psychometric realism. Psychometric realism refers to

the view that attributes characterizing the aim of psychological and educational measurement exist in the actual world *and* that claims about their existence can be justified (see Hood, 2009). Psychometric realism has both an ontological and epistemic component. Ontologically, a psychometric realist views attributes such as "personality," "critical thinking," and "intelligence" as entities that exist within the world. Importantly, a psychometric realist also believes that it is possible to justify claims about the existence of these entities. An antirealist would deny at least one of these positions. For example, an antirealist may deny the ontological status of attributes or the possibility of warranting claims about their existence.

There are many points of contention with respect to validity semantics in educational and psychological assessment (Moss et al., 2006; Newton, 2012). However, this discussion is delimited by the work of two prominent validity theorists. First, the argument-based approach to validity is described (Kane, 1992, 2006) given that this view coincides with many of the historical transformations characterizing validity semantics. Following this account is the critique offered by Borsboom et al., (2004), which falls under the instrument-based approach. With respect to the ontological, these views diverge in whether they require adherence to psychometric realism. Likewise, these perspectives deviate in their epistemological characteristics, such as the role accorded to truth in validity theory. Finally, the instrument-based approach has a relatively narrow validation focus when compared to the argument-based approach. Before proceeding to this examination, the following section provides a brief overview of the historical evolution of validity semantics in educational and psychological testing.

A Brief History of Validity Semantics

The application of statistical concepts, such as the theory of errors, to the measurement of mental phenomena can be traced to the later part of the 19th Century (see Traub, 2005). However, it was not until the early 1900's that validity, as a formal concept, became a point of vivid discussion. Efforts to articulate a theory of validity occurred within educational and psychological testing. Other academic disciplines share concerns with measurement error (Taylor, 1997); however, the "physical" sciences lack an analogous discourse aiming to formulate a formal theory of validity. Validity semantics, it would seem, is primarily a concern among academicians who have sought to apply measurement models to intraindividual and interindividual variation in social research. Numerous authors have provided an overview of the historical evolution of validity theory within these disciplines (see Kane, 2001, 2009; Lissitz & Samuelsen, 2007; Moss, Girard, & Haniford, 2006; Sireci, 1998, 2009). This section broadly outlines this development, while emphasizing some of the philosophical implications of distinct validity semantics.

Prior to the 1940's, predicting subsequent performance was a primary concern among test developers. During this time, various advancements in statistical techniques were gaining widespread acceptance in the social and behavioral sciences. For example, this period saw advancements in correlation coefficients (Pearson, 1920; Rodgers & Nicewander, 1988), and factor analysis was developed to address theoretical issues relevant to intelligence testing (Spearman, 1904). Each of these advancements however, had slightly different implications on validity semantics, which in turn, contributed to distinct positions toward psychometric realism. Covariation between two variables is fundamental to both prediction and causation, though concerns with the former reinforced a view of validity wherein predicting criterion variables were paramount. Tests were generally valid for any criterion that it predicted (Cureton, 1951; Guilford, 1946). At least two challenges existed with this view.

The first difficulty is the problem of the criterion (Chisolm, 1973; Amico, 1995), which pertains to the challenge of answering questions about what we know and how we know. To recognize instances of knowledge it seems that we must have a criterion or procedure. To illustrate this point, Chisholm (1973) refers to our ability to determine the quality of apples. If we wish to identify apple quality then we need a criterion to distinguish "good" apples from "bad" apples. We may choose to sort apples into different piles based upon their color, though any criterion is adequate for this example. The problem arises whenever we ask whether our criterion worked in that

color actually separated good apples from bad apples. How can we investigate our criterion without already knowing something about which apples are good and bad? In order to evaluate the criterion of color it seems that we need prior knowledge about apple quality, which would itself depend on a different criterion. Simply put, we appear caught within a vicious circle wherein recognizing instances of knowledge requires a criterion that cannot be evaluated without existent knowledge.

When applied to educational and psychological testing, this minimally implies that one must assume the validity of a criterion variable when making inferences about the validity of a newly developed instrument. Questions about the validity of a criterion variable require further assumptions about a new criterion, and this could easily lead to an infinite regress (Kane, 2001). This position towards validity therefore appears more reasonable when the criterion variable is intrinsically valid (Gulliksen, 1950), which may occur when using scores to predict performance on tasks that are either the same, or very similar to, specific demands within occupational and educational settings. This leads to the second challenge for predictive or criterion-related validity, which consists of identifying adequate criterion variables. In some situations, identifying an acceptable criterion is problematic. This is particularly the case for many concepts investigated within educational and psychological testing.

Around this same period, various authors discussed the importance of test content within validity theory. For example, there are connections between content validity and criterion related validity (Lissitz & Samuelsen, 2007). Tests are constructed for a specific purpose (Rulon, 1946), and focusing on test content reinforced the view that items/tasks constitute samples from a theoretical universe of possible tasks (see Sireci, 1998). An emphasis on test content, coupled with the prior development of factor analysis (Spearman, 1904), promoted a conceptual and statistical framework for advancing psychometric realism (Mulaik, 1987). In other words, the question of validity, though multifaceted, aimed at investigating "whether a test really measures what it purports to measure" (Kelley, 1927, p. 14). Factor analysis accords with the notion that real attributes underlie variation in observed scores. Though such techniques seem aligned with psychometric realism, many critics have argued that factor analysis is an instrumental tool that simplifies empirical observations (e.g. Anastasi, 1938). Under this latter view, extracted factors reflect useful ways to summarize observations without necessitating reference to actual entities. If extracted factors reference entities in the actual world, then such procedures have the potential to provide insight into the world, as it exists.

These early controversies set the stage for broad distinctions about the proper location of validity. However, a seminal article by Cronbach and Meehl (1955) eventually led to radical changes in validity semantics, which support a departure from psychometric realism. Cronbach and Meehl were concerned about situations wherein a target domain and/or a relevant criterion remained ambiguous. For example, if an instructor were creating a geometry test the course objectives may constitute a target domain from which item sampling occurs. However, other concepts of interest, such as anxiety, may cease to have a clear domain from which to sample items or an unambiguous criterion for investigating subsequent predictions. In these latter cases, Cronbach and Meehl argued that what is necessary is the establishment of *construct validity*. However, their conceptualization of construct validity relies upon the construction of a *nomological network*, and this network aligns with logical positivist aims to create a theory of scientific knowledge without metaphysical assumptions. Stated differently, construct validity as articulated by Cronbach and Meehl allows one to infer the meaning of unobservable constructs without requiring them to exist as an entity within the world.

To understand this effort, it is first necessary to make a distinction between theoretical and observational language (Carnap, 1950). Theoretical language may refer to such concepts as "temperature," "gravity," and "quarks," or to such terms as "personality," "intelligence," and "anxiety". The logical positivist movement reflects an effort to connect theoretical language to observational language via logical reconstruction. For example, anxiety is not directly observable, but instead anxiety is a theoretical concept used to account for empirical observations.

From a logical positivist perspective, anxiety may be viewed as a theoretical term standing in relation to observational language (e.g., John has a heart rate of 120 beats per minute) through correspondence rules (e.g. increases in heart rate reflect proportionate increases in anxiety). Construct validity consists of placing anxiety within a nomological network. This requires identifying law-like relationships between anxiety and other theoretical terms within an interlocking web or system. Theory may suggest law-like relationships between anxiety, depression, and self-esteem. A positive correlation between anxiety and depression, as well as a negative correlation between anxiety and self-esteem may provide therefore provide marginal support for such a system.

Cronbach and Meehl's (1955) view of construct validity reflects a strong departure from psychometric realism, given that one may infer the meaning of theoretical terms without requiring them to exist as entities within the world. Thus, "scientifically speaking, to 'make clear what something *is*' means to set forth the laws in which it occurs" (p. 290). Theoretical terms, such as anxiety, derive meaning from their placement within a nomological network. Altering one aspect of this network (e.g. anxiety is positively correlated with self-esteem) changes the meaning of theoretical terms within a system. With this movement away from realism, it is but a small step toward strictly locating validity as a property of score-based inferences and entailed uses of test scores. There is some ambiguity however, with respect to the proper location of validity as discussed in this article. Cronbach and Meehl readily identify construct validity as an interpretation to be defended. However, they were "not in the least advocating construct validity as preferable to other three kinds (concurrent, predictive, content)" (p. 300). Instead, construct validity is an additional consideration in testing. Though they indicate that "one does not validate a test, but only principle for making inferences" (p. 297), there are subtle distinctions between validity and validation. Validity may still be a property of tests, whereas validation refers to the process of investigating claims about a test. Nevertheless, their conceptualization of construct validity marks a pivotal turn in validity semantics.

It was not until the 1970's and 1980's that theorist forcefully emphasized interpretations and entailed uses of test scores as the proper location of validity. This view was perhaps most vehemently argued by Messick (1975, 1989), though similar positions can be found in various other sources (e.g. APA, AERA, NCME, 1974; Cronbach, 1971). Under this view, tests are neither valid nor invalid, but it is the proposed use and interpretation of test scores that encompasses validity. Stated differently, *inferences* from test scores are either valid or invalid. To understand the rationale for this movement, it is beneficial to consider the constructive-realist position as advocated by Messick (1998).

Messick (1998) was concerned with various philosophical criticisms, particularly the value-laden nature of empirical observations. Given that "theories can no longer be tested against facts" but are instead "relative to specific social practices of science" (Messick, 1998, p. 36), there are dangers in presuming that psychological concepts directly refer to a given reality. Messick however, did not abandon realist sympathies, and instead argued for an alternative account that combined constructivist criticism and psychometric realism:

> "In this constructive-realist view of psychological measurement, constructs represent our best, albeit imperfect and fallible, efforts to capture the essence of traits that have a reality independent of our attempt to characterize them. Just as on the realist side there may be traits operative in behavior for which no construct has yet been formulated, on the constructive side there are useful constructs having no counterpart in reality" (p. 35).

Messick draws a subtle distinction between constructs and the traits or attributes to which constructs may refer (Hood, 2009). Constructs may either refer or fail to refer to attributes that exist irrespective of human input. Consequently, Messick seems to adopt a form of realism wherein our theoretical terms may correctly describe unobservable attributes, though our epistemic accessibility to these attributes remains problematic. Messick (1989) however, sought to unify semantics under the label

"construct validity," which broadly concerns "an integration of any evidence that bears on the interpretation or meaning of the test scores" (p. 7). Thus, validity pertains to the degree to which evidence and theory support the adequacy and appropriateness of score-based inferences (Messick, 1989).

Many contemporary theorists have criticized this unified picture of validity. These criticisms stem from concerns about the ambiguity of the term "construct" and an inability of this approach to guide validation efforts (Borsboom, Cramer, Kievit, Scholten, & Franic, 2009; Kane, 2001). In other words, under this view it is difficult for test developers to articulate where validity evidence should begin and end. There also appears to be relative consensus that nomological networks, as originally articulated by Cronbach and Meehl (1955), have failed to be identified in education and the social sciences. Though it remains possible that undiscovered law-like relationships exist, such relationship have not, and perhaps never will be, identified in educational and psychological research. This led to a proliferation of what Cronbach (1988) refers to as weak validity programs that seek to establish a loose network of ill-defined interrelationships in validation research. Nevertheless, locating validity as a property of score-based interpretations and entailed uses has remained a consistent theme in contemporary validity semantics. This position is adopted within the most recent Standards for Educational and Psychological Testing (1999), which indicates that validity is "…the degree to which evidence and theory support an interpretation of test scores entailed by proposed uses" (APA, AEAR, & NCME, p. 9). This statement implies that validity is an open-ended evaluation that varies according to evidential support. As such, validity is not an all-or-nothing statement, but instead inferences are more or less valid according to existing evidence.

The argument-based approach to validity originated out of these historical developments (Kane, 2001). This approach aims to clarify validation efforts by requiring researchers to specify assumptions and interpretive inferences prior to seeking validity evidence. As previously mentioned however, Borsboom and colleagues (2004) have criticized locating validity as a property of score-based inferences. Under their view, validity is a binary function of truth residing within tests themselves. The following sections delineate each perspective by their ontological and epistemological characteristics, as well as their breadth of validation focus.

Validity Semantics under the Argument-Based Approach

The historical developments within validity theory, particularly the account of construct validity provided by Cronbach and Meehl (1955), along with the unified view of validity offered by Messick (1988), promoted specific principles that Kane (1992, 2006) argues are aligned with an argument-based approach. These developments support (a) validity as a property of interpretations and not tests, (b) validation consisting of an extended investigation, (c) consequences of testing as an aspect of this investigation, and (d) subjecting interpretations, assumptions, and proposed uses of scores to logical and empirical examination. Instead of placing construct validity as a unifying feature of validity theory, Kane (1992, 2013) provides a unified view of validity by locating interpretative and validity arguments at its' center. Succinctly put, individuals construct arguments for each score-based interpretation or entailed use of test scores. Validation consists of subjecting these inferences, and plausible alternatives, to both logical and empirical examination.

An interpretive argument "lays the network of inferences leading from the test scores to the conclusions to be drawn and any decisions to be based on these conclusions" (Kane, 2001, p. 329). Since validity resides with an interpretation, as opposed to a test, it is conceivable that multiple score-based interpretations and proposed uses exist for the same set of scores. If one developed 20 test items pertaining to statistical hypothesis testing, relatively simple evidence may be needed if these scores are primarily used as an indication of achievement within a specific course. For example, it may be possible to investigate the alignment of items to specific learning objectives, or estimate the generalizability of scores on observed items to a universe of possible items. Much stronger evidence would be required if one wished to interpret these scores as an indication of statistical aptitude, or if administration decided to use these scores for placement into academic programs. Consequently, the

content of an interpretative argument frames subsequent validation efforts since distinct score-based interpretations require different lines of evidential support (Kane, 2009).

Ontology. Locating validity as a property of interpretations, as portrayed in the argument-based approach, does not *necessitate* a realist or antirealist position toward the existence of educational and psychological attributes, traits, or skills. A psychometric realist, at least in the sense used within the present article, contends that testing procedures aim to measure entities in the actual world. As previously discussed, an anti-realist may deny this claim. For example, the collegiate learning assessment (CLA) is a performance-based measure aiming to assess critical thinking, analytical problem solving, and other higher-level cognitive skills (Klein, Benjamin, Shavelson, & Bolus, 2007). The realist would view "critical thinking" as an entity in the actual world and could conceivably argue that the CLA aims to assess variation in this entity. Denial of this position may occur in various ways. For example, the antirealist may argue that at best, "critical thinking" is a concept that is more or less useful for achieving valued aims and purposes. Alternatively, an antirealist may remain agnostic with respect to the existence of critical thinking as an entity that exists in the actual world. Put differently, an antirealist may view the existence of critical thinking as unimportant to the theory of critical thinking (Borsboom, 2005). The argument-based approach to validity remains uncommitted to either the realist or the antirealist viewpoint. This is reinforced by recognizing a distinction between score-based interpretations and entailed uses of test scores.

Including entailed uses of scores within the semantics of validity indicates that this concept incorporates the consequences resultant from testing procedures (Kane, 2012). The aims of assessment or testing may be indifferent to the ontological status of unobservable theoretical entities, attributes, or skills. For example, some researchers primarily use standardized tests to predict subsequent performance (Mattern, Kobrin, Patterson, Shaw, & Camara, 2009) or they may use test scores to select candidates for entrance into special programs. In this situation, it is the proposed use of the scores, not the ontological status of an unobservable attribute, which stands in need of validation. An a priori negation of the realist or antirealist positions does not therefore occur under the argument-based approach to validity. Put differently, the argument-based approach to validity remains absent of ontological commitments, at least prior to the collection of validation evidence.

Epistemology. Contrary to the ontological, which is generally focused on the "what" of existence, epistemological questions focus on the possibility, constituents, and limitations of human knowledge (Williams, 2001). As such, the justification of knowledge has taken a central place in epistemology as a philosophical discipline (Chisholm, 1989), and in an analogous way the process of validating score-based interpretations and entailed uses of test scores is of paramount importance within the argument-based approach. At least two epistemological implications are inferred from the argument-based approach: (a) appropriate evidence is a function of interpretative arguments and (b) validity is a tentative judgment employed with varying degrees of certainty. These features underscore the constructed, dynamic, and open-ended aspects of validating interpretative arguments (Kane, 1992) and result from locating validity within the realm of interpretation and the constraints imposed by specific interpretative arguments. Thus, each interpretative argument establishes boundaries or parameters that guide the validation process.

The argument-based approach is aligned with the view that "extraordinary claims require extraordinary evidence" (Sagan, 1980), whereas less ambitious claims require less extraordinary evidence. For example, "it is possible to answer questions about whether a person can perform a job, without having any deep understanding of how they perform the job" (Kane, 2009, p. 53). Assessment practitioners and researchers may therefore model patterns in observed data, in such a way, that useful predictions are facilitated without any understanding of the underlying mechanisms that account for observed score variation. This view is therefore aligned with the distinction between prediction and explanation (Pedhazur, 1997), yet once again, the content of the interpretative argument informs validation criteria. However, interpretative

arguments, at least broadly speaking, are examined for their coherence and plausibility given the available evidence (Kane, 2012). Claims to validity are not necessarily claims to truth, but instead reflect evaluative judgments about a given line of evidence.

Breadth of Validation Focus. Validation efforts under the argument-based approach are exclusively constrained by the *number* and *content* of score-based interpretations and proposed uses of scores. Once articulated, an interpretive argument for a set of scores can limit the evidence needed to support a claim. However, given that a potentially unlimited number of interpretations may be advanced for the same set of scores, or the possibility that the same scores may be used to make multiple decisions, the argument-based is considered broad in focus. For example, the CLA may be interpreted strictly as indicators of performance within a target domain or as a measure of critical thinking. Moreover, these scores could be used to make decisions about student placement or to establish benchmarks when making institutional comparisons. A consequence of placing validity as property of interpretations and uses is that each of these possibilities, once proposed, is included within the argument-based approach. This breadth of focus stands in sharp contrast to the instrument-based approach to validity.

Validity Semantics under the Instrument-Based Approach

Instead of locating validity as a property of score-based inferences, validity is located as a property of tests themselves under the instrument-based approach. This position is reminiscent of earlier theorists who indicated that tests themselves are either valid or invalid (Kelley, 1927). Borsboom's exposition of this view entails an ontological emphasis requiring commitment to psychometric realism if ever a validity claim is accepted (Hood, 2009). As discussed below, the ability to warrant causal inferences is the fundamental criteria for evaluating test validity. This focus on causal mechanisms restricts the breadth of validation focus to an examination of such processes. Hood (2009) has expounded upon both the ontological and epistemological characteristics of this view, thus the present section entails a reformulation and expansion of this work.

Before proceeding, a succinct summarization of Borsboom's validity semantics is in order. Borsboom, Van Heerden, and Mellenbergh (2003) provide the following definition of validity:

> Test X is valid for the measurement of attribute Y if an only if the proposition "Scores on test X measure attribute Y" is true (p. 323).

There are a few points to notice about this definition. First, validity is strictly located as a property of tests since it pertains to a single proposition indicating that scores measure an attribute. This definition would therefore naturally lead to questions about the semantics of measurement, and Borsboom (2005) relies upon the latent variable model to provide such a framework. The latent variable model contends that covariation between observed scores can be explained or accounted for by unobservable entities. The rationale for selecting the latent variable model becomes evident when considering the following requirements for validity claims:

> A test is valid for measuring an attribute if and only if (a) the attribute exists, (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure (Borsboom et al., 2004, p. 1061)

In contrast to the argument-based approach (Kane, 1992, 2001), wherein validity is a function of evidential support, Borsboom instead argues that validity is a function of truth itself (see Borsboom, et al., 2009). Consequently, the truth-value of validity claims is a function of two conditions. First, the attribute entailing the aim measurement must actually exist. Secondly, differences in this attribute are causally responsible for differences in observed scores.

Ontology. The conditions provided by Borsboom for establishing validity require the existence of psychological attributes. An individual who rejects psychometric realism must therefore reject all claims to validity under this approach (see Hood, 2009). For example, it is technically possible that an antirealist would accept Borsboom's position that a test is valid if

indeed it captures variation in real-attributes. However, to maintain an antirealist position, which may either deny the existence of these attributes or our ability to detect such attributes, the antirealist would have to reject that any test is indeed valid. To put this differently, the antirealist may accept the validity semantics yet discard either the ontological status of these attributes or the epistemic possibility of warranting such claims. The psychometric realist on the other hand can maintain consistency irrespective of whether a test is valid or invalid.

Locating validity as a property of tests themselves, at least as conceived by Borsboom, appears to necessitate psychometric realism if ever a test is deemed valid. Accepting measures of theoretical attributes, such as "critical thinking," "written communication," and "respect for diversity," as valid obliges one to accept the position that these are indeed entities within the actual world, as opposed to instrumental devices (see Duhem, 1954; Popper, 1963). Additionally, these attributes must exist in such a way that they act as efficient causes on observed scores. It should be clear however, that this idea of causation, though aligned with latent variable theory, does not necessarily coincide with causal efficacy within a single individual. For example, and as discussed by Borsboom, the five factor model of personality (McCrae & Costa, 1999), may account for variation between people, yet this does not imply that these factors reside within a single individual.

There are two important consequences of this perspective. First, mapping between-person and within-person variation is of central importance. There are dangers in presuming that between-person models align with within-person variation. If such models coincide, then no additional theoretical work may be necessary. However, if between-person models are different from within-person models then theories must account for such discrepancies. The second implication, which in some ways derives from the first, suggests that underlying processes accounting for between-person differences may be different than what occurs within a person. For example, differences in extraversion between people may account for patterns in item responses. However, extraversion may be constant within an individual. Since causality between X and Y requires X and Y to covary, a constant level of extraversion could not act as a causal force within a person (see Borsboom, Mellenbergh, & Van Heerden, 2003). Test validity would therefore pertain to processes occurring at both of these levels. This leads to a consideration of the epistemological characteristics of this position, which emphasizes an explication of process models wherein establishing causal inferences is essential.

Epistemology. Validity under this approach is inherently ontological and the truth of this ontological claim is independent of epistemological issues, or our ability to evaluate these claims (Borsboom, Mellenbergh, & Van Heerden, 2004). An instrument may be valid without any existing evidential support, or conversely an instrument may be invalid despite a line of seemingly strong supportive evidence. Borsboom and colleagues have argued that if the aim of measurement is to construct an instrument that is sensitive to variation in entities, then the truth of this ontological claim guarantees validity. Thus "validity is about ontology; validation is about epistemology" and the "two should not be confused" (p. 1063). Validity is therefore bound to the ontological status of theoretical entities and their causal connection to empirical observations. Without the existence of actual entities, validity also ceases to exist. Validation on the other hand, consists of the empirical explication of causal processes from an attribute to differences in observed scores. Establishing this causal connection appears to require the existence of strong theory and a priori knowledge about how variation in an attribute leads to changes in observed scores.

This position departs from the evidential criteria guiding most validation practices, which heavily relies on investigating correlation matrices. Consider the typical process of validation research. First, a researcher may create an instrument in collaboration with content experts who provide ratings about construct representation. Subsequent investigations typically examine the internal structure of the instrument (e.g. perhaps by conducting a factor analysis along with reliability estimates). The internal structure of the test may then be compared across different populations, with the same population across time, or across distinct environmental conditions. Finally,

correlations are sought with numerous variables to examine the external structure of the test. Generally speaking, if both the internal structure and external structure correspond with theoretical expectations, then researchers feel confident that they are measuring the intended attribute.

This line of evidential support is largely inadequate under the instrument-based approach. To see this, it is first necessary to be clear about measurement under Borsboom's approach. Various theories of measurement have been proposed (e.g. Stevens, 1946; Rasch, 1960) and much of this remains controversial (see Michell, 1999). However, Borsboom (2005) has argued that the practice of measurement requires the existence of entities that are responsible for producing variation in observed scores. Measurement is therefore both directional and causal (Borsboom et al., 2004). This naturally aligns with the latent variable model, the truth of which would essentially clinch the question of validity (Borsboom et al., 2009). However, how does one establish the truth of this model? Unfortunately, Borsboom has not provided many details about how this would work, though he has alluded to some possibilities.

An obvious choice may be to conduct a confirmatory factor analysis, which could examine how well a theoretical model fits patterns within observed data. However, adequate model fit does not itself demonstrate causality and the problem of underdetermination also poses challenges for claims to test validity. In other words, even with adequate model fit alternative models may equally account for the observed data. Such issues have led Borsboom to argue for process models specifying how variation in an attribute leads to responses on items or tasks. Instead of constructing an instrument and then trying to determine what it measures, instruments are constructed with theories detailing how differences in an attribute lead to specific responses. Borsboom illustrates how this may be done by referring to a balance scale task given to children (Borsboom et al., 2004). This task requires students to determine whether a balance scale will tip to one side, or remain equal, given the distribution of weights that are placed along the scale. Theories exist about specific strategies children use across developmental stages. Such theories allow one to construct tasks so that children relying on certain strategies will tend to fail a particular item. These developmental strategies can be translated into a model that investigates the extent to which latent class membership corresponds to patterns in observed responses (Jansen and Van der Maas, 1997). Validation under the instrument-based approach thus extends beyond methodological decisions detached from theoretical considerations. Theory is crucial for detailing processes leading to response behavior, which is the crux of validity under the instrument-based approach.

Breadth of validation focus. Locating validity as a property of tests themselves, and concurrently requiring one to establish causal relations between theoretical attributes and observed scores as a sole aim of validation, restricts the breadth of focus to these endeavors. Under the argument-based approach, the content of interpretative inferences informs evidential requirements in validation research. However, the validity semantics provided by Borsboom locates evidential criteria within the realm of causal relations. Thus, whether the SAT is useful for admission purposes would remain outside the purview of validity theory. This does not imply that such questions are unimportant or unworthy of investigation, but only that they are separate concerns from test validity. There are other implications of this position. For example, it is technically inappropriate to incorporate the consequences of testing as an aspect of validity theory. Tests may be valid in Borsboom's sense, yet still have undesirable consequences in various applied settings. Though consequences are important, the central question of validity would instead pertain to invariance of causal relations across different conditions and/or populations. In sum, the validity semantics under the instrument-based approach restricts validation efforts to a single proposition—scores on a particular test measure a specified attribute. Causal relationships constitute the heart of validity and consequently inform validation efforts.

Discussion

The concept of validity is fluid in educational and psychological testing. Scholars have diverged in validity semantics throughout the history of educational and

psychological testing. Broadly speaking, validity theory has witnessed a cyclical affinity with antirealist perspectives. However, scattered throughout this history are points of contention, division, and disagreement about the proper location of validity. Accepting that tests themselves are valid or invalid nicely coincides with psychometric realism, or the position that instruments indeed measure real attributes; whereas, locating validity with the realm of interpretation appears to more easily coalesce with antirealist affinities in that an antirealist may consistently accept some, but not all, test-score interpretations as valid. Nevertheless, strictly speaking, locating validity semantics as a property of inferences or instruments does not alone necessitate either of these views (Hood, 2009). What does seem important, at least for assessment practitioners, is that these divergent validity semantics amount to more than arbitrary affinities. Thus far, this discussion has largely been philosophical and theoretical in character. I therefore wish to conclude this discussion by briefly underscoring some of the pragmatic implications of this debate.

Before proceeding, I will make some final comments about the argument-based (Kane, 1992, 2013) and instrument-based (Borsboom et al., 2004) approaches to validity. The argument-based approach to validity is wide in scope, given that it incorporates multiple score-based interpretations and uses. The instrument-based approach restricts validity to one specific interpretation – scores on Test X are caused by variation in attribute Y. An individual who agrees with the argument-based approach may choose to adopt the score-based interpretation advocated by Borsboom. Both positions are therefore consistent, at least in this minimal respect. They do depart in important ways however. The argument-based approach views validity as an open-ended judgment that varies according to evidential support. Tests are not valid or invalid, but instead inferences derived from scores are more or less valid. The argument-based approach also incorporates the consequences of testing as an aspect of validity. Borsboom (2005) entirely rejects these aspects of the argument-based approach by articulating an account of validity informed by measurement theory. Various aspects of measurement theory remain controversial, so it is interesting to note that many theorists tend to neglect the semantics of measurement in their account of validity theory. Whereas Borsboom views validity as intricately connected to the meaning of measurement, the argument-based approach remains largely silent on such issues.

Clearly, assessment is concerned with more than formal "tests" per say, and the semantic differences described in this paper have implications well beyond formal testing. Educational and psychological assessment constitutes diverse aims and practices that definitely include testing, but it also incorporates alternative practices such as performance, portfolio, and classroom assessment. Though Borsboom discusses validity as a property of a test, it seems conceivable to broaden this view to other assessment practices. For example, scores are often given to student assignments, portfolios, oral presentations, and other samples of performance-based activities. The instrument-based approach to validity may therefore be applicable to any assessment activity aiming to measure variation in an attribute. The argument-based approach to validity seems applicable to most, if not all, assessment activities. Given the scope of assessment activities, does adherence to psychometric realism matter? How should we conceptualize validity within the context of assessment? These are challenging questions that lack a simple resolution.

Educational assessment is value-laden, in that assessment activities generally aim to investigate the achievement of desired student-learning outcomes. The importance of psychometric realism, as defined in this article, may depend upon the context in which assessment occurs. For the sake of simplicity, one may consider writing as a student-learning outcome. To assess this outcome student papers may be selected across the campus that are then scored by raters using a common rubric. These scores presumably reflect differences in an attribute, that of individual differences in writing. However, what evidence do we have that variation in the quality of writing, and not something else, actually influences observed scores? It seems, at least in cases such as this, that processes through which observed scores are manifest are important validity considerations. If for example, we find that assignment characteristics systematically elevate or deflate scores

then the integrity of data-driven decisions resultant from this procedure is questionable. A similar line of reasoning exists for institutional effectiveness, which minimally consists of an examination of value-added outcomes. Institutional effectiveness is an attribute that presumably acts as an efficient cause on learning outcomes. Mapping how variation in institutional experiences lead to differences in student learning outcomes is therefore crucial. This implies that the aim of many assessment processes may implicitly assume a form of realism.

Within other contexts however, the realist position remains irrelevant. Consider validation efforts of the SAT, a test that is widely used for admission purposes across many institutions of higher education. According to the argument-based approach, validity evidence is necessary for each proposed use and interpretation of these scores. If one uses the test *solely* to make predictive inferences about student success in college, then validity entails the degree to which evidence supports this entailed use of test scores. This specific use remains unconcerned with psychometric realism, or the position that observed scores reflect differences in an underlying trait or attribute. It is also worth noting that such inferences may be valid irrespective of test content. If self-reported frequency of coffee consumption were highly correlated with first-year GPA in undergraduate courses, then using these scores for admission purposes may very well be valid. Such a view resembles a pragmatic or instrumental orientation toward measurement (Stevens, 1946). In other words, whether scores are sensitive to variation in an underlying attribute is unimportant so long as these scores are capable of doing what we want them to do. Under Borsboom's approach, this line of evidence is irrelevant to validity, which strictly involves a causal relation between observed scores and an underlying attribute. Validation efforts would therefore exclusively consist of investigating underlying causal processes through which observed scores are manifest. This does not imply that evidence for entailed uses of test scores is unimportant, but such questions fall outside the scope of formal validity theory.

Finally, if Borsboom's criteria for test validity are accepted, then the validity of most existent tests in educational and psychological assessment remains suspect. This is not a point against Borsboom, but merely an indication of how accepting this position may radically alter validity claims within assessment practices. Perhaps the strength of Borsboom's approach, irrespective of whether one agrees with the validity semantics, is that his view underscores a need to investigate a relatively neglected area in educational and psychological assessment – namely, process models underlying observed score variation. Investigation of these processes may be invaluable for assessment practitioners. Furthermore, between-person models do not necessarily correspond to models within a person. This is an important consideration, given that distinct processes leading to observed scores may exist across these models. As assessment practices continue to evolve, between-person models may be insufficient for many of the tasks that lie ahead. However, the importance of this distinction largely depends upon the kinds of claims one wishes to make:

> "…that 30 per cent of interindividual variation in success in college may be predicted from the grade point average in high school, does not mean that 30 per cent of the exams you passed were predictable from your high school grades; and that there is a sex difference in verbal ability does not mean that your verbal ability will change if you undergo a sex change operation" (Borsboom, 2005, p. 77).

Both the argument-based approach and instrument-based approach to validity are consistent with an examination of such models. Thus the critical question, at least for assessment practitioners, is whether we wish to require a commitment to psychometric realism for accepting validity claims. Answering this question is well beyond the scope of the present article. Nevertheless, the field of educational and psychological assessment, to the extent that it wishes to address validity, may eventually need to contend with the implications of these divergent positions.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological tests*. Washington DC: American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1974). *Standards for educational and psychological tests*. Washington DC: American Psychological Association.

Amico, R.P. (1995). *Problem of the criterion*. Lanham, MD: Rowman and Littlefield Publishers.

Anastasi, A. (1938). Faculties versus factors: A reply to professor Thurstone. *Psychological Bulletin*, *35*, 391-395.

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

Borsboom, D., Cramer, A.O.J., Kievit, R.A., Scholten, A.Z., Franic, S. (2009). The end of construct validity. In R.W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 135-170). Charlotte, NC: Information Age Publishing.

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

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203-219. doi: 10.1037/0033-295X.110.2.203

Borsboom, D., Van Heerden, J., & Mellenbergh, G.J. (2003). Validity and truth. In Yanai, H., Okada, A., Shigemasu, K. Kano, Y., & Meulman, J.J. (Eds.), *New developments in psychometrics*: *Proceedings of the International Meeting of the Psychometric Society 2001* (pp. 321-328). Tokyo: Springer.

Carnap, R. (1950). *Testability and meaning*. New Haven, CT: Whitlock.

Chisolm, R.M. (1973). *The problem of the criterion*. Milwaukee, WI: Marquette University Press.

Chisolm, R.M. (1989). *Theory of knowledge*. Englewood Cliffs, NJ; Prentice-Hall.

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. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Cureton, E.E. (1951). Validity. In E.F. Lindquist (Ed.), *Educational measurement*. Washington DC: American Council on Education.

Duhem, P. (1954). *The aim and structure of physical theory*. Princeton: Princeton University Press.

Guilford, J.P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *6*, 427-439.

Gulliksen, H. (1950). Intrinsic validity. *The American Psychologist*, *5*, 511-517.

Hood, S.B. (2009). Validity in psychological testing and scientific realism. *Theory & Psychology*, *19*, 451-473. doi: 0.1177/0959354309336320

Jansen, B.R.J., & Van der Maas, H.L.J. (1997). Statistical tests of the rule assessment methodology by latent class analysis. *Developmental Review*, *17*, 321-357.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527-535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319-424.

Kane, M. (2006). In praise of pluralism. A comment on Borsboom. *Psychometrika*, *71*, 441-445. doi: 10.1007/s11336-006-1491-2.

Kane, M. (2009). Validating the interpretations and uses of test scores. In R.W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 39-64). Charlotte, NC: Information Age Publishing.

Kane, M. (2012). All validity is construct validity. Or is it? *Measurement*, *10*, 66-70. doi: 10.1080/15366367.2012.681977

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

Kelley, T.L. (1927). *Interpretation of educational measurements*. New York: World Book Company.

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: facts and fantasies. Evaluation Review, 31, 415-439. doi: 10.1177/0193841X07303318

Lissitz, R.W. (2009). *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing.

Lissitz, R.W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*, 437-448. doi: 10.3102/0013189X07311286

Mattern, K., Kobrin, J., Patterson, B., Shaw, E., & Camera, W. (2009). Validity is in the eye of the beholder: Conveying SAT research findings to the general public. In R.W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 213-40). Charlotte, NC: Information Age Publishing.

McCrae, R.R., & Costa, Jr., P.T. (1999). A five-factor theory of personality. In L.A. Pervin & O.P. John (Eds.), *Handbook of personality: Theory and research* (2nd Ed. pp. 139-153). New York: Guilford.

Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955-966.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed.). Washington, DC: American Council on Education.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*, 35-44.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge: Cambridge University Press.

Moss, P.A., Girard, B.J., & Haniford, L.C. (2006). Validity in educational assessment. *Review of Research in Education*, *30*, 109-162.

Mulaik, S.A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research*, *22*, 267-305.

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

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*, 25-45.

Pedhazur, E.J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). USA: Thomas Learning Inc.

Poli, R. (2010). Ontology: The categorical stance. In R. Poli & R. Seibt (Eds.), *Theory and applications of ontology: Philosophical perspectives* (pp. 1-22). New York: Springer.

Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London: Hutchinson.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.

Rodgers, J.L., & Nicewander, A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*, 59-66.

Rulon, P.J. (1946). On the validity of educational tests. *Harvard Educational Review*, *19*, 405-450.

Sagan, C. (1980). *Cosmos: A personal voyage* [Television Series]. Arlington, VA: Public Broadcasting Service.

Sireci, S.G. (1998). The construct of content validity. *Social Indicators Research*, *45*, 83-117.

Sireci, S.G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R.W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19-37). Charlotte, NC: Information Age Publishing.

Spearman, C. (1904). General intelligence: objectively determined and measured. *American Journal of Psychology*, *15*, 201-293.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, *103*, 677-680.

Taylor, J. R. (1997). *An introduction to error analysis: The study of uncertainties in physical measurements*. Sausalito, CA: University Science Books.

Traub, R.E. (2005). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, *16*, 8-14.

Williams, M.J. (2001). *Problems of knowledge: A critical introduction to epistemology*. USA: Oxford University Press.

Citation:

Hathcoat, John D. (2013). Validity Semantics in Educational and Psychological Assessment. *Practical Assessment, Research & Evaluation*, 18(9). Available online: http://pareonline.net/getvn.asp?v=18&n=9.

Author:

John D. Hathcoat is an Assistant Professor in Graduate Psychology and Assistant Assessment Specialist in the Center for Research and Assessment Studies at James Madison University. His scholarly interests include educational assessment, validity theory, and the analysis of error variance within assessment processes.

Email: johndhathcoat [at] gmail.com