Studies in Applied Linguistics & TESOL at Teachers College, Columbia University, Vol. 23, No. 1, pp. 18–36
Forum

# On Validity

**Pieter Lauwaert**
*Teachers College, Columbia University*

## ABSTRACT

The way in which validity has been conceptualized has changed throughout the years. The focus in validation studies shifted from evaluating distinct components of validity to developing a comprehensive argument for the use and interpretations of test scores. The argument-based approach to validity incorporates the distinct types of the componential approach, underscores Messick's attention to construct validity, highlights the need for evidence of the evidence-gathering approach, and combines it all into one logical argumentative structure. Such a comprehensive argument is now an indispensable part of validity studies (Dursun & Li, 2021).

## INTRODUCTION

Validity has always been one of "the most fundamental and important" concepts in language testing (Angoff, 1988, p. 19). But the way in which validity has been conceptualized has changed throughout the years. Consequently, different approaches have been used to pursue validity investigations. In this paper, I will demonstrate that at first, a componential approach to validity was used that viewed validity as consisting of distinct types: criterion (Cureton, 1950), content (Rulon, 1946), and construct validity (Cronbach & Meehl, 1955). Messick (1989b) unified these several types and demonstrated that they were all part of construct validity. Messick (1989a) advocated for an evidence-gathering approach to validity which was operationalized by Bachman and Palmer (1996) and Weir (2005). Because of the limited practicality of these checklist approaches, Kane (1992, 2006) proposed to use an argument-based approach to validity which later influenced the development of the assessment use argument (Bachman, 2005; Bachman & Palmer, 2010). Kane's Interpretive/Use argument was later expanded by Chapelle (2021) to make it more comprehensive. As such, the different approaches to validity have evolved greatly over the years.

## COMPONENTIAL APPROACH

### Criterion Validity

The history of test validation starts at the beginning of the 20th century after Karl Pearson invented the correlation coefficient in 1896 (Markus & Borsboom, 2013). In the next fifty years, the concept of validity referred to the extent to which a test measured what it is supposed to be measuring (Buckingham, 1921; Guilford, 1946; Garrett, 1947; Cureton, 1950).

Or as Rulon (1946) put it: "whether the test does the work it is employed to do" (p. 290). "If it does," Lado (1961, p. 321) said, "it is valid."

To operationalize this approach, Bingham (1937) proposed to compute the correlation of the test scores with "some other objective measure of that which the test is used to measure" (p. 214). In a similar vein, Guilford (1946) stated that "a test is valid for anything with which it correlates" (p. 429). Cureton (1950) coined the term 'criterion scores' to refer to this "standard against which the usefulness of the test scores is judged" (p. 623). The correlation between the test score and some criteria indicated the validity of the test. As Lado (1961) put it: "correlating the scores on a test with those of another test or criterion" (p. 30) demonstrated the validity of a test. This validity approach consisted of two types: predictive and concurrent validity (APA et al., 1954). The difference is that in the case of concurrent validity, the criterion is available at the same time as the test, whereas the criterion becomes available in the future in the case of predictive validity. Anastasi (1982) disagreed with this distinction by claiming that the difference between the two is "based, not on time, but on the objectives of testing" (p. 138). Concurrent validity would be used for tests that want to diagnose existing states of affairs whereas predictive validity should be reserved for tests that want to predict outcomes in the future. In 1966, these two types of validity, concurrent and predictive, were merged into 'criterion-validity' (AERA et al., 1966).

This criterion approach to validity had several limitations. Firstly, it might be difficult to find an adequate criterion that can be used (APA et al., 1954; Kane, 2006). Secondly, the external criterion itself might be invalid (Davies & Elder, 2005). This approach might also lead to an infinite regress of criteria that relate to each other (Bachman, 1990; Kane, 2006). Nor does it look at other sources of evidence and is, as such, limited (Messick, 1989b). Finally, if a weak correlation is found, it is not clear whether the test or the criterion is to blame (Chalhoub-Deville & O'Sullivan, 2020).

## Content Validity

Rulon (1946) argued that certain tests were "obviously valid" (p. 291) and did not need an external criterion. Instead, these tests, such as proficiency and achievement tests, only asked for a review by subject-matter experts to demonstrate that their content represented an appropriate sample of the domain. According to Cureton (1950), this "content validity" could be demonstrated by "a tabulation showing that the test content actually parallels and covers the course content" (p. 669). The Standards for Educational and Psychological Testing defined it in their first edition in 1954 as "showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn" (APA et al., 1954, p. 213). Likewise, Lado (1961) claimed that multiple-choice items needed to contain a language problem that was representative of a real-life problem.

This approach to validity had several limitations. It proved to be difficult to demonstrate a convincing correspondence between the test tasks and real-world tasks (Davies, 1984). Especially in the case of abstract constructs, such as aptitude and personality tests, because, as Angoff (1988) argued, "the domain cannot be adequately described" (p. 23). Kane (2006) claimed that the evidence for content validity tended "to be subjective and to have a confirmatory bias" (p. 19). The major limitation of content validity is that it does not account for how individuals actually perform on the assessment (Cronbach, 1971). Because content validity does not address these test scores or responses it, according to Messick (1989b), "does not qualify as validity at all" (p. 17). Likewise, Fulcher (1999) concludes that "using content validity as a major criterion in test design and evaluation has been mistaken" (p. 221).

## Construct Validity

In 1954, P.E. Meehl and R.C. Challman, as part of the subcommittee of the APA Committee on Psychological Test, first formulated the term construct validity (Chalboub-Deville & O'Sullivan, 2020). They defined the term as "demonstrating that certain explanatory constructs account to some degree for performance on the test" (APA et al., 1954, p. 214). The term 'construct' was defined by Cronbach and Meehl (1955) as "some postulated attribute of people, assumed to be reflected in test performance" (p. 283). It is the underlying trait that is presumed to be measured by the test. Cronbach and Meehl (1955), stated that this construct validity was to be investigated "whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured" (p. 282). So, whenever the criterion or content approach to validity was "insufficient to indicate the degree to which a test measures what it is intended to measure," test developers should try to assess the construct validity of a test (AERA & NCMUE, 1955, p. 16).

Cronbach and Meehl (1955) maintained that the construct needed to be part of a nomological network: "the interlocking system of laws which constitute a theory" (p. 290). Cronbach and Meehl (1955) claimed that construct validation is the development of such a network that "makes contact with observations and exhibits explicit, public steps of inference" (p. 291). As such, construct validity is the extent to which test performance is consistent with predictions that were made based on a theory. The construct defines which kind of data needs to be collected and these data are used to validate or revise the theory (Angoff, 1988). Strong construct validation, according to Cronbach (1988), wants to find "alternative explanations of the accumulated findings" (p. 14), whereas, in weak construct validation, any available evidence suffices to demonstrate validity.

To operationalize this approach, Campbell and Fiske (1959) developed an empirical test for construct validity. Their Multitrait-Multimethod matrix consists of intercorrelations that need to demonstrate convergent and discriminant evidence. The former indicates that two measures of the same construct are related, whereas the latter refers to two measures of two different constructs that should not be related. Or as Campbell and Fiske (1959) put it: "measures of the same trait should correlate higher with each other than they do with measures of different traits involving separate methods" (p. 104). According to Messick (1975), discriminant evidence is a stronger indicator of construct validity than convergent evidence because this indicates a refute of "plausible rival hypotheses" (p. 956).

This construct approach to validation had several shortcomings. The first is the unrealistic ambitions of this approach (Moss, 1992). Cronbach (1989) called it "pretentious" (p. 159) to try to develop strong construct validation studies in social and behavioral sciences because "this reaches centuries into the future" (p. 163). But even weak construct validation did not offer any guidance on the amount and type of evidence that needs to be collected (Norris, 2006). As a consequence, according to Kane (2001a), researchers were "highly opportunistic in the choice of validity evidence" (p. 323). Also, most educational practitioners were not skilled enough to perform this type of theoretical research (Norris, 2006).

## Face Validity

The 1940s saw the development of another type of validity, face validity (Angoff, 1988). Mosier (1947) defined face validity as "to appear practical, pertinent, and related to the purpose of the test" (p. 192). As such, Anastasi (1982) claimed it referred to the question "whether the test looks valid to the examinees who take it" (p. 136). But this approach to

validity was short-lived because it had several limitations. First of all, superficial judgments by laymen risk being wrong (Angoff, 1988). Cronbach (1984) pointed out that "good-looking" tests might have poor validity (p. 182). Face validity also implied that any other empirical verification was no longer necessary to demonstrate validity because the test seemed valid to the developer and the test takers (Mosier, 1947). Because of its ambiguous nature, Mosier (1947) urged that face validity "be banished to outer darkness" (p. 191). For those reasons, the American Psychological Association described face validity in 1974 as "not an acceptable basis for interpretive inferences from test scores" (APA, 1974, p. 26) and no longer used the term in the more recent editions.

Still, Anastasi (1982) maintained that face validity is "a desirable feature of tests" because it can increase the cooperation and engagement of the test taker (p. 136). Likewise, Davies (1977) stated that face validity "is not necessarily trivial from a practical point of view" (p. 60). Therefore, Angoff (1988) concluded that face validity can help to make a test more acceptable, but it is in itself not "a serious psychometric effort" (p. 24). In a similar vein, Cattell (1964) pointed out that face validity "perhaps still has a role, but in diplomacy rather than psychology" (p. 8).

### Trinitarian Doctrine

Thus, from the 1950s through the 1970s validity was being conceived as consisting of several 'types' (Angoff, 1988). As the AEPA (1955) framed it: "These four aspects of validity may be named content validity, predictive validity, concurrent validity, and construct validity" (p. 213). In 1966, concurrent and predictive validity were combined into one type, named "criterion-related validity" (AERA, 1966). These three types of validity: content, criterion, and construct, were being viewed, in the words of Guion (1980), as "something of a holy trinity representing three different roads to psychometric salvation" (p. 386).

This "trinitarian doctrine" had several shortcomings (Guion, 1980). Norris (2006) noted that this approach presumed that, once the validity of a test was established, the "validity inhered within the test regardless of how or with whom or why it was used" (p. 38). As such, there was no attention to the different uses or users of a test. This approach gave rise, according to Norris (2006), to a "mechanical and limited" view of validity in which test developers only used correlations or expert judgments to demonstrate the validity of the test (p. 38). Messick (1989b) pointed out that, because of the lack of clear choosing criteria, this approach to validity made "it possible to select evidence opportunistically and to ignore negative findings" (p. 33).

## UNIFIED APPROACH

Loevinger (1957) was the first to point out that these different types of validity should be unified under 'construct validity': "since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view" (p. 636). Loevinger (1957) regarded content and criterion validity as "possible supporting evidence" for construct validity (p. 653). Therefore, Loevinger concluded that construct validity is "the only kind of validity" (p. 641).

Messick (1975) advocated for the same unified framework when he stated that "all measurements should be construct-referenced" (p. 957). Because of their dependence on particular criterion variables and particular contents, Messick (1975) stated, "reliance upon criterion validity or content coverage is not enough" (p. 956). Instead, construct validity

should be viewed as "the unifying concept of validity that integrates criterion and content considerations" (Messick, 1980, p. 1015). The goal of construct validation, according to Messick (1980), is to determine "the meaningfulness or interpretability of the test scores" (p. 1015). As a consequence, construct validity subsumes content and criterion validity because information about content relevance and criterion-relatedness also addresses the score interpretation, which is the essence of construct validity (Messick, 1995). Thus, as pointed out by Markus and Borsboom (2013), in Messick's framework, the previous types of validity now became "distinct types of validity evidence for a single validity judgment" (p. 12). As such, Messick's theory is, according to Markus and Borsboom (2013), a "unified, but not unitary" theory because the several types of validity are subsumed under the umbrella of construct validity.

Messick (1980) proposed to use alternative descriptors for content validity, "content relevance and content coverage," and for criterion validity, "predictive utility and diagnostic utility" (p. 1015). Construct validity, according to Messick (1980), consists of twelve different types of validity: convergent, discriminant, trait, nomological, factorial, substantive, structural, external, population, ecological, temporal, and task validity. Messick (1995) defined two major threats to this construct validity: construct underrepresentation and construct-irrelevant variance. The former refers to assessments that fail to include all facets of the construct, whereas the latter is defined as excessive variance that is due to other aspects than the construct.

Messick criticized the previous approaches to validity because they failed to take into account value implications and social consequences of score interpretation and use (Messick, 1995). According to Messick (1992), test validity and social values are intertwined and thus, the "evaluation of intended and unintended consequences of any testing is integral to validation of test interpretation and use" (p.2). Validity and values are, according to Messick (1995), "one imperative" (p.749). Messick (1986) stated that this puts "not only a heavy ethical burden but a heavy interpretive burden" on the test users to include value consequences into their interpretation (p.18). By doing so, according to Davies and Elder (2005), Messick made "the link between testing arguments and the wider social and ethical turns" (p. 798). It underscored, according to Kunnan (1998), that the previous psychometric approaches were "not value-neutral" in themselves (p. 254). Thus, Messick's approach highlighted the importance of surveying different stakeholders because each can give a different interpretation of the meaning of the scores (Im et al., 2019).

# EVIDENCE-GATHERING APPROACH

## Messick's Matrix

Validity, according to Messick (1989b), is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). Thus, researchers need to collect evidence in order to evaluate the use and interpretation of a test. Messick (1989a) defined validity as the "summary of both the existing evidence for and the actual as well as potential consequences of score interpretation and use" (p. 5). As such, a test is not being validated, but rather, the inferences that are derived from the score. Thus, inferences need evidence to support the interpretation and to discount rival interpretations (Messick, 1992). According to Messick (1986), the most important issues of validity are "the interpretability, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of

social consequences" (p. 2). Validity is a matter of degree, and because of the changing social conditions, it is a continual process (Messick, 1989). In sum, Messick's evidence-gathering approach (1995) defined validity as "an evaluative summary of both the evidence for and the actual–as well as potential–consequences of score interpretation and use" (p. 472).

Messick's validity framework consisted of two facets of validity – test interpretation and test use–together with two types of evidence–the evidential basis and the consequential basis. As such, the justification for testing was linked to the function of testing (Kane, 2006). As can be seen in Table 1, this resulted in a matrix with four boxes that all consisted of construct validity and that highlighted, according to Messick (1995), the importance of "both meaning and values in both test interpretation and test use" (p.747). This progressive matrix started with construct validity in its first cell and relevance, value interpretations and social consequences were added to the following cells. Bachman (1990) listed different types of empirical evidence that could be collected for each of these cells. Kunnan (1998) found that most of the empirical studies, using Messick's framework, focused on the evidential basis for test interpretation whereas the consequential basis of test interpretation and test use was less often being researched.

**TABLE 1**
**Facets of validity as a progressive matrix**

| | Test Interpretation | Test Use |
|---|---|---|
| Evidential Basis | Construct Validity | Construct Validity + Relevance/Utility |
| Consequential Basis | Construct Validity + Value Implications | Construct Validity + Relevance/Utility + Value Implications + Social Consequences |

Although Messick (1989) highlighted the importance of test use and consequences, he did not offer any guidance on how to investigate these (Kane, 2006). According to Chalhoub-Deville and O'Sullivan (2020) it wasn't clear which groups should provide the evidence Messick was looking for. As Davies and Elder (2005) put it, the framework "need to be simplified or at least rendered more transparent to test users" (p. 810). Thus, Messick failed to prioritize the validity questions, and consequently, researchers didn't know what kind of evidence was enough (Chappelle and Voss, 2013; Shepard, 1993). Nor was it clear in Messick's framework how this validity research should be implemented and organized (Norris, 2006). Furthermore, Shepard (1993) criticized Messick's faceted framework because it gave the impression that "values are distinct from a scientific evaluation of test score meaning" (p. 426). Because construct validity is located in all the cells, Shepard (1993) critically stated that it is "not clear whether the terms name the whole or the part" (p. 427). Furthermore, it is unclear how to fully integrate the social dimension of a test into this framework (Roever & McNamara, 2006). Likewise, Wiley (1991) maintained that including social values into the framework risked overburdening the concept of validity: "the preoccupation with test use in recent discussions of validity has diverted attention from the fundamental roles of skills and task characteristics" (p. 105).

## The Test Usefulness Approach

The evidence-gathering approach of Messick influenced the test-usefulness approach of Bachman and Palmer (1996). Bachman and Palmer (1996) developed a metric to evaluate "all aspects of test development and use" (p.17). Their model consists of six test qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality. Impact, according to Bachman and Palmer (1996), refers to the effect tests have "on society and educational systems and upon the individuals within those systems" (p. 29). Because tests are developed and used in a specific context with certain values and goals, Bachman and Palmer (1996) claimed that we have to look at the consequences the test has "for both the individuals and the system involved" (p. 30). Bachman and Palmer (1996) posited that all of the qualities need to be addressed because prioritizing some to the exclusion of others is an "indefensible view" (p. 134). Thus, test developers should evaluate the combined effect of these test qualities on the usefulness of the test and try to maximize it. To do so, Bachman and Palmer (1996) developed a list of 42 questions that can be used "as a first step in setting minimum levels for the different qualities of usefulness" (p. 138). To determine the impact that a test has on individuals and society, they listed fifteen questions that range from the experience of taking the test, the feedback test takers get, and the decisions that are made, to the impact the test has on the teachers and on society.

This approach to validity was used by Spence-Brown (2001) to investigate the authenticity and interactiveness of a university-level Japanese language course in Australia. Spencer-Brown (2001) found that these two qualities have "implications for the validity of a task for learning, as well as its validity for assessment" and, therefore, should be analyzed more by researchers (p. 479). Chapelle et al. (2003) looked at all six qualities of the framework to develop a validity argument for a low-stakes web-based ESL course. Chapelle et al. (2003) claimed that this approach enabled them "to see how test validation theory informed practice at the initial stages of development" (p. 432).

This framework provides researchers with a systematic checklist that helps them decide what kind of evidence to look for in the validation process (Im et al., 2019). But, addressing all these questions may, according to Fulcher (2015), "mask any real problems with alternative explanations" (p. 119). Also, as Xi (2008) remarked, this approach did not prioritize the six qualities. As a consequence, according to Kane (2013), test developers might be tempted to focus on questions that are easily answered but ignore others that are more important, the so-called 'gilding-the-lily fallacy' (p. 19). Also, answering all these questions, according to Lewkowicz (2000), "may not always be possible or practical" because of time constraints (p. 50). Bachman (2005) later called this checklist approach no "coherent theory of test use" (p. 4). Nor did it indicate how construct validity and test use are directly related (Bachman, 2005). For those reasons, Bachman and Palmer (2010) adopted the argument-based approach in their later works (Fulcher, 2015).

## Socio-Cognitive Model

O'Sullivan and Weir (2011) refuted the test usefulness approach of Bachman (1990) because it was "impossible to operationalize […] due to its lack of clear prioritization" (p. 14). They also claimed that Bachman's approach lacked the cognitive processing dimensions of the various skills. Instead, according to O'Sullivan and Weir (2011), the socio-cognitive approach should be used because it is "the first systematic attempt to incorporate the social, cognitive and evaluative (scoring) dimensions of language use into test development and validation" (p. 20). This approach, developed by Weir (2005), specified five types of validity: context, theory-based, scoring, consequential, and criterion-related validity. The characteristics and cognitive processing skills of the test-takers should be an essential part of

the validity evidence because they "directly impact on the way the individuals process the test task" (p. 51). It views language use, as Lynda Taylor put it, "as both a socially situated and a cognitively processed phenomenon" (Shaw & Weir, 2007, p. 242). Therefore, O'Sullivan and Weir (2011) claim that this socio-cognitive approach goes "beyond the linguistic description of the language in use" and instead also incorporates the social context and psycholinguistic processing of language performance (p. 27). It puts "the individual [test-taker] at the centre" and forces test developers to look at how the test takers interact with the test items. In sum, this socio-cognitive approach tried to incorporate socio and cognitive elements into Messick's framework by listing what kind of validity evidence needed to be gathered for validation.

This approach has been used by Shaw and Weir (2007) to find validity evidence for the writing assessment of the Cambridge ESOL examinations. They found previous evidence-based approaches unhelpful because "the nature of the evidence to support claims and reasoning is not always clearly, explicitly, or comprehensively specified" (p. 243). In contrast, the socio-cognitive approach taught them what kind of evidence to look for, and it highlighted "the symbiotic relationship" between the tasks and the cognitive processes needed to successfully perform the tasks (p. 241). Khalifa and Weir (2009) adopted this approach to examine the reading tasks of the Cambridge examination suite. They found the socio-cognitive framework to be very useful because it attempts "to describe the constituent parts more fully and explicitly and to reconfigure validity to examine how these parts interact with each other" (p. 217). O'Sullivan (2009) used the socio-cognitive approach to link the CEFR-levels to other assessments. O'Sullivan (2005) adopted the approach to build test specifications and found it to be an efficient and practical way to ensure systematic changes across the different levels.

However, Fulcher (2015) critiqued this approach as a step back in time because it did "not appear to add much to what had gone before" (p. 117). Context validity is, according to Fulcher, nothing more than content validity plus a testing environment. Likewise, Weir's theory-based validity is similar to construct validity plus content knowledge. In a similar critique, Fulcher (2015) stated that Weir's consequential validity and criterion-related validity "map directly onto previous categories" of Messick. Fulcher (2015) also feared that this approach would not be easy to operationalize because of its complex and time-consuming nature. Furthermore, Fulcher (2015) also feared that this approach only focused on the perspectives of test developers instead of test takers. Therefore, Fulcher (2015) concluded that this "checklist research usually generates 'marginally relevant' information from a 'do-it-yourself kit of disjoint facts'" (p. 119).

# ARGUMENT-BASED APPROACH

## Interpretation/Use Argument

Validation researchers, according to Cronbach (1988), should develop a 'validity argument' that "must link concepts, evidence, social and personal consequences, and values" (p. 4). This validity argument is a coherent analysis of all the evidence for and against a proposed score interpretation and should include possible alternative interpretations. This resonated with Messick's (1989) idea of validity as a judgment of "the adequacy and appropriateness of inferences and actions based on test scores" (p. 13).

Kane (2006) used this concept of a validity argument and made it the organizing mechanism of his approach to validation research. Kane's (1992) argument-based approach to validation consists of an interpretive argument "with the test scores as a premise and the statements and decisions involved in the interpretation as conclusion" (p.527). The goal of the

interpretive argument is to make the proposed interpretation and use of the test scores explicit and to provide a framework for the validity argument (Kane, 2001b). In order to highlight the importance of score uses in determining score interpretation and to underscore the importance of score use in validation, Kane (2013) renamed the interpretive argument into "interpretation/use argument" (IUA). The argument-based approach also consists of a validity argument that evaluates the plausibility of the interpretive argument (Kane, 2001b). The clarity, completeness, and coherence of the interpretive argument need to be evaluated, and the plausibility of the inferences and assumptions need to be assessed (Kane, 2010). One validates a test-score interpretation by finding supporting evidence for the inferences and assumptions of the interpretive argument. Kane (2001a) stated that test developers should first state the interpretive argument and then develop a preliminary version of the validity argument by "assembling all available evidence relevant to the inferences and assumptions in the interpretive argument" (p. 330). The evidence, then, needs to be evaluated, and this might lead to a revision of the interpretive argument. These steps need to be repeated until all the inferences in the interpretive argument are plausible. As such, there are two main stages: a development stage in which the interpretive argument is being developed and an appraisal stage in which the argument is being critically challenged (Kane, 2013). These two steps are, according to Kane (2013), separated from each other "to encourage a careful and relatively complete specification of the proposed interpretation and use" (p. 19). Interpretive arguments are dynamic artifacts that can be adjusted based on their plausibility (Kane, 1992). The weakest parts of the interpretive argument should be addressed, and the most questionable assumptions should be analyzed (Kane, 1992). More ambitious claims need more support than less ambitious ones and are, thus, harder but more interesting to validate. Because interpretations change over time, the validity of the score interpretation is a matter of degree that can evolve over time. Kane (2001a) defined validity as "the clarification and justification of the intended interpretations and uses of observed scores" (p.339). As such, it is the proposed score interpretation and use that is being validated and not the test (Kane, 2013).

According to Toulmin (1958), making a claim carries with it the obligation to support the claim with evidence and trying to refute counterevidence. Using this idea of Toulmin (1958), Kane's model consists of inferences that start from a datum and make a claim. These inferences rely on a warrant, a general rule to infer claims from data, that require support. This support can either support the warrant, as a backing, or undermine it, a rebuttal. The interpretive argument includes a group of inferences of which the data are the claims of the previous inference. Kane (2013) described these inferences as the spans of a bridge indicating that "if one span falls, the bridge is out, even if the other spans are strongly supported" (p. 13).

**FIGURE 1**
**Kane's argument based approach to validation**

| Scoring inference | Generalization | Extrapolation | Decisions | |
|---|---|---|---|---|
| Observed performance | Observed score | Universe score | Target score | Decisions |

As can be seen in Figure 1, the interpretive argument consists of several inferences: scoring, generalization, extrapolation, and decisions (Kane, 1992; Kane, 2001b). The scoring inference links the observed performance to an observed score (Kane, 2013) and makes assumptions about the appropriateness of the scoring criteria. It assumes that the scoring rule was applied accurately and consistently. This inference is supported by evidence that indicates that the procedures were followed correctly (Kane, 1992). The generalization inference takes us from the observed score (test score) to the universe score and assumes that the sample was

representative. This inference can be supported by reliability or generalizability studies. The extrapolation inference links the universe score to the target score, which is the predicted performance in real-life. This inference assumes that there are relationships between the test performance and performances in a larger domain (Kane, 2013). Evidence for this inference can be found from criterion-related studies (Kane, 1992). The decision inference links the target score to the decisions and assumes that the consequences of the decisions are beneficial. The decision rule is the warrant for the decisions and, according to Kane (2013), it is "the capstone of the IUA for score-based decisions" (p. 46). Kane (2013) called a decision rule successful if it achieves its goal "at an acceptable cost and with acceptable consequences" (p. 47). The intended effects, adverse impacts, and unintended systemic effects determine how successful a decision rule is (Kane, 2013). Looking at the consequences is, according to Kane (2013), "a necessary component" of the validity process (p. 61). Still, Kane (2013) posited that the consequences should be limited to the group of test takers or to a subset of the population. Also, negative consequences in themselves do not count against the validity of the interpretation unless it indicates a defect in the argument.

If these interpretations do not involve any decisions, they are called "descriptive," whereas "decision-based" interpretations do include decision procedures (Kane, 2001b). As such, the argument consists of a "descriptive part," the first four inferences, and a "prescriptive part," the final inference (Kane, 2001a). The latter should be conducted by the test users and the test developers. The first four inferences are "semantic" because they make claims about what test scores mean whereas the last one is a "policy inference" because it refers to decision making (Kane, 2001b). The latter can be called successful if they achieve positive consequences at modest costs and with few negative side effects (Kane, 2001b).

As such, this argument-based approach can be applied to different types of test interpretations or use (Kane, 1992). By focusing on the weakest inferences, Kane (1992) posited that his approach might "lead to improvements in measurement procedures" (p. 534). It highlights the need for different kinds of analysis and evidence (Kane, 2010). Kane (2001a) argued that his framework provided a set of procedures that gave researchers "detailed guidance" while conducting validation studies (p. 340).

According to Markus and Borsboom (2013), the greatest advantage of Kane's approach is that it gives test developers a transparent framework that helps them answer the questions how to collect validity evidence. It forces test developers to look at potential rival explanations and to find support for their rebuttals (Chapelle & Voss, 2013). According to Markus and Borsboom (2013), this approach "provides more concrete guidance" to test developers on how to validate the uses of their test (p. 14).

According to Kane (2006), this argument-based approach to validity "reflects the general principles inherent in construct validity without an emphasis on formal theories" (p. 23). However, Im et al. (2019) argued that this missing aspect of construct validity is problematic because the IUA "do not have a phase to define constructs, which is critical to test design, before a scoring inference" (p. 11). Im et al. (2019) also claimed that the framework misses a methodological guideline to investigate decision-making inference. Fulcher (2015) criticized Kane for limiting the scope of the consequences, which "places some limited responsibility for test use upon test providers" (p. 112). O'Sullivan and Weir (2011) claimed that Kane's framework lacked attention to the social and cognitive aspects of language and that test developers might "focus on types of evidence that show their test in a good light" (p. 20). Likewise, Markus and Borsboom (2013) faulted Kane for not being concerned with the truth by only looking at justified realities.

## The Assessment Use Argument

According to Bachman (2005), test use is "at the heart of language assessment" (p. 2). Because previous approaches were not grounded in a coherent theory of test use nor did they provide procedures to investigate test use and consequence, Bachman (2005) developed an assessment use argument (AUA) which is a logical framework for linking assessment performance to test use. It consists of two parts: the utilization argument that links an interpretation to a decision and the validity argument that links assessment performance to an interpretation. Each argument consists of claims, warrants, and backings that support the intended uses and rebuttals of the unintended consequences of test use. The utilization argument has four warrants that are, according to Bachman (2005), "essential to the argument": relevance, utility, intended consequence, and sufficiency (p. 19). The interpretations of the scores need to be relevant to and useful for making the decision. Also, the intended consequences of using the test need to be beneficial, and the test should provide sufficient information for making that decision. The utilization argument also includes rebuttals articulating why another decision is made and what the potential unintended consequences are of using this test. Test developers should first develop the assessment utilization argument by articulating the intended uses of the test and then articulate the validity argument. Validation is, thus, in the words of Bachman (2005), "the process of collecting evidence or backing, to support the entire assessment use argument" (p. 16). Test developers need to defend the decisions that are made based on the test by giving credible evidence that supports their argument. Thus, the main reason for articulating an assessment use argument, according to Bachman (2005) is "so that we can be accountable to stakeholders" (p. 31).

Bachman (2005) claimed that the concerns for critical language testing (Shohamy, 2001) and the qualities of usefulness (Bachman & Palmer, 1996) and fairness (Kunnan, 2004) can all be incorporated in the framework. Therefore, Bachman (2005) claimed that the assessment use argument "can enable test developers and users to more clearly articulate (…) many of the qualities and concerns that language testers have discussed regarding test use" (p. 28). Pardo-Ballester (2010) used this assessment use argument to support a validity argument for an online Spanish listening exam. She found backings for three qualities of test usefulness: consistency, construct validity, and authenticity. Pardo-Ballester (2010) used Bachman's approach because it allowed her "to consolidate test design, development, scoring interpretations, and intended uses within a single model" (p. 140).

Since the AUA starts with the assessment records claim, Kim (2008) posited that this model is only useful for validation of existing tests. Kim argued (2008) that the impression is given that "test design decisions are only used as evidence in a retrospective way, that is, looking back at test design to evaluate arguments on test interpretations and use" (p. 36). Schmidgall (2017) criticized the framework because it consists of 39 potential warrants that need backing, which "implies a lot of documentation that could be difficult to create, maintain, and adapt" (p. 8).

Bachman and Palmer (2010) claimed that Kane's approach was confusing to test users and therefore suggested an updated version of Bachman's Assessment Use Argument. Bachman and Palmer (2010) posited that it consists of four general claims that each have warrants and rebuttals for which evidence needs to be provided. The consequence claim states that the consequence of using an assessment and of the decisions that are made are beneficial to all stakeholders. The decisions claim posits that the decisions take the community values into consideration and that they are equitable for the stakeholders. The interpretation claim states that the interpretations are meaningful, impartial, generalizable, relevant, and sufficient. The assessment records claim posits that the scores are consistent across different tasks, raters, and test takers. For each claim, several warrants and rebuttals were formulated by

Bachman and Palmer (2010) that "indicate the kinds of evidence that the test developers need to collect" to justify the use of the assessment (p. 95). According to Bachman and Palmer (2010), test developers are to be held accountable to the individuals who are affected by the use of the assessment and "must demonstrate through argumentation and the collection of supporting evidence" that the intended uses of the assessment are justified (p. 85). Thus, test developers need to articulate the specific statements that support the links between consequences and assessment performance and then collect the relevant evidence in support of the statements.

This approach is comprehensive because it captures more aspects of a test into validity studies than before (Schmidgall, 2017). Schmidgall (2017) claimed that it also helps "to clarify the implications of validity research" and can enhance the assessment literacy of the stakeholders because of its coherent framework (p. 3). Fulcher (2015) praised this approach because by highlighting the intended consequences and values of language testing, they embed "ethical considerations (…) into the earliest stages of test development" (p. 185). As such, they incorporate Messick's (1989) value implications and social consequences into their framework (Im et al., 2019). They also stress the importance of considering the values and perspectives of different stakeholders. By doing so, Im et al. (2019) describe this as "a more thorough process for the evaluation of inferences concerning decisions and consequences" than previous scholars (p. 14).

However, it seems as if Bachman and Palmer still used a checklist approach, which makes it difficult to operationalize (Im et al., 2019). Chapelle et al. (2010) maintained that a "taxonomy is not an argument" and that it does not force the developers to critically evaluate the strengths of the evidence (p. 9).

## The Expanded Interpretive/Use Argument

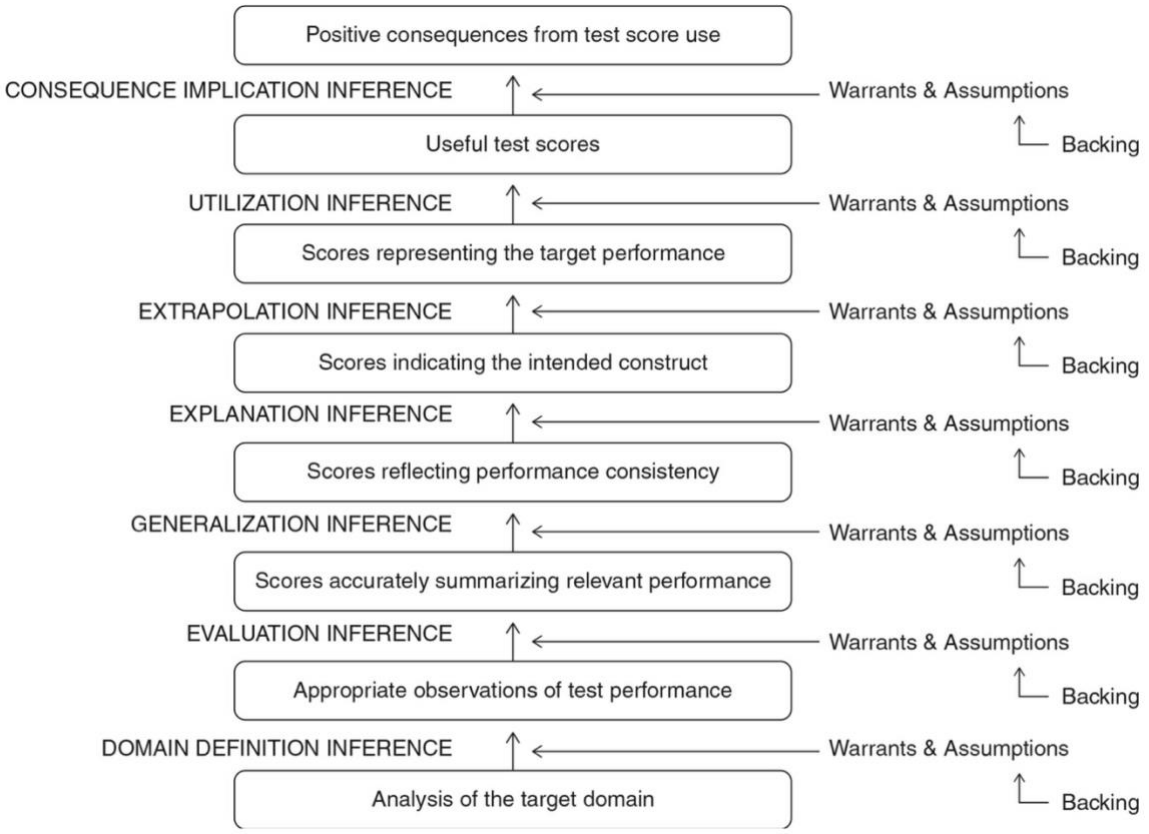
Chapelle et al. (2008) expanded the framework of Kane (1992) to develop an interpretive argument for the TOEFL test. Kane's framework missed a link to the underlying theoretical construct, and it did not "include test use, which needs to be a critical part of a validity argument" (p. 12). Therefore, the explanation inference was added to the framework that links the observed test performance to a construct. Kane's decision inference was renamed as the utilization inference, and it links the target score to the decisions for which the score will be used. The domain description inference was added to the framework, and it serves as the link between performance in the target domain and the observations of the performance in the test domain. Building on the bridge metaphor of Kane, Chapelle et al. (2010) viewed the backings of inferences as tokens that "are required to cross each bridge" (p. 9).

Chapelle et al. (2008) argued that this argument-based approach, in contrast with the evidence-gathering approach, provided them with "a way to organize the evidence and its implications" (p. 23). The addition of the explanation inference, according to Chapelle et al. (2010) "provided a place for the construct, without relying on it as a starting point or the central piece" (p. 12). Im et al. (2019) praised the addition of the domain description because it "allows test users to make better decisions to subsequently achieve the intended outcomes of a test" (p.12). Instead of looking for existing evidence, the approach by Chapelle et al. (2010) asks developers to "identify areas where additional research is needed" (p. 8). Chapelle et al. (2008) claimed that their approach forces test developers to look at validation from several perspectives and it provides a way to share the results with the different stakeholders "in a concise and transparent manner" (p. 349). Chapelle et al. (2011) called it the 'praxis step' in validation that makes validity research more "practical and useable" (p. 717). The

framework invites critics to challenge the assumptions and is, thus, according to Chapelle et al. (2008), a contribution "to the evolving concept of validation" (p. 320).

However, Im et al. (2019) argued that the role of the stakeholders was still too limited. The framework did not consider the value implications of the test scores, nor did it differentiate between decisions and consequences. According to Im et al. (2019), these decisions and consequences should be separated "to investigate how well selection decisions based on language test scores are supported and what consequences have been brought about" (p. 13). As can be seen in Figure 2, Chapelle (2021) added the consequence implication inference as "the final conclusion of a validity argument" (p. 40). It looks at how a test might benefit the users of the test and which benefits there are to society. As such, it underscores the claim that test score use should result in positive consequences for individuals and for society.

**Figure 2**
**Structure of a validity argument**




Chapelle and Voss (2021) maintained that this approach to argument-based validity could be used for several types of tests and at different stages in the development process. The approach could be applied to technology-mediated tests and is well suited for a mixed-method design. Chapelle and Voss (2021) claimed that by not being "so dependent on abstract constructs" (p. 328), the approach allowed for a more varied type of mixed-method research. Chapelle and Voss (2021) concluded that argument-based validity provides test developers with "tools for carrying out validation research in a way that is consistent with established professional perspectives" (p. 341). It underscores that validity is neither a yes or no decision about a test nor an objective deterministically derived result (Chapelle & Voss, 2021). Moreover, as Chapelle and Voss (2021) maintained, the framework provides the language testing community with a "shared common language" that might increase the assessment

literacy of the people (p. 344). As such, it does not try to change Messick's concept of validity; rather, "it provides a means of operationalizing" it (p. 342).

Xi (2008) and Purpura (2011) found that several quantitative research methods can be and have been used to develop a validity argument. Corpus studies, such as Biber et al. (2004) and Neff-VanAertselaer (2008), tried to find backings for the domain description inference. Many-facet Rasch measurement studies, such as Weigle (1998) and Eckes (2008), have been used to find evidence for the warrants underlying the scoring inference. (Multivariate) Generalizability studies, such as Kim (2008), were used to find support for the generalizability inference. Structural Equation Modeling studies, such as Phakiti (2008), tried to find evidence for the explanation inference. Correlation studies, such as Sawaki and Nissan (2009) and Xi (2007), looked at the extrapolation inference. Cognitive diagnosis studies, such as Sawaki (2009) and Jang (2009), were used to find support for the utilization inference. These studies led Purpura (2011) to conclude that "the validity argument offers multiple opportunities for the use of quantitative methods" (p. 739).

## CONCLUSION

In this paper, I demonstrated that the approach used to pursue validation studies has evolved throughout the years. At first, validity was viewed as consisting of several types: criterion (Cureton, 1950), content (Rulon, 1946) and construct validity (Cronbach & Meehl, 1955). In contrast to this componential approach, Messick (1980) showed that these types were all subsumed under construct validity, and he added the importance of values and consequences. To find evidence for the several facets of validity, Messick (1989) proposed using an evidence-gathering approach. Bachman and Palmer (1996) and Weir (2005) tried to operationalize this evidence-gathering approach through the test usefulness approach and socio-cognitive model, respectively. Kane (1992, 2006) criticized these checklist approaches for their limited practicality and instead proposed using an argument-based approach to validity. This approach influenced Bachman (2005) and Bachman and Palmer (2010) to develop an Assessment Use Argument. Later, Chapelle (2008, 2021) expanded Kane's Interpretive/use argument to make it more comprehensive by adding the explanation, domain definition, and consequence implication inference.

As such, the focus in validation studies shifted from evaluating distinct components of validity to developing a comprehensive argument for the use and interpretations of test scores. The argument-based approach to validity incorporates the distinct types of the componential approach, underscores Messick's attention to construct validity, highlights the need for evidence of the evidence-gathering approach, and combines it all into one logical argumentative structure. Such a comprehensive argument is now an indispensable part of validity studies (Dursun & Li, 2021). Consequently, validity scholars should use the argument-based approach to validity to test the developer's intended and actual meanings and uses of test scores.

## REFERENCES

American Educational Research Association & National Council on Measurements Used in Education. (1955). *Technical recommendations for achievement tests*. National Education Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1966). *Standards for educational and psychological testing*. American Psychological Association.

American Psychological Association. (1974). *Standards for educational and psychological testing*. American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement Used in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 57*(2), 1-38.

Anastasi, A. (1982). *Psychological testing.* Macmillan.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19–32). Lawrence Erlbaum Associates.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L.F. (2005) Building and supporting a case for test use. *Language Assessment Quarterly*, *2*(1), 1-34. https://doi.org/10.1207/s15434311laq0201_1

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Biber, D., Conrad, S. M., Reppen, R., Byrd, H. P., Helt, M., Clark, V., Cortes, V., Cosmay, E., & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus.* Educational Testing Service.

Bingham, W. V. (1937). *Aptitudes and aptitude testing.* Harper.

Buckingham, B.R. (1921). Intelligence and its measurement: A symposium – XIV. *Journal of Educational Psychology*, *12*, 271-275. https://doi.org/10.1037/h0066019

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105. https://doi.org/10.1037/h0046016

Cattell, R. B. (1964). Validity and reliability: A proposed more basic set of concepts. *Journal of Educational Psychology, 55,* 1-22. https://doi.org/10.1037/h0046462

Chalhoub-Deville, M. & O'Sullivan, B. (2020) *Validity. Theoretical developments and integrated arguments*. Equinox Publishing.

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. Sage Publishing.

Chapelle, C. A., Enright, M. K., Jamieson, J. M. (Eds.) (2008). *Building a validity argument for the test of English as a foreign language*. Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3-13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2011). *Building a validity argument for the test of English as a foreign language*. Taylor and Francis.

Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing*, *20*(4), 409–439. https://doi.org/10.1191/0265532203lt266oa

Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 3) (pp. 1-17). Wiley-Blackwell.

Chapelle, C.A., & Voss, E. (2021). *Validity argument in language testing. Case studies of validation research*. Cambridge University Press.

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

Cronbach, L. J. (1984). *Essentials of Psychological Testing (*4th ed.). Harper and Row.

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

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory and public policy* (pp. 147-171). University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302. https://doi.org/10.1037/h0040957

Cureton, E. E. (1950). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). American Council in Education.

Davies, A. (1977). The construction of language tests. In J. P. B. Allen, & A. Davies (Eds.), *Testing and experimental methods. The Edinburgh course in applied linguistics (Vol. 4).* Oxford University Press.

Davies, A. (1984). Simple, simplified, and simplification: What is authentic? In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 181-198). Longman.

Davies, A. & Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*. Lawrence Erlbaum.

Dursun, A., & Li, Z. (2021). A systematic review of argument-based validation studies in the field of language testing (2000–2018). In C. Chapelle, & E. Voss (Eds.), *Validity argument in language testing: Case studies of validation research* (*Cambridge applied linguistics*, pp. 45-70). Cambridge University Press. https://doi.org/10.1017/9781108669849.005

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics, 20*(2), 221-236. https://doi.org/10.1093/applin/20.2.221

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Routledge.

Garrett, H. E. (1947). *Statistics in psychology and education.* Longmans, Green.

Guilford, J. P. (1946). New standards for test evaluation. *Educational & Psychological Measurement, 6,* 427-438. https://doi.org/10.1177/001316444600600401

Guion, R.M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11,* 385-398. https://doi.org/10.1037/0735-7028.11.3.385

Im, G. H., Shin, D. & Cheng, L. (2019). Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. *Lang Test Asia, 9*(14). https://doi.org/10.1186/s40468-019-0089-4

Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading ability. *Language Assessment Quarterly*, *6*(3), 210–238. https://doi.org/10.1080/15434300903071817

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. (2001a). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. (2001b). *The role of policy assumptions in validating high-stakes testing programs*. Paper presented at the Annual meeting of the American Educational Research Association. Seattle.

Kane, M. (2006). Validity. In Brennan, R. L., National Council on Measurement in Education., & American Council on Education. (2006). *Educational measurement*. Praeger Publishers.

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182. https://doi.org/10.1177/0265532209349467

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1-73. https://doi.org/10.1111/jedm.12000

Khalifa, H. & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

Kim, J. Y. (2008). *Development and validation of an ESL diagnostic reading-to-write test: An effect-driven approach.* University of Urbana-Champaign.

Kunnan, A.J. (1998). *Validation in language assessment. Selected Papers from the 17th Language Testing Research Colloquium, Long Beach.* Lawrence Erlbaum Associates.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27-48). Cambridge University Press.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. Longmans, Green and Company.

Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing*, 17(1), 43–64. https://doi.org/10.1177/026553220001700102

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694. https://doi.org/10.2466/pr0.1957.3.3.635

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning.* Taylor & Francis.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955-966. https://doi.org/10.1037/0003-066X.30.10.955

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012-1027. https://doi.org/10.1037/0003-066X.35.11.1012

Messick, S. (1986). *The once and future issues of validity: assessing the meaning and consequences of measurement*. Educational Testing Service.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103.). American Council on Education and Macmillan.

Messick, S. (1992). *The interplay of evidence and consequences in the validation of performance assessments. Research report*. Educational Testing Service.

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. https://doi.org/10.1037/0003-066X.50.9.741

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7,* 191-205. https://doi.org/10.1177/001316444700700201

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229–258. https://doi.org/10.3102/00346543062003229

Neff-VanAertselaer, J. (2008). Arguing in English and Spanish: A corpus study of stance. *University of Cambridge ESOL Examinations Research Notes*, *33*, 28–33.

Norris, J. (2006). From test validation to validity evaluation in educational assessment. In R. Grotjahn & G. Sigott (Eds). *Language testing and evaluation*. Lang.

O'Sullivan, B. & Weir, C. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: Theories and practices.* Palgrave Macmillan.

O'Sullivan, B. (2005). *Levels specification project report*. Internal report, Zayed University, United Arab Emirates.

O'Sullivan, B. (2009). *City & Guilds Achiever Level IESOL Examination (B1) CEFR Linking Project Case Study Report*. City & Guilds Research Report.

Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: Test usefulness evaluation. *Language Assessment Quarterly, 7*(2), 137-159. https://doi.org/10.1080/15434301003664188

Pearson, K. (1896). VII. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 187*, 253-318.

Phakiti, A. (2008). Strategic competence as a fourth-order factor model: A structural equation modeling approach. *Language Assessment Quarterly*, *5*(1), 20–42. https://doi.org/10.1080/15434300701533596

Purpura, J. (2011). Quantitative research methods in assessment and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*. Routledge.

Purpura, J. (2021). A rationale for using a scenario-based assessment to measure competency-based, situated second and foreign language proficiency. In M. Masperi, C. Cervini, & Y. Bardière (Eds.), *Évaluation des acquisitions langagières: du formatif au certificatif*, *mediAzioni, 32*, A54-A96. https://doi.org/10.7916/zy23-wt92

Roever, C. & Mcnamara, T. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics*, 16, 242 – 258. https://doi.org/10.1111/j.1473-4192.2006.00117.x

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review*, *16*, 290-296.

Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, *6*(3), 239–263. https://doi.org/10.1080/15434300903079562

Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL iBT listening section* (Research Report, No. 09-02). Educational Testing Service.

Schmidgall, J. (2017). *Articulating and evaluating validity arguments for the TOEIC ® tests*. ETS Research Report Series.

Shaw, S., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge University Press.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405–450. https://doi.org/10.3102/0091732X019001405

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Routledge.

Spence-Brown, R. (2001). The eye of the beholder: Authenticity in an embedded assessment task. *Language Testing, 18*(4), 463–81. https://doi.org/10.1177/026553220101800408

Toulmin, S. (1958). *The uses of arguments*. Cambridge University Press.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287. https://doi.org/10.1177/026553229801500205

Weir, C. J. (2005). *Language test validation: An evidence-based approach*. Palgrave.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in the social sciences: A volume in honor of Lee J. Cronbach*. Erlbaum.

Xi, X. (2007). Validating TOEFL iBT Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly*, *4*(4), 318–351. https://doi.org/10.1080/15434300701462796

Xi, X. (2008). Methods of test validation. In E. Shohamy and N. H. Hornberger (Eds), *Encyclopedia of language and education* (pp. 177–196). Springer.

**Pieter Lauwaert** is a doctoral student in Applied Linguistics at Teachers College, Columbia University. His research interests focus on areas of second language assessment, including second language assessment validation. Correspondence should be sent to his email at pl2666@tc.columbia.edu.