Language Testing in Asia

REVIEW Open Access

Critical review of validation models and practices in language testing: their limitations and future directions for validation research

Gwan-Hyeok Im[1*], Dongil Shin[2] and Liying Cheng[1]

* Correspondence: gwan.im@queensu.ca
[1]Faculty of Education, Duncan McArthur Hall, 511 Union Street, Queen's University, Kingston, ON K7M 5R7, Canada
Full list of author information is available at the end of the article

Abstract

Purpose and background: The purpose of this paper is to critically review the traditional and contemporary validation frameworks—the content, criterion, and construct validations; the evidence-gathering; the socio-cognitive model; the test usefulness; and an argument-based approach—as well as empirical studies using an argument-based approach to validation in high-stakes contexts to discuss the applicability of an argument-based approach to validation. Chapelle and Voss (2014) reported that despite the usefulness and advantages of an argument-based approach for test validation, five validation studies using this approach were found in a search from two major journals—*Language Testing* and *Language Assessment Quarterly*. We reviewed the validation approaches in language testing and extended the search for empirical studies that used an argument-based approach in five language testing journals including ProQuest Dissertation and Theses. By doing so, this paper aims to provide validation researchers with each approach's conceptual limitations and future directions for validation research. For validity arguments to be defensible, this paper suggests that various validity evidences be required, involving multiple test stakeholders.

Implications: By comparing variations of an argument-based approach and reviewing eight representative studies out of 33 empirical validation studies using an argument-based approach, this paper presents the following implications for future researchers to consider: (a) defining test constructs and relevant test tasks through domain analysis; (b) inviting multiple test stakeholders to test validation; (c) investigating the intended and actual interpretations, decisions, and consequences; (d) considering social, cultural, and political values to be embedded; and (e) employing multiple methods beyond statistical analyses using test scores.

Keywords: Validity, Validation, An argument-based approach, Stakeholders, Multiple methods

Introduction

Standardized tests have been used as yardsticks of language ability or proficiency for the purposes of admissions, graduation requirements, overseas assignment, and hiring and promotion. For example, the Test of English for International Communication (TOEIC) is used for hiring and promotional decisions, and the International English

Language Testing System (IELTS) Academic and the Test of English as a Foreign Language (TOEFL) are used for university admissions. Standardized language tests have high-stakes for test stakeholders and generate tremendous consequences. Therefore, it is important to find out the extent test scores are interpreted as intended in terms of an indicator of test takers' language ability or proficiency, whether valid decisions have been made based on interpretations of test scores, and the consequences of using the test (i.e., making decisions based on test scores). To facilitate evaluation, validation frameworks play an important role to guide processes for collection of validity evidence.

The primary purpose of this paper is to review different approaches to test validation studies and to discuss the applicability of a widely used contemporary validation framework, which is an argument-based approach to validation (Bachman, 2005; Bachman & Palmer, 2010; Chapelle, Enright, & Jamieson, 2008; Kane, 1992, 2006, 2013). The argument-based approach addresses limitations in the traditional frames of validations (i.e., content, criterion-related, and construct validities) and makes Messick's (1989) construct validation more practical. Specifically, this paper reviews not only the traditional and contemporary validation frameworks but also empirical studies that investigated test validity in high-stakes contexts using an argument-based approach to inform researchers on both advantages and disadvantages for validation frameworks and on various methods for validation research. Finally, this paper discusses conceptual limitations of an argument-based approach and provides future directions for validation research.

Chapelle and Voss (2014) clearly articulate the evolution of test validity and validation in language testing research over the past few decades by focusing on the works of prominent scholars (e.g., Bachman, 2005; Bachman & Palmer, 1996; Carroll, 1980; Kane, 1992, 2001, 2002, 2004, 2006, 2013; Kane, Crooks, & Cohen, 1999; Messick, 1989). The evolution consists of the following validation approaches: (1) the one question and three validities (i.e., content, criterion-referenced, and construct validities to answer whether a test measures what it intends to measure), (2) the evidence-gathering, (3) the test usefulness, and (4) argument-based approach. In addition to the abovementioned approaches, the socio-cognitive model (Weir, 2005) is also reviewed in this paper because it has been applied to a number of comprehensive validation projects for the several recent large-scale, high-stakes applications, such as the Cambridge examination (e.g., Geranpayeh & Taylor, 2013; Khalifa & Weir, 2009; Shaw & Weir, 2007; Taylor, 2012), the Aptis testing system (O'Sullivan, 2012), the General English Proficiency Test (GEPT) in Taiwan (Wu, 2011) and the Test of English for Academic Purposes (TEAP) in Japan (Nakatsuhara, 2014; Taylor, 2014; Weir, 2014). This paper critiques different validation frameworks using all five of the aforementioned approaches to help validation researchers and graduate students to understand validation frameworks and provide them with practical guidelines for collection of validity evidence.

Early views on validation

This section reviews and discusses the early views on test validation, including the one question approach (content, criterion-referenced, and construct validations), the evidence-gathering, the test usefulness, and the socio-cognitive model.

The one question and three validities approach

A traditional approach to designing a test that accurately measures the abilities of test takers is for the test designer to answer this fundamental question: "does the test measure what it claims to measure?" (Lado, 1961, p. 321). This one question approach must be supported by three types of validity evidence (content, criterion-referenced, and construct). The concepts of validities are well articulated in the first and second editions of *Educational Measurement* (Cronbach, 1971; Cureton, 1951) as well as other related publications (American Psychological Association, American Educational Research Association,, and National Council on Measurement in Education, 1966; Carroll, 1980; Cronbach & Meehl, 1955)—all of which are cornerstone documents in guiding validity studies.

Content validity refers to how relevant and representative the test items are to the tasks in the target domain of interest. Typically, content validity is systematically evaluated by experts in the domain (Carroll, 1980). *Criterion-referenced validity* pertains to how test scores are correlated to scores of other (existing) scores hypothesized to measure the performance in the target domain (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014; Cureton, 1951; Fulcher, 2015). Correlations between test scores and criterion measures are provided as evidence for predicting the test taker's performance in the real-life settings. *Construct validity* refers to how well test scores conform to the theoretical test constructs (American Psychological Association, 1954; Cronbach & Meehl, 1955). Evidence through statistical analyses is collected for construct validation (Chapelle & Voss, 2014), the effect of which is to make test scores the object of validation (Fulcher, 2015). Construct validity has been regarded as a fundamental consideration in test design (Loevinger, 1957) because operationalizing test constructs is the main concern in construct validity and "content and criterion validities are all essentially ad hoc" (Loevinger, 1957, p. 636). Construct validation was used as a general framework for validation (Anastasi, 1986; Embretson, 1983; Guion, 1977; Kane, 2006; Messick, 1988, 1989) and had subsumed the content and criterion validations.

Despite the prevalence of validation practices based on the traditional validities, researchers have noted that it may lead to insufficient validity evidence for the meaning of test scores (Fulcher, 2015; Messick, 1989; Shepard, 1993). The validities do not include collecting evidence for the use of test scores and consequences for high-stakes testing, which are more evident in language testing (Fulcher, 2015). Regarding content validity, it is difficult to judge how representative the test items are of the abilities in the target domain of interest because experts may have differing judgments on the items (HajiPourNezhad, 2003). Evidence to support content validity also provides little information on how test scores are interpreted and used as a prediction of performance in real-life settings, due to the emphasis on test formats (Shulha & Wilson, 2009). Evidence to support criterion-referenced validity may not be sufficient to infer the test taker's ability because there may be a large number of factors related to predicting future performance. For example, a high score on a written driving test, which is highly reliable with different types of driving test scores, may not be sufficient evidence of a safe driver with a clean driving record (Shepard, 1993). Construct validity has issues of inherent difficulties in defining abilities in terms of the abstract nature of constructs (Kane, 2013). Test designers often focus on the evidence to prove the usefulness of

their tests in examining a construct without investigating evidence that could rebut their claims about test taker's abilities based on test scores (Cronbach, 1988). To address this issue in construct validity, Messick (1989) broadened the scope of construct validity to include two threats to construct validity: construct underrepresentation (test constructs' narrow reflection of actual performance in the real-life settings, e.g., lack of interaction in a listening test), and construct irrelevant variance (unrelated factors to test constructs, e.g., test-wiseness). The limitations of the traditional approaches are the major reasons why contemporary scholars have increasingly turned to Messick's (1989) evidence-gathering approach to test validation. The evidence-gathering approach to test validation has broadened the scope of validation to include social aspects of a test, which will be discussed in the following section.

Evidence-gathering

Rather than focusing on the quality of the test, like in positivistic approaches, Messick (1989) defined validity as "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, italics in original) in interpretive approaches to validation. According to Messick (1989, p. 13), "integrated" means that validity is a unitary concept to include subjective judgment based on values (i.e., "evaluative judgment"). "The degree" means that validity is a matter of degree required to provide "empirical evidence and theoretical rationales" to support the adequacy of inferences and actions (i.e., psychometric property of a test: evidential basis in Table 1) and appropriateness of inferences and actions (i.e., values and consequences: the consequential basis in Table 1).

Messick (1989) provides two facets of validity—*test interpretation* and *test use*—along with two types of evidence—the *evidential basis* and the *consequential basis*—as opposed to the three types of validity provided the traditional approach. This definition calls for involving multiple inquiry methods and various validity evidences (Kane, 2001; Messick, 1989; Moss, 1996), and making arguments regarding how well test scores are interpreted and used. What is validated in the one question approach is how well a test measures what it intends to measure, whereas the evidence-gathering approach finds out how well the intended meaning and use of test scores are supported by evidence and theoretical rationales (Messick, 1996). In other words, instead of a test alone, interpretations and uses of test scores should be validated, because a single test can be interpreted differently for different purposes (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). In addition, depending on score use contexts, the meaning of test scores may be understood differently by stakeholders; therefore, test scores may be used in a different way from the intended use (Koch & DeLuca, 2012; Macqueen, Pill, & Knoch, 2016; Zumbo, 2015). Accordingly, evidence for the intended score meaning and use is

Table 1 Facets of validity as a progressive matrix (Messick, 1989, p. 20)

	Test interpretation	Test use
Evidential basis	Construct validity (CV)	CV + Relevance/utility (R/U)
Consequential basis	CV + R/U + value implications (VI)	CV + R/U + VI + social consequences

gathered to ensure the defensibility of the argument about the test taker's abilities and decisions made about them (Messick, 1989).

The framework in Table 1 (Messick, 1989, p. 10) represents a progressive matrix regarding different facets of validity. The foundation of the evidence-gathering approach, which appeared in the third edition of *Educational Measurement*, is, at its core, construct validity, which also utilizes the one question and three validities together into one framework for collecting evidence for evaluation of score meaning. Then, in order to better make decisions in particular contexts, it added the relevance and utility of test scores to the approach. Specifically, statistical analyses for differential item functioning (DIF) are conducted to flag items that may discriminate a certain group of test takers to ensure absence of bias (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

In addition to the psychometric trait of a test, Messick's (1989) evidence-gathering approach includes value implications and social consequences into test validation. Value implications can be defined as values attached to test interpretation and test use (Im & McNamara, 2017), although Messick (1989) defined value implications as values only attached to the test interpretations. For example, Im and McNamara (2017) reported in a study of Korean university admissions officers' ($n = 20$) understanding of TOEIC Listening and Reading (TOEIC LR) scores that some admissions officers viewed the scores as an indicator of the student's efforts and/or diligence, while others viewed the scores as a reflection of former residence in English-speaking countries, or as test-wiseness. Im and McNamara (2017) also found that values were embedded in test use, as TOEIC LR scores were employed due to social pressures attached to the use of TOEIC LR scores. Regarding the use of TOEIC LR scores, the admission officers considered lower test fees and higher testing site availability of the TOEIC LR than other English tests. In addition, they noticed that students who received admission with TOEIC LR scores had higher possibility to get a job after graduation because they already met the company requirements for English language ability, and they tended not to drop out their school. Graduates' employment rate, drop-out rate, and financial stability were highly related to university admissions using TOEIC LR scores to meet the criteria in the annual university evaluation. These value implications affect test users' understanding of score meaning and use of test scores as well as social consequences derived from the test use. Messick (1989) suggests that the evidential basis and the consequential basis are not exclusive within the unified view.

Messick's (1989) evidence-gathering approach has been influential on educational assessment and language testing in terms of broadening the scope of validation to include the consequential basis, and various validation approaches have been developed based on Messick's evidence-gathering approach, e.g., test usefulness (Bachman & Palmer, 1996), the socio-cognitive model (Weir, 2005), and the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999, 2014) for including the consequential basis and fairness. Furthermore, the concept of the consequential basis has enabled validation researchers to consider multiple purposes of a test (e.g., Koch & DeLuca, 2012; Stobart, 2009) as well as test stakeholders' values in score meaning and use (e.g., Elder, 1997; Im & McNamara, 2017; McNamara & Roever, 2006).

Despite its influential work on test validation, Messick's evidence-gathering approach may lead to potential misperceptions because, as shown in Table 1 above, construct validity relates to all four facets of validity, where "it is not clear whether the term labels the whole or the part" (Shepard, 1993, p. 427). Value implications are also located in the related test interpretation and test use in consequential basis, which gives "the impression that values are distinct from a scientific evaluation of test score meaning" (p. 427). Furthermore, his approach has been criticized in that it does not fully provide validation researchers with a methodological guideline in terms of *how much* of *what kind* of evidence should be collected during test validation (Chapelle, Enright, & Jamieson, 2010; Chapelle & Voss, 2014; Kane, 2011). Messick (1995, 1996) indeed provided the methodological guideline for test validation, pointing out six aspects of construct validity and sources of evidence to each aspect: (a) content aspect (i.e., evidence of relevance, representativeness of test item, or tasks to the target domain), (b) substantive aspect (i.e., theoretical rationales and empirical evidence for the test taker's performance), (c) structural aspect (i.e., evidence for the validity of scoring criteria or rubrics), (d) generalizability aspect (i.e., evidence for the consistent score), (e) external aspect (i.e., relationships with criterion and other criterion measure score), and (f) consequential aspect (i.e., evidence and rationales for evaluation of intended and unintended consequences). Messick (1995) noted that corresponding evidence to all of the six aspects should be collected in terms of a unified concept of construct validity. Nevertheless, he also noted that an inability of gathering corresponding evidence to the six aspects does not invalidate score meaning and use, but validation researchers need to provide an argument for such omissions (Messick, 1995).

Messick's (1989, 1995, 1996) evidence-gathering approach clearly offers *how*, *how much*, and *what kind* of validity evidence should be collected during test validation. However, how to integrate the evidential basis and the consequential basis into validation still remains unresolved (McNamara & Roever, 2006). Contemporary validation researchers try to integrate Messick's (1989) consequential basis into validation approaches with more logical, practical processes for test validation, although integration of the consequential basis into validation is a controversial issue (Crocker, 1997; Linn, 1997; Mehrens, 1997; Popham, 1997). The following sections discuss major validation approaches following Messick's (1989) evidence-gathering approach that addresses the limitations of previous approaches.

Test usefulness

Advocating the perspective of Messick's (1989) evidence-gathering approach to validation, Bachman and Palmer (1996) adopted the test usefulness approach to provide a more explicit guideline for *how much* of *what kinds* of validity evidence should be collected. Their approach includes six qualities: *construct validity*, *reliability*, *authenticity*, *interactiveness*, *impact*, and *practicality*. The six qualities to be balanced have a total of 42 questions, each of which needs to be addressed during the validation processes. If not all questions within each quality are addressed, the intended meaning of test scores is considered indefensible.

Regarding addressing all of the questions during the validation processes, Weigle (2002) contends that the degree to which each quality is minimally accepted depends

on the testing contexts. In other words, the qualities that should be weighted depend on test purpose, cost, time, and contexts.

Bachman and Palmer's (1996) approach is useful in terms of providing a checklist about what qualities must be investigated and what kinds of evidence must be collected. However, addressing all questions within each quality may be a burden on validation researchers, in terms of time and difficulty answering them (Lewkowicz, 2000) and might "mask any real problems with alternative explanations" (Fulcher, 2015, p. 119) once researchers have answered all of the questions. This may account for why Bachman and Palmer (2010) adopt an argument-based approach to validation in their later work (Fulcher, 2015).

Socio-cognitive model

The most comprehensive checklist approach is evident in Weir's (2005) socio-cognitive model, which was further developed by O'Sullivan and Weir (2011). It provides *what kind* of elements (evidence) need to be considered (collected) during test development and validation. Initially, Weir (2005, pp. 44–47) specified five types of validity categories: *context validity*, *theory-based validity*, *scoring validity*, *consequential validity*, and *criterion-related validity*. Context and theory-based validities pertain to defining test constructs and relevant test tasks for test specification, which correspond to Messick's (1989) content aspect (i.e., evidence of relevance, representativeness of test item, or tasks to the target domain) and substantive aspect (i.e., theoretical rationales for the test taker's performance with empirical evidence) of construct validity. Weir (2005) added the testing environment to Messick's content aspect. Aspects of the test taker's cognitive processing elements associated with test tasks were specified under the theory-based validity (see Weir, 2005, p. 46). In addition, the test taker characteristics such as physical (e.g., age, ailments, and disabilities), psychological (e.g., memory, personality, and emotional state), and experiential (e.g., familiarity with testing) characteristics are included as validity evidence because these characteristics affect the meaning of test scores.

Except for the addition of testing environment, however, the socio-cognitive model does not seem to add much to Messick's (1989) construct validity (Fulcher, 2015), other than providing exhaustive elements (i.e., evidence) to be considered (collected) during test development and validation. For example, the scoring validity pertains to the structural aspect (evidence for the validity of scoring criteria or rubrics) and the generalizability aspect (evidence for the consistent score), as the criterion-related validity and the consequential validity correspond to the external aspect (relationships with criterion and other criterion measure score), and the consequential aspect (evidence and rationales for evaluation of intended and unintended consequences), respectively.

The socio-cognitive model provides a more specific and explicit guideline for *what kinds* of validity evidence should be collected. However, like Bachman and Palmer's (1996) test usefulness approach, it may not be easy to operationalize because it may be time-consuming and complex for every element to be addressed (Fulcher, 2015). Furthermore, it seems that the elements listed in the theory-based validity only reflect the perspectives of test developers or language specialists, as they specify test takers' cognitive processing elements. There still remains a question of whether the socio-cognitive

model encourages the test taker's involvement for defining test constructs. Another limitation is that the content knowledge added to theory-based validity could be either relevant or irrelevant to the construct of the test, depending on its purpose (Fulcher, 2015).

In summary, the approaches to validation that have been reviewed so far have limitations in terms of insufficiency of evidence for the meaning of test scores, their use, and their consequences. Hence, an argument-based approach to validation was adopted by Kane (1992, 2001, 2002, 2004, 2006) and Kane et al. (1999), and it has been widely accepted and revised by scholars (e.g., Bachman, 2005; Bachman & Palmer, 2010; Chapelle et al., 2008; Kane, 2013; Mislevy, Steinberg, & Almond, 2003). The following section examines various applications of an argument-based approach to validation in more detail.

Contemporary views on validation

The checklist approach is identified as "promiscuous empiricism" (Fulcher, 2015, p. 117) due to its complexity and time-consuming work; an argument-based approach proposes an opposite strategy. Addressing issues of practicality in test validation, Kane (1992) first adopted an argument-based approach to validation, using Toulmin's (1958, 2003) argument model, which is evident in Kane's (2006) work in the fourth edition of *Educational Measurement*. The argument-based approach has been one of the most influential approaches to validation because it provides a simple, systematic process for how validation researchers structure validity arguments, linking validity evidence for the development and use of a test. Researchers can flexibly determine what claims they want to make based on test scores and what kind of evidence to collect to support the claims, depending on testing contexts and test uses (Chapelle, Chung, Hegelheimer, Pendar, & Xu, 2010; Kane, 2013). In this section, the variations of an argument-based approach are compared and contrasted with Messick's (1989) evidence-gathering approach. This comparison will help identify the strengths and weaknesses of each variation.

Terms that commonly appear within the argument-based approach are claim, data, a warrant with backing evidence, a rebuttal with rebuttal evidence, and an inference. Figure 1 below provides a simple example to explain the terms. To illustrate, let us say a student comes to the library everyday (data). You may conclude he or she is a

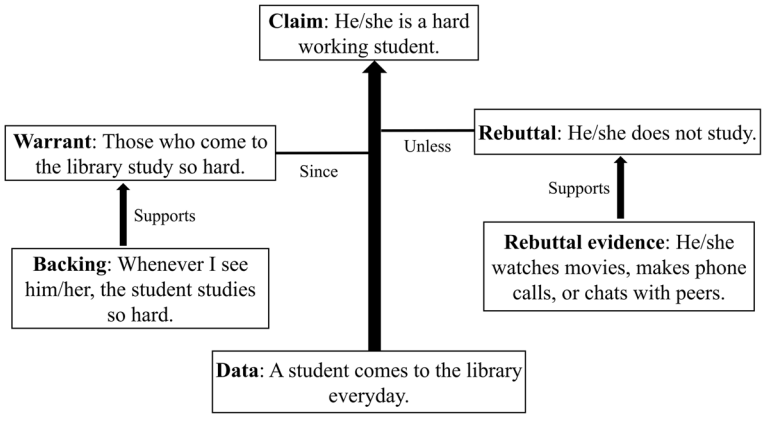

Fig. 1 Example of an argument model

hardworking student (claim). To support your conclusion, there should be a warrant, i.e., a general rule, which would be that those who come to the library study so hard (warrant). Then you should provide evidence to support the warrant, which would be that whenever you see him/her, the student studies so hard (backing), unless the student does not study (rebuttal), watching a movie, making phone calls, or chatting with peers (rebuttal evidence).

Applying the argument model to testing, a claim is made based on data (e.g., the test taker's performance on a test). The claim is supported with a warrant, which refers to "a law, generally held principle, rule of thumb, or established procedure" (Chapelle et al., 2008, p. 6). The warrant is not supported by itself, but needs backing evidence (e.g., validity study). The claim is supported if any alternative explanations that may be used as a rebuttal are refuted (e.g., a bias review). An inference links data to claim.

However, there are variations in using the terms by different theorists. For example, while Kane (1992, 2006, 2013) and Chapelle et al. (2008) use both claims and inferences, Bachman and Palmer (2010) and Mislevy et al. (2003) use claims. Figure 2 below provides comparisons of the applications of an argument-based approach to Messick's (1989) validation framework. It is worth reviewing the various applications in light of Messick's (1989) evidence-gathering approach to identify the strengths and limitations each application has. White, dotted, gray, and checkered ovals correspond to Messick's (1989) facets of validity: construct validity, relevance/utility, value implications, and consequences, respectively. Furthermore, corresponding inferences or claims among the applications of an argument-based approach are arranged in Fig. 2. For example, the *universe score* claim in Kane (2006, 2013) corresponds to the *expected score* claim in Chapelle et al. (2008) and the *assessment records* in Bachman and Palmer (2010).

Kane's (1992, 2006, 2013) interpretive argument and IUA

Following Cronbach's (1988) "five perspectives on validity argument" (p. 330), Kane (2006) identified two types of arguments: an *interpretive argument* followed by a *validity argument*. The interpretive argument relates to specifying types of claims or inferences regarding the intended meaning and use of test scores in a coherent way (Kane, 2004). The validity argument refers to evaluation of the interpretive argument with empirical evidence to support (i.e., backing evidence) or challenge the claims (i.e., rebuttal evidence) (Kane, 2004, 2013). Put simply, the interpretive argument outlines what steps in validation need to go through in what way, while the validity argument refers to how well evidence supports or challenges the interpretive argument. These interpretive and validity arguments have been supported by contemporary validity researchers (e.g., Bachman, 2005; Bachman & Palmer, 2010; Chapelle et al., 2008 in language testing; Kane, 1992, 2006, 2013; Mislevy et al., 2003 in educational assessment), and various applications have been observed in test validation frameworks.

Figure 2 illustrates Kane's (2006) interpretive argument pertaining to specifying claims or inferences[1] proposed by the major validation researchers. Construct validity in Messick's (1989) evidence-gathering approach is mainly focused on in Kane's (2006) interpretive argument as well as other applications. As shown in the second row of Fig. 2, the interpretive arguments specify the scoring inference that links observation as data (test takers' performance on a test) to observed score (test scores), the

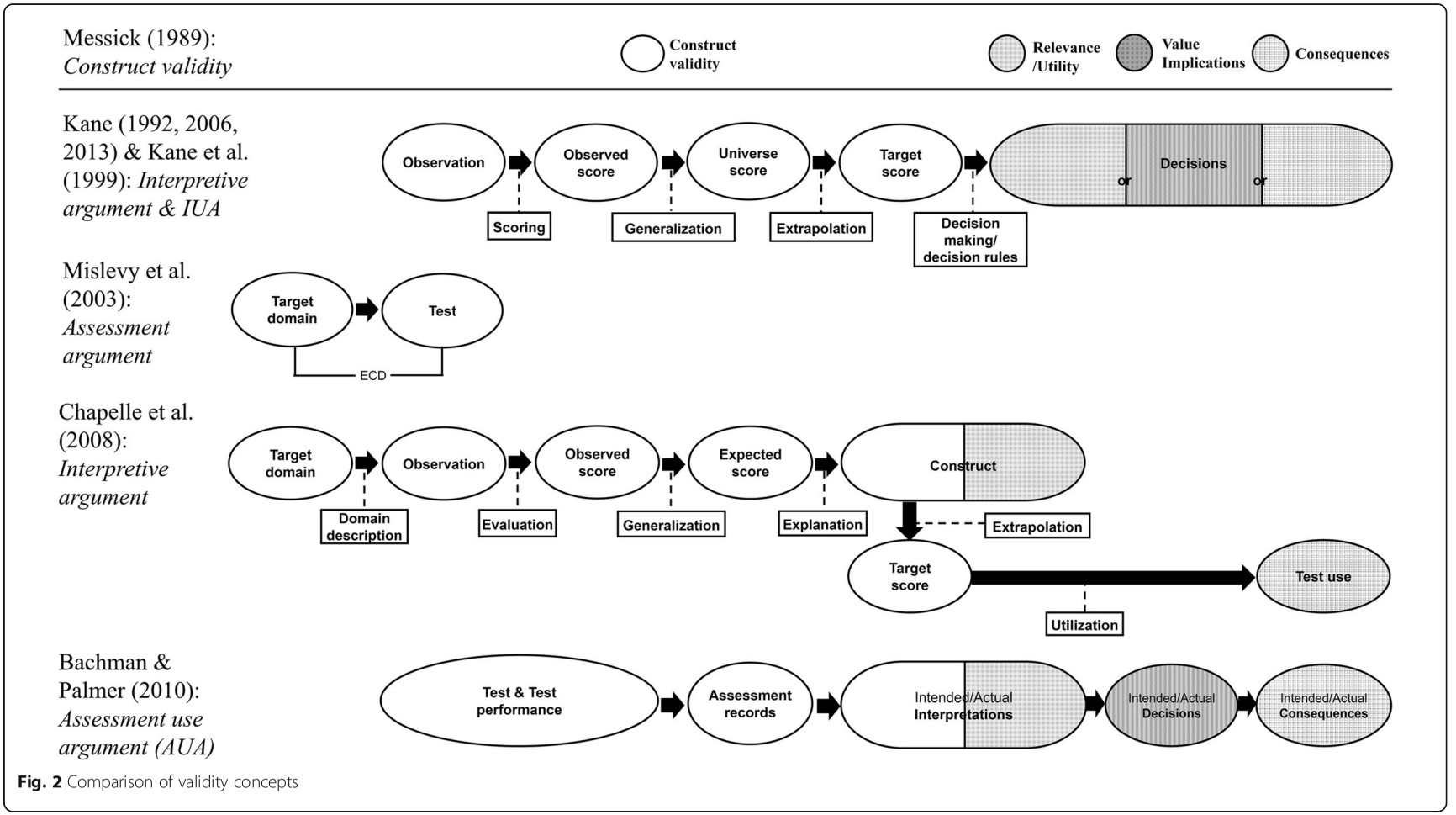

Fig. 2 Comparison of validity concepts

generalization inference that links the observed score (test score) to the universe score (consistent score), the extrapolation inference that links the universe score (consistent score) to the target score (the predicted performance in the real-life) as evidence for test use, and finally the decision-making inference that links the target score to decisions as a claim. For a validity argument, evidence in support of the interpretive arguments can be collected by experts' judgments on scoring rubrics, procedures for the scoring inference, reliability studies for the generalization inference, and correlation studies for the extrapolation inference. The decision-making inference is examined in terms of utility, values, or consequences. The criteria employed depend on whether a decision based on test scores results in positive impact with a reasonable cost and few negative side effects (i.e., unintended consequences).

Kane's (2002, 2006) interpretive argument, however, does not provide an explicit methodological guideline for validation of the decision-making inference. In his later work, Kane (2013) has changed the term *interpretive argument*, to *interpretation and use argument* (IUA) to make an equal emphasis between score meaning and score use. Almost nothing has been changed regarding types of inferences. However, Kane (2013) changed the term of decision making (or decisions) into decision rules and further specified the types of consequences as evaluation criteria: (a) intended outcomes (the extent to which intended outcomes are achieved), (b) adverse impact (discriminative impact on groups), and (c) systemic effect (positive or negative effects on teaching and learning). This contributes to providing a conceptual guideline for evaluating consequences of test use, although the IUA does not provide an explicit methodological guideline for validation of the decision-making inference.

In addition, in Kane's interpretive argument and IUA, the substantive aspect of construct validity Messick (1995, 1996) provided (i.e., theoretical rationales for the test taker's performance with empirical evidence) is missing. Kane (2013) focuses on defining test constructs based on only observed performance in a specific domain without considering theoretical construct definitions, due to the difficulties in conceptualizing constructs. Therefore, Kane's interpretive argument and IUA do not have a phase to define constructs, which is critical to test design, before a scoring inference.

Mislevy et al.'s (2003) evidence-centered design

With a focus on defining test constructs, Mislevy et al. (2003) highlight the analysis of knowledge and task types in the target domain of interest in terms of providing steps for test design in accordance with the test's purpose (Chapelle et al., 2008; Mislevy et al., 2003). They proposed an evidence-centered design (ECD), which made explicit the domain analysis in terms of defining theoretical constructs based on the analysis of performance as well as identifying types of tasks to elicit the performance in the target domain.

This is a complementary approach to operationalize test constructs in order for test scores to reveal more accurate information on language abilities in language testing. As illustrated in the third row of Fig. 2, the ECD, however, does not include the context of test score use, e.g., the value implications and consequences in Messick's (1989) framework, and therefore, it is seen primarily as a procedure for test specification (McNamara & Roever, 2006). Due to this limitation, empirical validation studies based

on Mislevy et al.'s (2003) ECD were not included in the review of empirical studies using an argument-based approach.

Chapelle et al.'s (2008) interpretive argument

Addressing the limitations in the previous validation frameworks, Chapelle et al. (2008) have adapted and applied approaches of Kane (1992, 2001), Mislevy et al. (2003), and Bachman (2005) to language testing in a validation study on the Test of English as a Foreign Language internet-Based Test (TOEFL iBT). As shown in the fourth row of Fig. 2, following the same terminologies found within Kane's (1992, 2001) interpretive argument, Chapelle et al. (2008) added the *domain description* inference for identifying both language proficiency and tasks on Mislevy et al.'s (2003) ECD approach—as part of the construct definition—to Kane's (1992, 2006) interpretive argument. Additionally, Chapelle et al. (2008) made explicit the *explanation* inference in their validation approach for examining whether test scores account for the test taker's performance from theoretical perspectives. Chapelle et al. (2008) also replaced the decision inference (Kane et al., 1999) with *utilization* (Bachman, 2005), which specifically characterizes the four types of warrants for a claim about a decision: utility, relevance, sufficiency, and outcomes.

Chapelle et al.'s (2008) interpretive argument provides holistic and systematic processes for test development and test validation as it starts with the domain description inference linked to subsequent inferences and ends with the utilization inference for evaluating decisions and consequences of testing. One of the strengths in Chapelle et al.'s (2008) interpretive argument is that having the domain description and explanation inferences that require providing evidence for defining and evaluating test constructs, respectively, allows test users to make better decisions to subsequently achieve the intended outcomes of a test.

For validity arguments, for example, Chapelle et al. (2008) conducted domain analysis as the first step for two weeks through collecting written and spoken texts from students and teachers from five universities in the USA in order to identify language use of academic tasks. The researchers also used a survey where participants (i.e., students and faculty members) selected which statements of academic tasks were more important. Additionally, researchers in applied linguistics identified language proficiency for academic tasks. Chapelle et al. (2008) developed tasks with experts' judgments on language proficiency required for academic tasks and with test performance of students on task types. Next, the researchers went through statistical analyses such as measures of consistency, generalizability, factor analysis, correlation analysis of test scores, and raters' rating processes including observation of students' test taking processes. At the end of the processes, they investigated the impact of the TOEFL iBT using interviews with test stakeholders and observations of students. This is a comprehensive validation study that included various test stakeholders and various methods for collecting evidence.

However, Chapelle et al. (2008) put the stakeholders' involvement into a limited dimension of test validation. When they used test scores for TOEFL validation, it may not lead to investigating the alignment between the intended and actual meaning and use of test scores in given contexts, which will further result in limitations in investigating value implications in the contexts. In addition, their interpretive argument has only

one inference (i.e., the utilization inference) for evaluating both decisions and consequences together. This may lead to an evaluation of either decisions or consequences. By distinguishing between decisions and consequences, researchers will be able to investigate how well selection decisions based on language test scores are supported and what consequences have been brought about.

Bachman and Palmer's (2010) assessment use argument

Based on an argument-based structure used in Kane (1992) and Mislevy et al. (2003) and Messick's (1989) unified concept of validity, assessment use argument (AUA) that focused on test use was first proposed by Bachman (2003, 2005) and further developed by Bachman and Palmer (2010). The initial version of the AUA (Bachman, 2005) consisted of two arguments: the *assessment validity argument* that links test performance to the meaning of test scores, and the *utilization argument* that links the score meaning to a decision about a test taker. Instead of laying out types of inferences, Bachman (2005) specified types of claims and warrants to be support by evidence. The initial version of the AUA predominantly focused on the utilization argument, specifying four types of warrants for claims of decisions: (a) relevance, (b) utility, (c) sufficiency, and (d) intended consequences. However, as shown in the last row of Fig. 2, the recent version of the AUA (Bachman & Palmer, 2010) further specified types of claims for consistent test scores (i.e., *assessment records* claim), score meaning (i.e., *interpretations* claim), test use (i.e., *decisions* claim), and consequences (i.e., *consequences* claim). As shown in interpretations, decisions, and consequences claims in Fig. 2, Bachman and Palmer (2010) emphasize the evaluation of the *intended* and *actual* interpretations, decisions, and consequences. This evaluation articulates benefits derived from test stakeholders for validation studies (Fulcher, 2015).

In terms of the relation of the AUA to other researchers' (e.g., Chapelle et al., 2008; Kane, 1992, 2006, 2013; Mislevy et al., 2003) approaches, the assessment records claim relates to the claim of the universe score and the expected score (i.e., consistency of test scores) in relation to the generalization inference in Kane's (2006) and Chapelle et al.'s (2008) interpretive arguments respectively. The interpretations claim corresponds to the claim of the target score in Kane (2006, 2013) and the claims of construct and target score in Chapelle et al. (2008) in relation to the extrapolation and explanation inferences and further includes *relevance* in Messick's evidence-gathering approach, has five warrant categories to be justified with evidence: (a) *meaningfulness*, (b), *generalizability*, (c) *impartiality*, (d) *relevance*, and (e) *sufficiency* (see Bachman & Palmer, 2010). Meaningfulness relates to construct validation in terms of the extent to which test scores conform to theoretical constructs. Generalizability pertains to the connection between test tasks and the target domain tasks, also relating to content validity. Impartiality relates to the same interpretations of test scores for all groups of test takers. Finally, relevance and sufficiency inform decision making in all facets of assessment.

Messick's (1989) value implications and social consequences are specifically manifested in warrants for the decisions and consequences claims (value sensitive and equitable decision(s) and beneficial consequences respectively), while Kane (1992, 2006) put them together in the decision-making or decision rules inference. *Values sensitivity* means that values in the community needs to be taken into consideration for decision-

making while *equitability* includes that test takers who are in similar levels of ability have the same chances of being classified into the same group (Bachman & Palmer, 2010). In particular, value sensitivity corresponds to Messick's (1989) value implications that invite social, cultural, and political values to the validity discussions. Bachman and Palmer (2010) further point out that values of different stakeholders need to be considered to bring about beneficial consequences. By having value sensitivity in decisions claim, Bachman and Palmer (2010) try to help language testers understand the contextual factors that may influence interpretations and uses of test scores. For supporting the claim of consequences, Bachman and Palmer (2010) reason that the use of test scores and decisions based on the scores bring about beneficial consequences to test stakeholders. The AUA provides a useful, more thorough process for the evaluation of inferences concerning decisions and consequences with a more emphasis on test use, by distinguishing the decisions and consequences.

However, Bachman and Palmer's (2010) AUA starts with assessment records claim which pertains to providing evidence for consistent test scores. A validity argument for test constructs and test tasks' representativeness of the target language use (TLU) domain is provided as backing for the subsequent claim, i.e., interpretations claim, which "might give an impression that the validity argument is formulated after a test is designed and administered, and that test design decisions are only used as evidence in a retrospective way" (J. Y. Kim, 2008, p. 36). In addition, Bachman and Palmer (2010) explicitly provided an exhaustive list of warrants to support the claims (Schmidgall, 2017) . It appears that Bachman and Palmer (2010) still adopt a checklist approach, which seems difficult to operationalize in test validation. Each claim except for assessment records claim has a number of warrants to be supported by evidence: one warrant for assessment records claim, 17 warrants for interpretations claim, two warrants for decisions claim, and five warrants for consequences claim. It would be helpful to evaluate almost all warrants to ensure the validity of meaning and use of test scores. However, providing a list of warrants may not be related to the nature of an argument-based approach, as Chapelle et al. (2010) points out that "a taxonomy is not an argument, and in working with a taxonomy one is not prompted to look at the strength of the evidence or to organize it in a way that presents a validity argument" (p. 9).

Empirical studies using an argument-based approach

For a review of empirical studies based on argument-based approaches, we extended the scope of search by adding three more journals[2] to the search of Chapelle and Voss (2014) as well as doctoral dissertations using database, ProQuest Dissertation, and Theses, published or completed from 1992 to 2016.[3] However, the studies using ECD approach (Mislevy et al., 2003) were not included for the search because the approach does not include validity arguments for test use, which is often considered necessary in test validation (Kane, 2013). In total, 33 journal articles and dissertations were found, and in this paper, we review two journal articles (i.e., Brooks & Swain, 2014; Chapelle, Chung, et al., 2010) and six dissertations (i.e., Hsu, 2012; Kadir, 2008; Lim, 2009; Liu, 2014; Sun, 2016; Tominaga, 2014) which were conducted for high-stakes testing. The reasons for selecting these studies for analysis are that the eight studies are representative of each variation of an argument-based approach and include conventional and unique methods such as statistical analyses using test scores, interviews with

stakeholders, conversation analysis, and recordings of speaking activity to compare test takers' performance on a test and in the target domain. Reviewing all empirical studies from the search may lead to redundancies.

Overall, most of the empirical studies except for Sun (2016) did not collect validity evidence from multiple stakeholders considering contextual factors (DeLuca, 2011; Shepard, 2000) that affect the meaning of test scores and the test use. A variety of instruments were used including language tests, questionnaires, interviews, conversation analysis, experts' (i.e., applied linguists) judgments, and recordings of in-class and out-of-class speaking activity. However, most of the studies focused on the psychometric properties of the test with reliability and the traditional concept of test validity (accuracy of test scores) with limited involvement of stakeholders.

It was found that some (e.g., Kadir, 2008; Sun, 2016) went through the whole processes of evaluation while the rest partially evaluated the certain inferences or claims in each application: the scoring inference (e.g., Hsu, 2012; Lim, 2009; Tominaga, 2014); the extrapolation inference (e.g., Brooks & Swain, 2014; Liu, 2014); and the domain description inference to the extrapolation inference (e.g., Chapelle, Chung, et al., 2010). For example, Hsu (2012), Lim (2009), and Tominaga (2014) only evaluated the scoring inference (the evaluation inference in Chapelle et al. (2008) interpretive argument) with different methods based on different applications of an argument-based approach.

Drawing on Chapelle et al.'s (2008) interpretive argument, Lim (2009) investigated the effects of questions in the writing component of the Michigan Language Assessment Battery (MELAB) and the effects of raters on test takers' responses. Researchers used statistical analyses such as Multi-Facet Rasch analysis to examine item difficulty and raters' bias, including information about consistency of test scores (Bachman, 2004) and analysis of variance to examine whether there is a mean difference among types of writing questions. It was reported that different prompts and different raters did not threaten the validity of test scores in the writing components.

Other than investigating the raters' bias based on test scores, Hsu (2012) investigated how raters' perceptions on the World Englishes (McArthur, 2001) affected test scores in the International English Language Testing System (IELTS) speaking test, drawing on Kane's (1992, 2002, 2004) interpretive argument. A questionnaire and interviews with raters were used, and it was reported that raters generally had positive attitude toward World Englishes, but ratings by those who had positive and negative attitudes toward World Englishes showed significant difference in ratings.

While the two studies evaluated the scoring inference by looking at raters' bias, Tominaga (2014) compared the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) Japanese scoring criteria to test takers' speaking performance in the test regarding discourse features in Japanese using conversation analysis, drawing on Kane's (2006) interpretive argument. It was found that there were discrepancies between level descriptors in the ACTFL OPI rating scale and test takers' actual performance on the test. It was suggested that the ACTFL level descriptors reflect the test takers' ability to extend their speech with interaction, not simply providing the length of speech to differentiate levels. This is a compelling study in terms of investigating the alignment between scoring criteria and the test taker's actual performance on the test. However, this study only looked at the test takers' performance on the test, not their actual performance in real-life settings.

This limitation can be addressed in the study on the extrapolation inference (i.e., evaluation of the predicted performance in real-life settings in Kane's (2006, 2013) and Chapelle et al.'s (2008) interpretive arguments). Moving beyond correlation between a test and the existing tests that measure the same constructs, Brooks and Swain (2014) investigated the extent to which test tasks represent the actual tasks in the target domain, directly collecting test takers' (i.e., international graduate students) actual language use. International graduate students' recordings of in-class and out-of-class speaking activities were compared to the students' performance on the TOEFL iBT speaking tasks, and the students' perceptions of the speaking test tasks and their speaking in their academic studies were examined. Similarities and differences between performance in the speaking test and in-class were found, as the researchers provided more defensible validity evidence in terms of evaluation of predictability of test scores in real-life settings beyond the correlation studies.

Rather than comparing the performance in the test and the actual performance in real-life settings, Liu (2014) compared test tasks and scores in two tests. A task-based reading test, based on can-do statements in the Common European Framework of Reference (CEFR), was developed. The tasks and scores of the test were then compared to those of General English Proficiency Test (GEPT). Drawing on Bachman and Palmer's (2010) AUA, experts' judgments were used for generalizability and confirmatory factor analysis using test scores of college students for meaningfulness. It was reported that in comparison of the GEPT reading test to the task-based reading test, the GEPT reading tasks did not match the target language use tasks because of inadequate sampling of the target domain, the GEPT scores were not generalizable to the TLU performance, and the test scores did not provide sufficient information about language abilities. However, this study only used test scores to investigate the constructs of GEPT and experts' judgments on GEPT tasks. This is not much different from the traditional approaches to validation in terms of limitations in the content validity and criterion-related validity. Rather, Brooks and Swain (2014) study on extrapolation is meaningful in terms of including the actual test-takers' performance in real-life settings into test validation.

In relation to defining test constructs to the evaluation of the extrapolation inference, Chapelle, Chung, et al. (2010), drawing on Chapelle et al.'s (2008) interpretive argument, provided validation processes and corresponding evidence in the test development of an academic grammar test designed by the Iowa State University. The researchers reported the following: test items were developed through identifying grammatical features and relevant test tasks by reviewing previous second language acquisition studies (the domain description inference); scoring rubrics were developed and revised based on analysis of test scores (the evaluation inference); generally test scores were reliable (the generalization inference); there were differences among groups at different proficient levels (the explanation inference); and there were positive, moderate correlations between the test and the TOEFL scores (the extrapolation inference).

However, theoretical constructs of language proficiency and corresponding test tasks to the constructs may call for feedback of stakeholders who are actual language users in the target domain as framed in Weir's (2005) socio-cognitive model. Bachman and Palmer (1996) suggest that test developers and testing researchers collect feedback on operational versions of tests from key test stakeholders (e.g., test-takers, raters, and others). For example, Brown (1993) discusses test takers' feedback on test items in an

oral proficiency test of Japanese for the tourism and hospitality. Because the test takers had relevant work experience with the test domain (the tourism and hospitality), their feedback on the test item was collected before the official implementation of the test.

Recently, scholars in language testing (e.g., Bachman & Palmer, 1996, 2010; Cheng & DeLuca, 2011; DeLuca, Cheng, Fox, Doe, & Li, 2013; Fox & Cheng, 2007, 2015; Hamp-Lyons, 2000; Moss, Girard, & Haniford, 2006; Shohamy, 2001) have advocated collecting stakeholders' perspectives for validation. The AUA (Bachman & Palmer, 2010) requires validation for both intended and actual score meaning and score use. Furthermore, involvement of multiple stakeholders into test validation is crucial for the validity and validation studies (Moss, 1996; Moss et al., 2006). As the contemporary validity theorists such as Kane (2013) and Messick (1989) noted, validity changes over time and context-bounded. This definition of the validity requires ongoing validation studies and involving relevant test stakeholders to investigate contextual factors (social, political, or cultural factors in a particular context) that may affect test stakeholders' understanding and uses of test scores. In addition, as different stakeholders may interpret and use test scores in different ways (Koch & DeLuca, 2012; Macqueen et al., 2016; Zumbo, 2015), validation studies need to include multiple stakeholders' perspectives to support or challenge the intended score meaning and use in particular contexts. Involvement of multiple stakeholders will strengthen validity arguments by disclosing sources of evidence that may threaten the validity of score interpretation and use in score use contexts (Moss et al., 2006).

Few empirical studies that used an argument-based approach involved multiple stakeholders in test validation studies. Although, Kadir (2008) investigated the scoring, generalization, extrapolation, and decision-making inferences in the English Language Proficiency Assessment (ELPA) used for hiring in the Malaysian public service, drawing on Kane's (2006) interpretive argument. Interviews with the test administrators and officers as policy makers in the Public Service Department were used as well as a questionnaire regarding test takers' perceptions of English language use at the workplace, test-taking experience, and the impact of ELPA. It was reported that ELPA test administrators found the test useful for hiring purposes, despite some limitations about overall management of the test and certain aspects of scoring procedures. The stakeholders also had positive attitude toward potential impact of the test.

However, Kadir (2008) did not investigate the stakeholders' actual understanding of test scores and the social, political, and cultural factors that may affect their interpretations and uses of test scores, including unintended consequences, which are defined as side effects derived from intended uses of test scores (Shepard, 1997). As Kane (2002, 2006) noted, the criteria employed for evaluation of a decision depend on whether the decision based on interpretation of test scores results in positive impact, with a reasonable cost and few negative side effects. It is required to provide validity evidence for how much and what kind of unintended consequences are derived from the test use.

Sun (2016) also involved multiple stakeholders in a validation study using an argument-based approach. He conducted the most comprehensive study on the intended/actual meanings and uses of Chinese English Test-4 (CET-4) that was designed and used for enhancing English education for Chinese college students. The meaning of the test scores and use by relevant test stakeholders (the test developers, test users, and test takers) were investigated. Interviews were used with the test

developers to lay out the intended meanings and uses of the CET-4 scores as baseline evidence and with test users to investigate how they interpreted the scores and how and why they used the scores in their contexts. A questionnaire was also used to investigate university students' perceptions of the effects of the CET-4 on their English learning. From interviews with the test users, the CET-4 scores were used differently from uses intended by the test developers as the scores were used for hiring purposes in business and for controlling access to social resources and for gatekeeping to government positions, other than for enhancing English education in educational contexts. The test use in educational contexts was consistent with intended purpose of the CET-4, while the test use in business and government contexts resulted from test users' extended interpretation of test scores regarding English proficiency.

It was also found that the score use was intertwined with social values or contextual factors in the Chinese context as the test scores were perceived as an indicator of the student's efforts rather than their English proficiency and as an association with the student's quality in terms of meeting basic requirements in university education. However, Sun (2016) did not investigate the consequences in business contexts as his research focus was the evaluation of the consequences in educational contexts. From the questionnaire responses, it was found that students' learning was consistent with the CET-4's intended outcomes in terms of helping students engage in test preparation, measuring intended constructs, promoting students' achievement, and motivating students in their learning. On the other hand, unintended consequences of test use in the educational context were reported. Students perceived the test scores were instrumentally used for high-stakes decisions, which was not included in the intended purposes of the CET-4. This is related to social factors that affect students' perceptions of test importance and test taking purposes (Bachman, 2007; Fox & Cheng, 2007; Ryan, 2002). These social factors (i.e., the score use for high-stakes decisions) exerted more effects on students' test preparation practices, leading to cramming strategy. Furthermore, students' test preparation did not significantly predict their performance in the CET-4. With these unintended consequences, the test use for enhancing English education in China was not supported by findings regarding consequences of the CET-4.

Sun's (2016) study is limited in terms of a lack of further investigation of test constructs regarding English proficiency that stakeholders believe are important in their contexts, although his research focus was the washback of the CET-4. This could be achieved by collecting feedback from stakeholders who are actual language users in the real-life settings, e.g., university students in the academic settings.

Discussion

In this discussion section, five points are discussed based on the review of traditional and contemporary validation framework and eight empirical validation studies: (a) defining test constructs and relevant tasks through domain analysis (Chapelle et al., 2008; Mislevy et al., 2003), (b) involving multiple stakeholders (Moss et al., 2006), (c) investigating the test developers' intended and test users' actual meanings and uses of test scores (Bachman & Palmer, 2010), (d) value implications (Im & McNamara, 2017), and (e) employing multiple or mixed methods beyond statistical analyses (e.g., Kadir, 2008; Sun, 2016) for validation. These five points may be critical to support the intended meanings and uses of test scores to bring about positive consequences in the score use contexts.

Conducting (a) the domain analysis is the most important aspect for test design to identify the language knowledge, skills, and abilities valued in the language use situations and the relevant tasks to elicit them. From the review of validation models in this paper, it was found that two validation models (i.e., Chapelle et al., 2008; Mislevy et al., 2003) explicitly specify the inferences for the domain analysis, although other validation models imply the importance of domain analysis. Without specifying the explicit inference for the domain analysis as an initial step before the test design, validation may be carried out only to prove the usefulness of a test in terms of accurately measuring what the researchers defined.

In language testing, there is not a consensus model for language proficiency that explains performance in real-life settings (Fulcher, 2015) because theoretical conceptualizations of language proficiency (e.g., Bachman, 1990, 2010; Bachman & Palmer, 1996, 2010; Canale, 1983; Canale & Swain, 1980; Chomsky, 1965; Hymes, 1972) have not been clearly defined nor empirically supported. Due to the difficulties in conceptualizing language proficiency and accordingly validating intended score meaning and use in language testing, language testers are drawn to Kane's (2006) interpretive argument in terms of test validation for "context-laden meaning of language and prediction to a specific domain" (Fulcher, 2015, p. 109). Kane's (1992, 2006, 2013) interpretive argument and IUA, however, may lead to narrower definitions of test constructs, limited to only a particular domain.

Faced with a lack of a strong theory for language proficiency, specifying the inference for the domain analysis in a validation model may help test designers and validation researchers focus more on defining theoretical constructs and evaluating the definitions to more accurately explain and predict the test taker's performance in the language use contexts based on test scores (Fulcher, 2015). However, Elder, McNamara, Kim, Pill, and Sato (2017) point out that test constructs in language testing have been defined by language specialists, and Elder and McNamara (2016), Jacoby and McNamara (1999), and H. Kim and Elder (2009) commonly call for listening to actual language users who have insights into the language use situations for defining language proficiency in specific-purpose language testing. The justification for this perspective is that language users' perspectives reflect actual language proficiency valued in the real-world. During the domain analysis, researchers can identify the language knowledge, skills, and abilities and their relevant tasks through working with actual language users in the target domain of interest.

Along with the importance of the domain analysis, the second point is (b) to involve multiple stakeholders into validation. As discussed in the previous section, actual language users in the target domain can be involved for evaluating test items during the validation such as Brown's (1993) study. In addition, including multiple stakeholders' perspectives into validation has been advocated by scholars in language testing (e.g., Bachman & Palmer, 1996, 2010; Cheng & DeLuca, 2011; DeLuca et al., 2013; Fox & Cheng, 2007, 2015; Hamp-Lyons, 2000; Moss et al., 2006; Shohamy, 2001).

However, validation research has mainly evaluated test-takers' performance on a test (e.g., Chapelle, Chung, et al., 2010; Lim, 2009; Liu, 2014; Tominaga, 2014). Collecting validity evidence using their performance is the necessary condition to support the intended interpretations of test scores. This, however, is not satisfactory for the validity of interpretations of the scores because of the variability of interpretations in a given context. Multiple stakeholders need to be involved in validation because different

stakeholders interpret and use test scores in different ways (Koch & DeLuca, 2012; Macqueen et al., 2016; Zumbo, 2015). In addition, validation models should provide a guideline to collect validity evidence from multiple stakeholders to detect any potential factors that may threaten the validity of interpretations and uses of test scores in a particular context (Moss et al., 2006).

Accompanying multiple stakeholders is (c) to investigate the test developers' intended and test users' actual meanings and uses of test scores, by juxtaposing them. As we reviewed the traditional and contemporary views on validation, only Bachman and Palmer's (2010) AUA explicitly calls for investigating the intended and actual interpretations, uses and consequences. Bachman and Palmer (2010) focus more on test use by adopting Messick's (1989) construct validity in terms of value implications and consequences. As the contemporary validity theorists such as Kane (2013) and Messick (1989) noted, validity changes over time and is context-bounded. This definition of validity requires ongoing validation studies and involving relevant test stakeholders to investigate contextual factors (social, political, or cultural factors in a particular context) that may affect test stakeholders' understanding and uses of test scores.

Nevertheless, the variations of an argument-based approach do not fully address (d) the value implications. While construct, content, and criterion-referenced validities were the main foci in one question and three validities approach in terms of validating psychometric properties of a test, contemporary views on validation since Messick's (1989) seminal work in validity have included social aspects of testing such as the value implications and social consequences. Some studies such as Ginther and Elder (2014), Im and McNamara (2017), and O'Loughlin (2011, 2013) investigated main interpreters' understanding and uses of test scores in terms of how they perceive test scores and how they use the scores. From these investigations, the value implications of the test users were found, and their values affected their uses of test scores. The value implications are no longer separated from validity, consistent with Messick's (1989) views on the interpretive approach to validity. Although some researchers (e.g., Mehrens, 1997; Popham, 1997) argue that it is too complex to discuss social aspects of testing in test validation, it is necessary to look at main interpreters' value implications if validation researchers aim to validate the interpretations and uses of test scores. Testing is social practices (Cronbach, 1988; McNamara & Roever, 2006). As seen in Im and McNamara (2017), admission officers' views on TOEIC LR scores related to the use of the scores for university admissions because of the social pressures such as lower test fees and higher testing site availability of the TOEIC LR compared to other English tests.

This calls for (e) using either multiple methods or mixed methods beyond quantitative inquiries in the test validation. Using different methods would be required to investigate in-depth understanding and uses of test scores in a given context. For example, discourse analyses can be used to identify language use between interlocutors by investigating their interactions in specific and cultural contexts (Abbuhl & Mackey, 2008). Content analysis using documents published by test developers and interviews with them can be used to investigate the intended interpretations and uses of test scores. Lee and Greene (2007) pointed out that different methods are required to understand various dimensions of evidence (i.e., consistencies and variations of perspectives) collected from multiple stakeholders when investigating the validity of interpretations and uses of test scores in a given context.

This perspective in a validation inquiry advocates pluralism to evaluate the validity in terms of viewing the validity as the social practices from multiple worldviews (Jang, Wagner, & Park, 2014). Using a variety of methods for validation, either multiple or mixed methods, would provide more in-depth understanding of interpretations and uses of test scores in the score use contexts like Kadir (2008) and Sun (2016) reviewed in this paper.

The five points discussed in this paper are not new, and have been discussed for the past decades. However, what we suggest is that validation models may need to provide guidelines for investigating social aspects of testing and involving multiple stakeholders into validation if the validation theorists follow the validity theories of Cronbach (1988) and Messick (1989). Accordingly, empirical validation studies should involve multiple stakeholders into validation and use different methods to disclose any potential sources of invalidity of interpretations and uses of test scores in a given context.

Conclusions

This paper not only illustrated the traditional and contemporary approaches to validation but also critically reviewed empirical studies that used an argument-based approach to validation, mostly in the high-stakes contexts. Early views on validation focused on evaluating psychometric properties of a test (i.e., construct, content, and criterion-related validities) and have limitations in terms of sufficiency of evidence for the meaning of test scores, their use, and their consequences in the one question and three validities as well as methodological or practicality issues in the evidence-gathering, test usefulness, and socio-cognitive models. On the other hand, the contemporary perspective on test validity and validation such as an argument-based approach to validation (Kane, 1992, 2006, 2013) goes beyond the traditional construct-focused evidence by offering validation researchers a systematic process for test validation in terms of how validation researchers structure validity arguments, linking validity evidence for the development and use of a test.

Considering this usefulness of an argument-based approach, we have focused on reviewing variations of the approach (e.g., Bachman & Palmer, 2010; Chapelle et al., 2008; Kane, 1992, 2006, 2013; Mislevy et al., 2003) and empirical studies which used the approach. Through the review, we suggest that various validity evidences be required for validity arguments to be defensible. Validation researchers need to consider (a) defining test constructs and relevant tasks through domain analysis (Chapelle et al., 2008; Mislevy et al., 2003); (b) involving multiple stakeholders (Moss et al., 2006); (c) investigating the intended and actual meaning, use, and consequences of testing (Bachman & Palmer, 2010); (d) investigating social, cultural, and political values embedded in test scores and test use; and (e) employing multiple or mixed methods beyond statistical analyses (e.g., Kadir, 2008; Sun, 2016).

In language testing, operationalizing test constructs and relevant test tasks to elicit the constructs through domain analysis is of great importance when designing a test which reflects the actual use of language in the target domain of interest. Domain analysis can be carried out by reviewing the existing literature (Chapelle, Chung, et al., 2010). Constructs of the language use, however, may vary from context to context. Therefore, there might not be a comprehensive theory that explains every language use. To address this issue of difficulty in defining test constructs, there may be a need to call

for feedback from the actual language users in the target domain of interest (Bachman & Palmer, 1996). Involving multiple stakeholders may also enable validation researchers to be able to disclose any potential invalidity of score interpretations and uses of a test, as different stakeholders may have different interpretations and uses of a test in a given context. Investigating the intended and actual score meaning, use, and consequence facilitate the evaluation of the validity by comparing the intended to the actual score meaning, use, and consequences. Indeed, the intended score meaning, use, and consequences are the baseline evidence, which would be a guideline for what to investigate for the subsequent investigation of the stakeholders' actual score meaning, use, and consequences of a test. When investigating stakeholders' actual score meaning, use, and consequences, value implications may be an important aspect in validation because their values affect their uses of test scores, and accordingly bring about unintended consequences. However, the majority of validation approaches we have reviewed in this paper do not fully provide validation procedures for investigating value implications. Value implications cannot be separate from test validation (Messick, 1989) as they shape the understanding (i.e., meaning) and use of test scores. There is an urgent need to address value implications in test validation with providing a conceptual validation framework and a methodological guideline, considering the sociopolitical aspects of testing. Lastly, ever since Messick's (1989) seminal work in validity, an interpretive approach to validation has been adopted by validation researchers. However, validation research still tended to rely on quantitative methods. Different stakeholders may have their own perspectives on test scores, and these perspectives can be investigated using either multiple or mixed methods for validation research.

Endnotes

[1]The ovals in Fig. 2 represent either data or claims. The squares represent types of inferences that link data to claims

[2]Assessing Writing, Language Testing in Asia, and Papers in Language Testing and Assessment

[3]The time period for this manuscript was between 1992 and 2016. The year 1992 was when Michael T. Kane first proposed an argument-based approach. There are more studies that used an argument-based approach from 2017 to 2019 such as Becker (2018), Ikeda (2018), Knoch and Chapelle (2018), and Mendoza and Knoch (2018), but these fall outside the scope of this paper.

Abbreviations

ACTFL: The American Council on the Teaching of Foreign Languages; APA, AERA & NCME: American Psychological Association, American Educational Research Association, & National Council on Measurement in Education; AUA: Assessment use argument; CEFR: The Common European Framework of Reference for Languages; CET-4: Chinese English Test-4; ECD: Evidence-centered design; ELPA: The English Language Proficiency Assessment; GEPT: The General English Proficiency Test; IELTS: The International English Language Testing System; IUA: Interpretation and use argument; MELAB: The Michigan Language Assessment Battery; OPI: The Oral Proficiency Interview; TEAP: The Test of English for Academic Purposes; TLU: The Target language use; TOEFL iBT: The Test of English as a Foreign Language internet-Based Test; TOEIC LR: The TOEIC Listening and Reading; TOEIC: The Test of English for International Communication

Funding

There was no funding for this manuscript.

Availability of data and materials

Not applicable.

Authors' contributions
All authors contributed to this review paper through collecting the relevant literature, writing, reviewing and revising the manuscript. Specifically, GHI collected the previous literature, synthesized it and made the first draft of the manuscript. DS revised the first draft and edited the discussion and conclusions sections and LC finally reviewed and revised the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Faculty of Education, Duncan McArthur Hall, 511 Union Street, Queen's University, Kingston, ON K7M 5R7, Canada.
[2]Department of English Language and Literature, Chung-Ang University, 84 Heukseok-ro, Seoul 06980, South Korea.

Received: 13 January 2019 Accepted: 10 June 2019
Published online: 10 August 2019

References

Abbuhl, R., & Mackey, A. (2008). Second language acquisition research methods. In K. A. King & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Research methods in language and education* (Vol. 10, pp. 1–13). Dordrecht: Springer.

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

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

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: American Psychological Association.

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

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1–15.

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

Bachman, L. F. (2003). Constructing an assessment use argument and supporting claims about test taker-assessment task interactions in evidence-centered assessment design. *Measurement: Interdisciplinary Research and Perspectives, 1*, 63–65. https://doi.org/10.1207/S15366359MEA0101_03

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

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

Bachman, L. F. (2007). What is the construct? The dialectic abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–72). Ottawa: University of Ottawa.

Bachman, L. F. (2010). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

Becker, A. (2018). Not to scale? An argument-based inquiry into the validity of an L2 writing rating scale. *Assessing Writing, 37*, 1–12. https://doi.org/10.1016/j.asw.2018.01.001

Brooks, L., & Swain, M. (2014). Contextualizing performances: comparing performances during TOEFL iBT TM and real-life academic speaking activities. *Language Assessment Quarterly, 11*, 353–373. https://doi.org/10.1080/15434303.2014.947532

Brown, A. (1993). The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10*, 277–301. https://doi.org/10.1177/026553229301000305

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–27). London: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47. https://doi.org/10.1093/applin/I.1.1

Carroll, B. (1980). Specifications for an English language testing service. In J. C. Alderson & A. Hughes (Eds.), *Issues in language testing. ELT documents 111* (pp. 66–110). London: The British Council.

Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing, 27*, 443–469. https://doi.org/10.1177/0265532210367633

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. London: Routledge.

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

Chapelle, C. A., & Voss, E. (2014). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1079–1097). Chichester: Wiley. https://doi.org/10.1002/9781118411360.wbcla110

Cheng, L., & DeLuca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational Assessment, 16*, 104–122. https://doi.org/10.1080/10627197.2011.584042
Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.
Crocker, L. (1997). Assessing content representativeness of performance assessment exercises. *Applied Measurement in Education, 10*, 83–95. https://doi.org/10.1207/s15324818ame1001_5
Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington: 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: Erlbaum.
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. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington: American Council on Education.
DeLuca, C. (2011). Interpretive validity theory: Mapping a methodology for validating educational assessments. *Educational Research, 53*, 303–320. https://doi.org/10.1080/00131881.2011.598659
DeLuca, C., Cheng, L., Fox, J., Doe, C., & Li, M. (2013). Putting testing researchers to the test: An exploratory study on the TOEFL iBT. *System, 41*, 663–676. https://doi.org/10.1016/j.system.2013.07.010
Elder, C. (1997). What does test bias have to do with fairness? *Language Testing, 14*, 261–277. https://doi.org/10.1177/026553229701400304
Elder, C., & McNamara, T. (2016). The hunt for "indigenous criteria" in assessing communication in the physiotherapy workplace. *Language Testing, 33*, 153–174. https://doi.org/10.1177/0265532215607398
Elder, C., McNamara, T., Kim, H., Pill, J., & Sato, T. (2017). Interrogating the construct of communicative competence in language assessment contexts: What the non-language specialist can tell us. *Language and Communication, 57*, 14–21. https://doi.org/10.1016/j.langcom.2016.12.005
Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197. https://doi.org/10.1037/0033-2909.93.1.179
Fox, J., & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test takers. *Assessment in Education: Principles, Policy and Practice, 14*, 9–26. https://doi.org/10.1080/09695940701272773
Fox, J., & Cheng, L. (2015). Walk a mile in my shoes: Stakeholder accounts of testing experience with a computer-administered test. *TESL Canada Journal, 32*, 65–86. https://doi.org/10.18806/tesl.v32i0.1218
Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. New York: Routledge.
Geranpayeh, A., & Taylor, L. (Eds.). (2013). *Examining listening: Research and practice in assessing second language listening*. Cambridge: Cambridge University Press.
Ginther, A., & Elder, C. (2014). *A comparative investigation into understandings and uses of the TOEFL iBT® Test, the International English Language Testing Service (Academic) test, and the Pearson Test of English for graduate admissions in the United States and Australia: A case study of two university contexts* (TOEFL iBT research report no. 24, ETS research report no. RR-14-44). Princeton: Educational Testing Service. https://doi.org/10.1002/ets2.12037
Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement, 1*, 1–10. https://doi.org/10.1177/014662167700100103
HajiPourNezhad, G. (2003). An approach to the validation of judgments in language testing. In T. Newfields, S. Yamashita, A. Howard, & C. Rinnert (Eds.), *Proceedings of the 2003 JALT Pan-SIG Conference held at Tokyo Keizai University on May 10–11, 2003* (pp. 80–84) Retrieved March 15, 2019 from http://jalt.org/pansig/2003/HTML/HajiPourNezhad.htm
Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System, 28*, 579–591. https://doi.org/10.1016/S0346-251X(00)00039-7
Hsu, H.-L. (2012). *The impact of world Englishes on language assessment: Rater attitude, rating behavior, and challenges (IELTS)*. (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses database. (3571158).
Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Middlesex: Penguin Books.
Im, G.-H., & McNamara, T. (2017). Legitimate or illegitimate uses of test scores in contexts unrelated to test purposes. *English Teaching, 72*, 71–99. https://doi.org/10.15858/engtea.72.2.201706.71
Ikeda, N. (2018). *Measuring L2 oral pragmatic abilities for use in social contexts: Development and validation of an assessment instrument for L2 pragmatic performance in university settings* (Unpublished doctoral thesis). The University of Melbourne, Melbourne, Australia.
Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18*, 213–241. https://doi.org/10.1016/S0889-4906(97)00053-7
Jang, E. E., Wagner, M., & Park, G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics, 34*, 123–153. https://doi.org/10.1017/S0267190514000063
Kadir, A. K. (2008). *Framing a validity argument for test use and impact: The Malaysian public service experience*. (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses database. (3337680).
Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535. https://doi.org/10.1037/0033-2909.112.3.527
Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x
Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*, 31–35. https://doi.org/10.1111/j.1745-3992.2002.tb00083.x
Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspective, 2*, 1351–1370. https://doi.org/10.1207/s15366359mea0203_1
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement (4th ed.)* (pp. 17–64). Washington: American Council on Education/Praeger.

Kane, M. T. (2011). Validating score interpretations and uses: Messick lecture Language Testing Research Colloquium, Cambridge April 2010. *Language Testing, 29*, 3–17. https://doi.org/10.1177/0265532211417210
Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. https://doi.org/10.1111/jedm.12000
Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x
Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.
Kim, H., & Elder, C. (2009). Understanding aviation English as a lingua franca: Perceptions of Korean aviation personnel. *Australian Review of Applied Linguistics, 32*, 23.1–23.17. https://doi.org/10.2104/aral0923
Kim, J. Y. (2008). *Development and validation of an ESL diagnostic reading-to-write test: An effect-driven approach*. (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses database. (3337823).
Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*, 477–499. https://doi.org/10.1177/0265532217710049
Koch, M. J., & DeLuca, C. (2012). Rethinking validation in complex high-stakes assessment contexts. *Assessment in Education: Principles, Policy and Practice, 19*, 99–116.
Lado, R. (1961). *Language testing*. London: Longman.
Lee, Y., & Greene, J. (2007). The predictive validity of an ESL placement test. *Journal of Mixed Methods Research, 1*, 366–389. https://doi.org/10.1177/1558689807306148
Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing, 17*, 43–64. https://doi.org/10.1177/026553220001700102
Lim, G. S. (2009). *Prompt and rater effects in second language writing performance assessment*. (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses database. (3392954).
Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*, 14–16. https://doi.org/10.1111/j.1745-3992.1997.tb00587.x
Liu, H.-m. (2014). *Investigating the relationships between a reading test and can-do statements of performance on reading tasks*. (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses database. (3607916).
Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, Monograph Supplement(9), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635
Macqueen, S., Pill, J., & Knoch, U. (2016). Language test as boundary object: Perspectives from test users in the healthcare domain. *Language Testing, 33*, 271–288. https://doi.org/10.1177/0265532215607401
McArthur, T. (2001). World English and world Englishes: Trends, tensions, varieties, and standards. *Language Teaching, 34*, 1–20. https://doi.org/10.1017/S0261444800016062
McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. London: Blackwell Publishing.
Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practices, 16*, 16–18. https://doi.org/10.1111/j.1745-3992.1997.tb00588.x
Mendoza, A., & Knoch, U. (2018). Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing, 35*, 41–55. https://doi.org/10.1016/j.asw.2017.12.003
Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale: Lawrence Erlbaum.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education & Macmillan.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749. https://doi.org/10.1037/0003-066X.50.9.741
Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241–256.
Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of assessment arguments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62. https://doi.org/10.1207/S15366359MEA0101_02
Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher, 25*(20–28), 43. https://doi.org/10.3102/0013189X025001020
Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109–162. Retrieved from http://www.jstor.org/stable/4129771
Nakatsuhara, F. (2014). *A research report on the development of the Test of English for Academic Purposes (TEAP) speaking test for Japanese university entrants—Study 1 & Study 2*. Retrieved from https://www.eiken.or.jp/teap/group/pdf/teap_speaking_report1.pdf
O'Loughlin, K. (2011). The interpretation and use of proficiency test scores in university selection: How valid and ethical are they? *Language Assessment Quarterly, 8*, 146–160. https://doi.org/10.1080/15434303.2011.564698
O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing, 30*, 363–380. https://doi.org/10.1177/0265532213480336
O'Sullivan, B. (2012). *Aptis test development approach (ATR-1)*. Retrieved from British Council website: https://www.britishcouncil.org/sites/default/files/aptis-test-dev-approach-report.pdf
O'Sullivan, B., & Weir, C. (2011). Language testing and validation. In B. O'Sullivan (Ed.), *Language testing: Theory and practice* (pp. 13–32). Oxford: Palgrave.
Popham, W. J. (1997). Consequential validity: Right concern—Wrong concept. *Educational Measurement: Issues and Practices, 16*, 9–13. https://doi.org/10.1111/j.1745-3992.1997.tb00586.x
Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice, 21*, 7–15. https://doi.org/10.1111/j.1745-3992.2002.tb00080.x
Schmidgall, J. E. (2017). *Articulating and evaluating validity arguments for the TOEIC® tests* (ETS Research Report 17-51). Retrieved from https://files.eric.ed.gov/fulltext/EJ1168723.pdf
Shaw, S., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450. Retrieved from http://www.jstor.org/stable/1167347

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(5–8), 13, 24. https://doi.org/10.1111/j.1745-3992.1997.tb00585.x

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*, 4–14. https://doi.org/10.3102/0013189X029007004

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

Shulha, L., & Wilson, R. (2009). Rethinking large-scale assessment. *Assessment Matters, 1*, 111–134.

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research, 51*, 161–179. https://doi.org/10.1080/00131880902891305

Sun, Y. (2016). *Context, construct, and consequences: Washback of the college English test in China*. Retrieved from ProQuest Dissertations & Theses database. (10155303).

Taylor, L. (Ed.). (2012). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge: Cambridge University Press.

Taylor, L. (2014). *A report on the review of test specifications for the reading and listening papers of the Test of English for Academic Purposes (TEAP) for Japanese university entrants*. Retrieved from https://www.eiken.or.jp/teap/group/pdf/teap_rlspecreview_report.pdf

Tominaga, W. (2014). *Validating the scoring inference of the Japanese OPI ratings: The use of extended turns, connective expressions, and discourse organization*. (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses database. (3648599).

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

Toulmin, S. (2003). *The uses of argument* (Updated ed.). Cambridge: Cambridge University Press.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

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

Weir, C. J. (2014). *A research report on the development of the Test of English for Academic Purposes (TEAP) writing test for Japanese university entrants*. Retrieved from https://www.eiken.or.jp/teap/group/pdf/teap_writing_report.pdf

Wu, R. Y.-F. (2011). *Establishing the validity of the General English Proficiency Test reading component through a critical evaluation on alignment with the Common European Framework of Reference* (Unpublished doctoral thesis). University of Bedfordshire, Bedfordshire.

Zumbo, B. D. (2015, November 5). *Consequences, side effects and the ecology of testing: Keys to considering assessment 'in vivo'*. Keynote address, the annual meeting of the Association for Educational Assessment–Europe (AEA–Europe), Glasgow, Scotland.