Language Teaching Research Quarterly **EUROKD**

2025 (51), 273–283
https://doi.org/10.32038/ltrq.2025.51.05

How Many Language Testing Publications Use "the Unqualified Phrase 'the Validity of the Test'"?

Haeun Kim*
School of Languages and Linguistics, The University of Melbourne, Australia
Shireen Baghestani
Department of English, Iowa State University, United States

Correspondence
Email: haeun.hannah.kim@unimelb.edu.au

Abstract
The *Standards for Educational and Psychological Testing* (AERA et al., 2014) states that "It is incorrect to use the unqualified phrase 'the validity of the test'" (p. 11). Although the *Standards* clearly states that it is incorrect to use the phrase "validity of the test" because "it is the interpretations of test scores for proposed uses that are evaluated, not the test itself" (p. 11), many authors still use this terminology. This study examines how frequently this occurs, why this may occur, and how to interpret this phenomenon. First, examination of articles published in *Language Testing* and *Language Assessment Quarterly* between 2011–2022 resulted in 233 articles being identified as including the expression "validity of + test" at least once. Next, the context around the occurrences of "use(s)" and "interpretation(s)" within these articles was analyzed to determine whether the author(s) referred to test interpretation and use. This was interpreted as evidence that the authors were familiar with the *Standards*' definition of validity, even though they used language that contradicted the *Standards'* guidelines. This study sheds light on the extent to which authors adhere to the *Standards*' guidelines and potential factors contributing to deviations from the recommended terminology.

ARTICLE HISTORY
Received: 01 July 2024
Revised: 17 October 2025
Accepted: 16 November 2025

KEYWORDS
Validity, Language Testing, Standards for Educational and Psychological Testing

How to cite this article (APA 7th Edition):
Kim, H., & Baghestani, S. (2025). How many language testing publications use "the unqualified phrase 'the validity of the test'"? *Language Teaching Research Quarterly*, *51*, 273–283. https://doi.org/10.32038/ltrq.2025.51.05

[1]Introduction

As a doctoral student in Professor Carol Chapelle's seminar course on validation at Iowa State University, one of the first things we read was Messick's (1989) seminal chapter on

[1] This paper is part of a special issue (2025, 50-51) entitled: In honour of Carol A. Chapelle's contributions to language assessment and learning (edited by Christine Coombe, Tony Clark, and Hassan Mohebbi).

validity. We learned the fundamental lesson that (1) it is not the test instrument itself but the interpretation and use of test scores that are validated, (2) validity is a matter of degree which is dependent on the evidence accumulated for or against the proposed interpretation or use, and (3) validity is a unitary concept. Drawing on Messick's (1989) definition of 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) and the argument-based approach to validation (Kane, 2013; Chapelle, 2021), students of Professor Carol Chapelle have come to understand validity as a context-dependent concept that very much depends on how a test is used and the evidence (e.g., reliability, content relevance) that supports its intended uses.

Not long after, we started to notice that scholars in the field of language testing often use phrases that seem to contradict this understanding of validity. For example, expressions such as the "the validity of the test" and "test validity" suggest that validity is an inherent characteristic of the test itself, rather than a quality connected to the interpretations and uses of its scores. Additionally, phrases such as "reliability and validity" seem to challenge the notion that validity is a unitary concept, as these two terms are presented in parallel—as if they are separate concepts—when, in fact, reliability is a type of empirical evidence that is necessary to support claims about test score interpretations for particular uses.

Knowing that the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014; hereinafter referred to as the *Standards*) goes as far as to say that "it is incorrect to use the unqualified phrase 'the validity of the test'" (p. 11), we became interested in how frequently language testing researchers use such language in their writing and why this may occur. Some might argue that the use of these phrases simply stems from a lack of awareness regarding contemporary validity concepts in educational measurement, suggesting a need for improved language assessment literacy among professionals. However, other factors such as communicative efficiency and the influence of various competing frameworks could also explain the variation in expressions that language testing scholars use.

Therefore, in conducting this study, focus was given to not only finding the frequency of occurrences for expressions such as "the validity of the test" but also analyzing the contexts in which such expressions appear to see how often authors use such expressions even when they seem to be aware of contemporary validity concepts.

Using corpus linguistics methods, this study examines the following research questions:

RQ$_1$: How many language testing articles published between 2011 and 2022 ascribe the attribute of validity to tests?

RQ$_2$: Among these instances, how many articles demonstrate authors' awareness that validity should be discussed in terms of test score interpretation and use?

Literature Review

Messick's (1989) definition of validity, as reflected in the *Standards*, is widely accepted among language testing scholars. However, it does not represent a universal consensus. There are other ways of conceptualizing validity, which may have gathered more attention in the past, that some scholars continue to reference. Additionally, the *Standards* largely reflects a North American perspective on validity, as it was written by three major organizations in the United States: the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). This may be why scholars based in Europe or other parts of the world sometimes turn to alternative frameworks of validity.

Among the many different ways of defining validity, the one that many language testers probably have heard of would be Cureton's (1951) which characterizes validity as indicating "how well the test serves the purpose for which it is used" (p. 621). At this time, validity was viewed as a characteristic of a test (Chapelle, 2021) that has two aspects—relevance and reliability. Cureton (1951) states, "To be valid—that is, to serve its purpose adequately—a test must measure something with reasonably high reliability, and that something must be fairly closely related to the function it is used to perform" (Cureton, 1951, p. 622). The main method used to answer questions related to validity was to estimate the correlation between test scores and criterion scores. Validity was seen as a property of the test that could be statistically measured. Therefore, it is not difficult to find phrases such as "the validity of the test" in publications around this time period.

Another way of conceptualizing validity which was widely adopted in the 1950s and 1960s is the multiple validities approach. The 1996 edition of the *Standards* reflects this view, distinguishing different types of validity, such as criterion-related validity (i.e., concurrent validity and predictive validity), content validity, and construct validity. However, as educational measurement scholars increasingly focused on the construct validity model (Cronbach & Meehl, 1955; Cronbach, 1971), concepts like content validity and criterion-related validity were reframed as types of evidence that support the interpretations and uses of test scores. This unitary view of validity was further elaborated by Messick (1989) in the third edition of *Educational Measurement*, and this definition continues to be used by modern day scholars in language testing. It was also at this time when people's conceptualization of what gets validated shifted from being the test itself to the interpretations and uses of test scores.

Considering how the definition of validity in educational measurement has evolved over the last century, it is understandable why scholars would express their concerns over those who continue to attribute the property of validity to tests. Sireci (2009), for instance, argues that "to claim that validity refers simply to demonstrating that a 'test measures what it purports to measure' or that it is an inherent property of a test is to ignore at least 70 years of research on validity theory and test validation as well as the

consensus *Technical Recommendations* and *Standards* that have existed since 1954" (p. 28). The 2014 edition of the *Standards* reinforces this perspective, explicitly stating that referring to "the validity of the test" is incorrect and unqualified. Nevertheless, the reality is these traditional conceptualizations of validity, which Chapelle et al. (2024) calls the "one question, three validities approach" (p. 68) is still persistent among nonspecialists, and it is not uncommon to see language testing scholars use phrases such as "the validity of the test" or "test validity" in academic publications.

Weir's (2005) framework of validity may also contribute to the use of the phrase by scholars. Although he ascribes to a unitary definition of validity, he also distinguishes between different types of validity (e.g., scoring validity, consequential validity, context validity). The expression "validity of + test" also appears frequently in his 2005 book. To this day, many scholars, especially in the UK, still reference this framework. This may be one of the things driving the continued use of the phrase "validity of + test."

Methods

Data Collection

While there are several expressions similar to "the validity of the test" such as "validity of + assessment," "validation of + test/assessment," "validate + test/assessment," "test validity," or "valid test," in this study, we focus on the phrase "validity of + test," which the *Standards* (AERA et al., 2014) explicitly states as being "incorrect" and "unqualified" (p. 11). To identify articles including this phrase, research articles published between 2011 and 2022 in *Language Testing (LT)* and *Language Assessment Quarterly (LAQ)* were searched using the term "validity," which resulted in 470 articles. The search term "validity" was used, rather than "validity of," to capture the uses of combined noun phrases with the preposition *of* such as the "validity and/or reliability of the test."

The PDF files of these research articles were converted to text files with AntFileConverter 1.2.1 (Anthony, 2017), and a custom Python script was used to clean the data. Reference lists and appendices following the references section were removed, and words split by hyphens at line breaks were connected. With the cleaned data, AntConc 4.3.1 (Anthony, 2024), a concordancing software, was used to find 3,646 hits of the word "validity" within 449 research articles—233 from *LT* and 216 from *LAQ.* Then, the Keyword in Context (KWIC) results from the AntConc were saved as a text file and imported into a Google Sheets spreadsheet for subsequent analysis.

Data Analysis

First, to identify instances matching the phrase "validity of + test," each KWIC line containing the word "validity" was manually reviewed by the researchers. As illustrated in Figure 1, test names such as TOEFL and OSSLT, which contain "test" in their non-abbreviated forms, were also included when they followed the phrase "validity of." Given the clear-cut nature of this method, the researchers divided the data into two (1,823 hits

each) and reviewed each KWIC line to determine whether the word "validity" and the preposition "of" was used to modify the noun "test" (i.e., whether the head noun of the prepositional complement is "test"). The distribution of the occurrences was also analyzed by journal and year to see if there were any notable trends.

Figure 1

Identifying the Instances of "Validity of + Test"

	A	B	C	F	H
1	Left Context	Hit	Right Context	Expression	File
2	l test and three other paired tests as a threat to the	validity	of group oral test and LANGUAGE ASSESSMENT	validity of + test	LAQ_00001.pdf
18	e confidence in the meaningfulness, reliability, and	validity	of several of the aviation language tests currently a	validity of + test	LAQ_00002.pdf
19	ce can be had in the meaningfulness, reliability and	validity	of several of the aviation language tests currently a	validity of + test	LAQ_00002.pdf
29	t performance] suggests problems of reliability and	validity	of test itself," (Ahmed, 2016, para. 6). Hidden syllal	validity of + test	LAQ_00004.pdf
30	listening and speaking skills raises the question of	validity	of the tests when curricular goals are taken into ac	validity of + test	LAQ_00004.pdf
44	riptions could serve as one basis for defending the	validity	of the test. Validity is traditionally defined as the de	validity of + test	LAQ_00023.pdf
83	e there is indeed some score-based support for the	validity	of TOEFL iBTTM as a measure of speaking ability	validity of + test	LAQ_00024.pdf
107	a solid link with the domain is a prerequisite for the	validity	of an LSP test, and such links need to be made ex	validity of + test	LAQ_00025.pdf
127	ests. Together with similar studies of the predictive	validity	of CEFR-based entrance tests, the study contribut	validity of + test	LAQ_00032.pdf
128	ntributes to the body of research into the predictive	validity	of language entrance tests, adding to this research	validity of + test	LAQ_00032.pdf
129	and Bridgeman (2012) investigated the predictive	validity	of the TOEFL-Internet Based Test (TOEFL iBT¬Æ)	validity of + test	LAQ_00032.pdf
130	Storch, and Lynch (1999) compared the predictive	validity	of TOEFL and IELTS and found positive correlatior	validity of + test	LAQ_00032.pdf
131	this volume). To date, no studies into the predictive	validity	of university entrance requirements or entrance tes	validity of + test	LAQ_00032.pdf
132	nto consideration when investigating the predictive	validity	of university entrance tests (Wang, Choi, Schmidga	validity of + test	LAQ_00032.pdf
155	evidence challenging the inferences underlying the	validity	of the OSSLT for L2 students. In Table 1, major fin	validity of + test	LAQ_00041.pdf
156	NTS 57 EQAO provides evidence for the construct	validity	of the OSSLT from two perspectives: (a) construct	validity of + test	LAQ_00041.pdf

Next, the KWIC lines for the hits "use(s)" and "interpretation(s)" within the articles where the phrase "validity of + test" was identified were examined qualitatively to determine whether the authors discussed validity with regard to test score interpretation and use in other parts of the article. This was interpreted as evidence that the authors were familiar with Messick's (1989) definition of validity, even though they used language that contradicted the *Standards* guidelines. For instance, as illustrated in Figure 2, LAQ_00135 (i.e., Li & Suen, 2012) was one of the articles where we found the authors using the expression "validity of + tests." However, the analysis of KWIC lines for "use(s)" and "interpretation(s)" revealed the authors referring to validity as a "concept that involves an overall evaluation of the evidence for the proposed interpretations and uses of test scores" (Li & Suen, 2012, p. 295). At the end of this line, the authors even cited the works of Kane (2006) and Messick (1989), which clearly shows their awareness of the Messick's definition of validity.

Due to the more interpretive nature of this analysis, we recognized the importance of ensuring inter-coder reliability before individually coding the KWIC lines for the hits "use(s)" and "interpretation(s)." Both researchers independently coded 69 out of the 119 articles, and reached a high percentage agreement of 86.7%. The coding of a particular article was considered to be in agreement when both researchers either coded at least one KWIC line as 'Y' or neither did. Disagreements were resolved through discussion. The

remaining 50 articles were then divided and coded individually, with each researcher coding 25 articles. The left and right context size was limited to 20 tokens, so the researchers consulted the original text file of the research articles when necessary (e.g., when the meaning was unclear because the beginning or end of the sentence was cut off).

Figure 2

Identifying Author Awareness of Messick's (1989) Definition of Validity

	A	B	C	D	E	H
1	Left Context	Hit	Right Context	Expression	Reference to score interpretation and use having the property of validity	File
946	ritten in English. The fairness and	validity	of large-scale tests used for ELLs is thus of great c	validity of + test	N	LAQ_00135.pdf
947	thesis" (Zuriff, 2000) to justify the	use	of test accommodations. The hypothesis states tha		N/A	LAQ_00135.pdf
948	the proposed interpretations and	uses	of test scores (Kane, 2006; Messick, 1989). Fairne		Y	LAQ_00135.pdf
949	English Language Learners The	use	of accommodations has been widely proposed as a		N/A	LAQ_00135.pdf
950	-based recommendations for the	use	of accommodations in large-scale assessments. P		N/A	LAQ_00135.pdf
951	ontexts, for example, whether the	use	of accommodations is comparable across schools		N/A	LAQ_00135.pdf
952	ELLs and non-ELLs, whereas the	use	of dictionaries, extra time, and extra time with gloss		N/A	LAQ_00135.pdf
953	of the evidence for the proposed	interpretations	and uses of test scores (Kane, 2006; Messick, 198		Y	LAQ_00135.pdf
954	ast test accommodation literature	uses	the term validity and the close relationship betwee		N/A	LAQ_00135.pdf
955	erms & Conditions of access and	use	can be found at https://www.tandfonline.com/actior	N/A	N/A	LAQ_00135.pdf
956	between fairness and validity, we	use	terms such as fairness or fairness and validity thro		N/A	LAQ_00135.pdf

Results

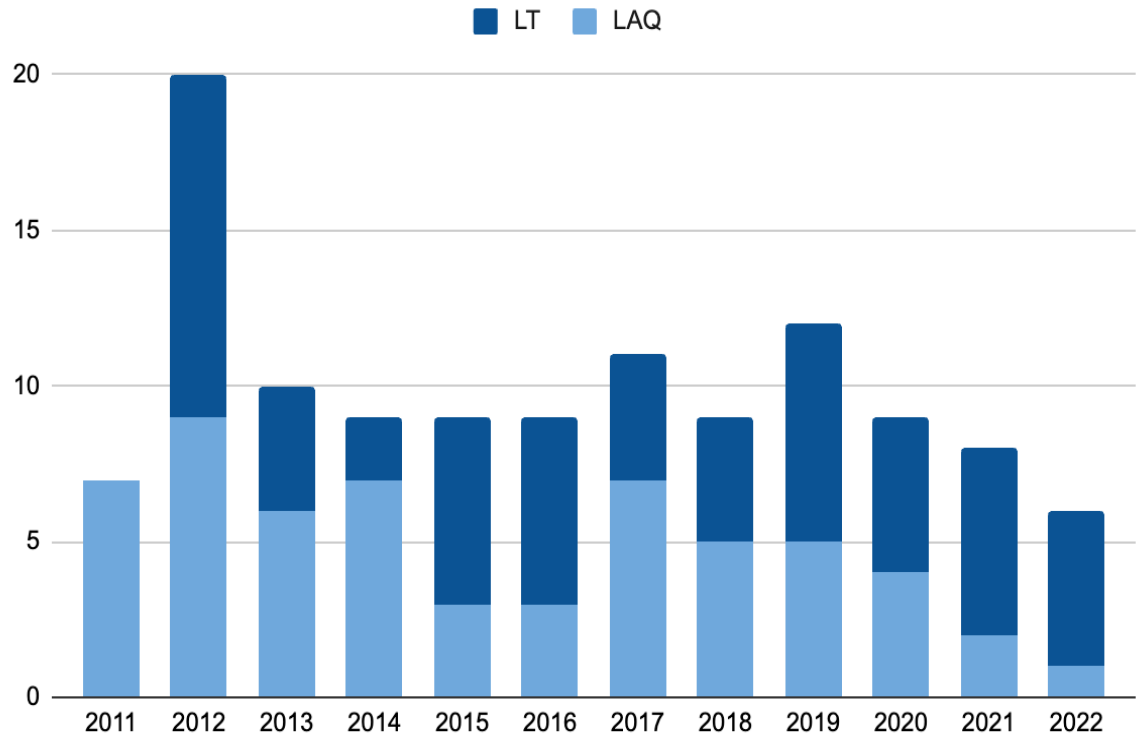
The first research question examined how many language testing articles published between 2011 and 2022 ascribe the attribute of validity to tests. To answer this question, we focused on identifying articles that included the phrase "validity of + test." The results showed that 119 articles (26.5%) out of the 449 research articles that had any reference to "validity" included the phrase "validity of + test." The proportion of articles that included this expression was slightly higher in *LAQ* (59 out of 216 articles = 27.3%) than in *LT* (60 out of 233 articles = 25.7%), but both were close to 25% which indicates that a substantial number of authors choose to use this expression. While there were eight authors whose name appeared twice as first author, the majority of the articles that used the expression "validity of + test" was written by different authors, which indicates that the findings were likely not skewed by the overrepresentation of a certain author's publication.

The distribution of articles by year from 2011 to 2022 is presented in Figure 3. One interesting point to note is that the number of articles including the phrase "validity of + test" was notably higher in 2012 compared to subsequent years, especially in the *LT* journal. Since this period is around the time the 2014 revision of the *Standards* was published, the change could be due to the editors' and reviewers' conscious effort to discourage the use of the phrase. However, the number of articles containing the phrase "validity of + test" in *LAQ* between 2011 and 2014 remained quite consistent between 2011 and 2014. It was only the number of articles in *LT* that changed quite drastically

between these years. This indicates that, in general, about a quarter of authors who discuss validity in their articles use the expression "validity of + test" each year.

Figure 3

Distribution of Articles Referencing 'Validity of + Test' in Language Testing (LT) and Language Assessment Quarterly (LAQ) from 2011 to 2022



The second research question looked at how many of these articles included expressions that suggest the authors' awareness of the *Standards*' statement that validity should be discussed in terms of test score interpretation and use. To answer this question, the KWIC lines for "use(s)" and "interpretation(s)" were examined. Among the 119 research articles that included the phrase "validity of + test," we found 32 articles (26.9%)—15 from *LT* and 17 from *LAQ*—mentioning score interpretation and use in the context of discussing validity. This shows that the authors' decision to use the phrase "validity of + test" should not be simply interpreted as the authors' ignorance of contemporary validity concepts. In Table 1, we present four articles (two from each journal) to exemplify our point.

Table 1

Examples of Research Articles Where the Authors' Awareness of Messick's (1989) Definition of Validity is Evident Even Though the Phrase "Validity of + Test" was Used

File ID (Authors)	Use of "Validity of + Test"	Reference to Score Interpretation and Use When Discussing Validity
LAQ_00148 (McNamara & Ryan, 2011)	*Recent discussions of the use of language tests in controversial social and political contexts have highlighted complex issues in the theory of the validity of tests.*	*The concern for fairness corresponds to the evidential basis of test score interpretation and test use, the upper row of Messick's famous validity matrix (Messick, 1989; see Table 1).*
LAQ_00197 (Timpe-Laughlin & Choi, 2017)	*This exploratory study responds to the call for more receptive, sociopragmatics assessment research by introducing and exploring the validity of a receptive pragmatics test: the American English Sociopragmatic Comprehension Test (AESCT).*	*Validity in assessment describes "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11).*
LT_00180 (Oller, 2012)	*Moreover, as demonstrated in many different empirical contexts the ecological validity of language tests and measures of discourse processing in general is enhanced by respecting narrative-like episodic organization and it is reduced by disrupting it (Oller, Chen, Oller, & Pan, 2005).*	*Let me begin by summing up what I think are the critical points of Kane's broad and practical approach to validating score interpretations and uses.*
LT_00248 (Yan et al., 2019)	*This study examines the construct validity of an ITA speaking test from a fairness perspective.*	*Fairness is concerned with standardization of testing procedures and consistency of test functionality (i.e., whether and to what extent score interpretations are equally valid for all examinee groups).*

Note. For each article, only one example of each occurrence is presented in this table.

Discussion

The study findings provide empirical evidence to support our initial observations regarding the prevalent use of the phrase "validity of + test" among scholars in the field of language testing, including authors, peer reviewers, and journal editors. At least one out of four articles published in *LT* and *LAQ* that referenced validity in any way included the phrase "validity of + test" which the *Standards* explicitly states as being incorrect and unqualified. Given that it has been some time since debates on validity were at their peak, one thought that came to our mind was that the number of scholars using this phrase may have increased over time, contributing to its overall frequency in the literature. However, the use of the phrase was found to be fairly consistent across the years.

Additionally, the analysis results of the KWIC lines for "use(s)" and "interpretation(s)" act as another counter evidence towards this idea, as we found that many of the authors who use the phrase "validity of + test" also included language which suggested that they were actually aware of Messick's (1989) definition of validity. As shown in Table 1, many of

these authors are highly regarded scholars in the field of language testing, which adds further support to our argument that the use of the phrase may not simply be due to the lack of knowledge or expertise on the topic.

Using the phrase "validity of + test" and showing knowledge of Messick's definition of validity suggests authors who are aware of the guidelines presented in the *Standards* may knowingly deviate from the suggested terminology when discussing validity. This may be due to a desire to conserve words, as saying "the validity of the interpretations and uses of test scores" is lengthy. Authors who use this shortened expression may believe it implies the full definition; in other words, saying a test is valid is to imply that the interpretations and uses of test scores for a particular purpose are appropriate. It is also possible that authors use this language out of habit without purposefully deviating from the *Standards*, although this is more likely to happen in impromptu speech than in writing or planned speech.

Another reason for the prevalence of the phrase "validity of + test" may be that the authors conceptualize validity using different frameworks from Messick's. For instance, a common pattern we noticed in the data was the tendency to see the phrase "validity of + test" preceded by an adjective (e.g., "ecological/predictive/concurrent/cognitive validity of the test"). This phrasing indicates that the authors ascribe to a multiple validities approach. Although the multiple validities approach is viewed as outdated (Chapelle et al., 2024; Sireci, 2009), our data show that many authors conceptualize validity in this way to the present day and that reviewers and editors of language testing journals are open to accepting language that reflects the multiple validities approach.

There may be many reasons, including the ones discussed in this section, why people use the expression "validity of + test" even though the *Standards* actively discourages it. In some cases, authors who use this expression may not be aware of the *Standards* or Messick's definition of validity. However, our study findings show that authors may use this as a shorthand or because they follow alternative validity frameworks. Based on the findings, one may argue that it may be more appropriate to say that the use of the phrase "validity of + test" must be accompanied by evidence showing the author's awareness of the full definition rather than saying the use of the phrase is "incorrect." Nevertheless, how to interpret this gap between the recommended terminology and actual scholarly practice in the field should be opened up for further discussion among language testing scholars.

Conclusion

Our findings confirm the prevalent use of the expression "validity of + test" among language testing scholars, with approximately one fourth of authors who mention validity using this expression. Moreover, our findings suggest that approximately one fourth of authors who use this expression know and acknowledge the *Standards*' and Messick's

definitions of validity. We hypothesize that convenience is the main driver behind the use of this expression; in addition, authors would not use it if they perceived it as being incorrect. It may be more accurate instead to view the expression "validity of + test" as an abstraction rather than an incorrect statement. In other words, "test" may be a shorthand for saying "inference about a test taker based on a score." The expression "a valid test" would therefore mean "a valid inference about the test taker based on a score" (score interpretation), while "valid test use" would express a "decision based on an inference about the test taker based on a score" (score use). We see this usage of "test = inference about a test taker based on a score" occurring in other expressions as well. For example, "test use" is commonly used even though it would be more accurately expressed as "score use." Likewise, the expression "the test is not valid" could be a shorthand for saying that the inference being made about the test taker based on their score is incorrect due to some fault or aberration (e.g., a fault in the test instrument and/or how was administered or scored, an aberration in the test taker causing them to perform worse or better than they usually would). Future research could investigate this hypothesis by asking language assessment researchers about their use of expressions that attribute the property of validity to tests.

ORCID

https://orcid.org/0000-0002-1113-0551

https://orcid.org/0000-0002-9967-1044

Publisher's Note

The claims, arguments, and counter-arguments made in this article are exclusively those of the contributing authors. Hence, they do not necessarily represent the viewpoints of the authors' affiliated institutions, or EUROKD as the publisher, the editors and the reviewers of the article.

Acknowledgements

Not applicable.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CRediT Authorship Contribution Statement

Haeun Kim: Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Project Administration

Shireen Baghestani: Conceptualization, Methodology, Validation, Writing - Review & Editing, Visualization

Generative AI Use Disclosure Statement

The authors did not use any AI tools in this manuscript.

Ethics Declarations

World Medical Association (WMA) Declaration of Helsinki–Ethical Principles for Medical Research Involving Human Participants

Not applicable.

Competing Interests

The authors have no competing interests,

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

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

Anthony, L. (2017). AntFileConverter (Version 1.2.1) [Computer Software]. Waseda University. Available from https://www.laurenceanthony.net/software

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Waseda University. Available from https://www.laurenceanthony.net/software

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. Sage Publications. https://doi.org/10.4135/9781071878811

Chapelle, C., Voss, E., & Kim, H. (2024). Designing evaluations for validation of language assessments. In A. Kunnan (Ed.), *The concise companion to language assessment* (pp. 67–79). Wiley-Blackwell.

Cronbach, L. J. (1971). *Essentials of psychological testing* (3rd ed.). Harper & Row.

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* (1st ed., pp. 621–694). American Council on Education.

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

Li, H., & Suen, H. K. (2012). Are test accommodations for English language learners fair? *Language Assessment Quarterly*, *9*(3), 293–309. https://doi.org/10.1080/15434303.2011.653843

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly*, *8*(2), 161–178. https://doi.org/10.1080/15434303.2011.565438

Oller, J. W. (2012). Grounding the argument-based framework for validating score interpretations and uses. *Language Testing*, *29*(1), 29–36. https://doi.org/10.1177/0265532211417212

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

Timpe-Laughlin, V., & Choi, I. (2017). Exploring the validity of a second language intercultural pragmatics assessment tool. *Language Assessment Quarterly*, *14*(1), 19–35. https://doi.org/10.1080/15434303.2016.1256406

Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Yan, X., Cheng, L., & Ginther, A. (2019). Factor analysis for fairness: Examining the impact of task type and examinee L1 background on scores of an ITA speaking test. *Language Testing*, *36*(2), 207–234. https://doi.org/10.1177/0265532218775764