**Language Teaching Research Quarterly** **EUROKD**

2025 (50), 159–168
https://doi.org/10.32038/ltrq.2025.50.11

# Test Validation: Beyond Arguments

Constant Leung*
*King's College London, UK*
Jo Lewkowicz
*University of Warsaw, Poland*
*King's College, London, UK*

**Correspondence**
**Email:** constant.leung@kcl.ac.uk

**Abstract**
The discussion in this article is organized in three parts. In the first part we acknowledge the significant contributions made by Carol Chapelle to the development of the argument-based approach to validation in English Language testing. The move away from a focus on the test itself (construct and content) to the use of test scores for validation purposes has been a significant conceptual shift. In the second part we suggest that productive operationalization of the argument-based approach rests on a stable and profession-wide taken-for-granted construct with specifiable and scalable features. Until recently the widely accepted construct of proficiency in internationalized English Language testing has been associated with the concept of communicative competence. However, this concept has been complexified by research in contingency in interactional language use and flexible multilingualism, some of which have been encapsulated in the expanded notion of language proficiency in the 2020 iteration of the CEFR Companion Volume. Language proficiency in this emergent dispensation embodies contingent, non-scalable and non-prescribable language use. Some of the questions and issues for test validation, including the argument-based approach, arising from this destabilization are discussed in the final section.

**ARTICLE HISTORY**
Received: 10 August 2024
Revised: 08 November 2025
Accepted: 15 November 2025

**KEYWORDS**
Language Proficiency, Test Validation, CEFR (Companion Volume), Contingent Language Use

**How to cite this article (APA 7th Edition):**
Leung, C., & Lewkowicz, J. (2025). Test validation: Beyond arguments. *Language Teaching Research Quarterly*, *50*, 159–168. https://doi.org/10.32038/ltrq.2025.50.11

## [1]Test Validation – A Milestone of Development

Carol Chapelle's notable contribution to the field of applied linguistics and English language testing has been broad and significant: one need look no further than at the range of papers included in this volume to understand the impact Chapelle has had on the field. In this paper, we focus on an important aspect of that contribution which is on the validation of large-scale psychometric tests. In so doing, we aim to strengthen

[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).

scholarship and academic discussion of key concepts that underpin language testing and assessment, namely that of validity and validation.

**Argument-Based Approach to Validation**

Validity which lies at the heart of all psychometric testing endeavours is a quality that affects the value of a test and acts as an indicator of the abstract concept it claims to measure (Davies et al., 1999). In educational assessment its importance has long been acknowledged (Newton & Shaw, 2014); in English language testing discussions of validity date back to the 1960s. However, our understanding of validity has changed over time and in a number of publications (e.g. Chapelle, 2021; Chapelle & Lee, 2021). Chapelle has charted its historical development and shown how the different iterations of validity and validation have been articulated in seminal resources such as Educational Measurement (1951-2006) and successive versions of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Validity in large-scale standardized English language testing can be charted over three time periods. In the 1960s, validity was viewed as an inherent attribute of a test, but one that failed to distinguish adequately between different types of validity (i.e. between content, construct, criterion-related, or face validity), and, possibly more importantly, failed to specify the types of evidence needed for a test to be validated (Chapelle, 2021). A shift in focus away from the test itself to test takers' performance on the test came in the late 1980s when Messick (1989) proposed a theoretical framework for establishing validity, one that added both a social and cultural dimension to the validation process. Messick (1989, p. 13) defined validity as 'an overall evaluative judgement of the degree to which evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores.' Validity was thus understood as a complex, unitary concept with construct at its centre. This approach, though largely theoretical, has been widely adopted by researchers in the field of English language testing (Chapelle & Voss, 2013).

An allegedly more practical approach to establishing validity was put forward by Kane in several publications (e.g. 1992, 2006, 2013). His argument-based approach to validation requires the specification of any 'proposed interpretations and uses of test results by laying out the inferences and assumptions leading from observed performances to the conclusions and decisions based on the performances' (Kane, 2006, p. 23, cited in Chapelle et al.,2010, p. 5). Characteristic of this approach is that it is context-specific and dependant on validity evidence from multiple sources gathered via 'a research program consisting of activities whose findings need to be integrated into a logical conclusion about the validity of test-score interpretation for particular uses', Chapelle (2021, p. 3).
Kane moved away from defining validity to detailing the validation process through developing a validity argument. At the outset he suggests that there is a need to

determine the claims to be made about test scores and their intended use, noting these are to be made within the social context of a given test, taking account of the test participants. The claims 'emerge' from the interpretative use argument formulated for each specific test and they are then investigated during the process of validation. Thus, empirical evidence is gathered to support claims about the test scores, such as the scores' relevance and reliability. This approach relies on a number of inherent assumptions underpinning the interpretation of test scores and their use. A key assumption is that we know what we are meant to be assessing, and test instruments can be developed to tap into the area/s of knowledge, skills and abilities of interest. In the English language testing literature, the term 'target language use' (TLU) is generally used to represent the focal area of interest (Bachman, 1991, among others). The domain of TLU is the basis of the formulation of construct.

Application of the argument-based approach to validation is comprehensively illustrated in the influential work of Chapelle and her colleagues during the validation of the 2005 Internet-based Test of English as a Foreign Language (TOEFL iBT) (see, Chapelle, Enright & Jamieson, 2008). They suggest the approach recognises the multifaceted meaning of test scores and, that the specific guidance and the conceptual infrastructure of Kane's argument-based approach enabled their reaching a conclusion regarding the adequacy of test score interpretation and test use (Chapelle, 2011, p. 20). In contrast to more traditional approaches to validation, the argument-based approach does not require the test construct to be defined at the start, prior to validation:

> 'Rather than prescribing what constructs must be, argument-validity takes into account the importance of the different possible meanings for various test scores by providing testers with tools to define the inferences in the validity argument in a manner that expresses the intended substantive meanings of the construct.' (Chapelle, 2021, p. 55)

It downplays, but does not eliminate, the need for defining the construct. Construct, thus, becomes one of the claims the test scores can insinuate, and not the key basis of the validation process. Echoing the claims of informal logic of argument (Toulmin, 1958/2003), the chain of reasoning is as follows: based on the test drawn up from the TLU claims are made about intended test scores and their use; scores are then observed, and inferences are drawn on their facets, e.g., their reliability, social consequence of use, etc. based on collected evidence during the validation process. Hence, the way the interpretative argument is specified makes clear how the validity argument can be questioned, weakened, limited, or refuted by research that supports or refutes the claims.

Following Kane's line of argument, an interesting issue arises if test takers of a particular test produce low level performance on the test. Does it mean that they are not able to perform adequately in terms of the thematic or focal area of knowledge and skills, i.e., on

the underlying construct of the test items? To address this question, we need to establish the correspondence between the test items and the construct being tested. Without knowing what the test items are tapping into, it is difficult to see what claim can be made in terms of evidence of the test's validity: the scores tell us nothing about the test. Without paying attention to the construct, it would be difficult to deal with such low-level performance. A similarly perplexing question that needs to be addressed relates to the emerging and expanded notion of language proficiency which we discuss discussed below.

**Fluid Foundations for Validation**

It can be argued that the viability and utility of the argument-based approach to test validation rest on the bedrock of stable and widely taken-for granted constructs ready for operationalization. In the case of English Language Teaching (ELT) as a transnational enterprise since the late 1970s, it would be accurate to say that the conceptualizations of language proficiency assessment frameworks and associated test instruments have been modelled on the concept of communicative competence. Building on the anthropological-ethnographic sensibilities emanating from the works of Hymes (1964/1972, 1974), Gumperz (1964), and many others, ELT adopted a pedagogic approach to language learning and language proficiency that is meant to be informed by the ways in which language is actually used by people in real-life activities (as opposed to a structural orientation that prioritizes the learning of vocabulary and grammar with some exemplifications of usage). The paper by Canale and Swain (1980) on communicative approaches to language teaching that appeared in the augural issue of *Applied Linguistics* can be seen in retrospect as a totemic publication representing the paradigm that has continued to frame ELT theorizing and research in the Anglophone academy and professional practices to this day. The conceptual and pedagogic concerns for mirroring real-life language use have meant that sociocultural and pragmatic conventions of language use in curriculum-relevant English-speaking communities are integrated into teaching content and proficiency descriptors alongside vocabulary and grammar. Within this purview of language use there is an implicit assumption that the ways in which language in real-life communication, both spoken and written, are 'describable' and stable in actual instances of use. In many ways this paradigmatic view of communicative competence has come to be seen as comprising immanent features of language use in ELT. This assumption is clearly reflected in the following proficiency descriptors:

Band 9

Under Fluency and Coherence: 'speaks coherently with fully appropriate cohesive features'

Under Lexical Resource: 'uses idiomatic language naturally and accurately'

Under Grammatical Range and Accuracy: 'produces consistently accurate structures apart from '"slips"' characteristic of native speaker speech'

IELTS Speaking (public version, https://assets.ctfassets.net/unrdeg6se4ke/mXLIv4Gi5tRicQNXuLkDK/3ef7e99b87f9d8d683cb4827c1cd4135/speakingbanddescriptors.pdf, accessed November 2024):

Descriptors similar to those cited above can be found in other well-established internationally recognized English proficiency frameworks, rating scales and teaching materials.

Needless to say, language teaching and assessment, particularly in formal educational settings, cannot cover the totality of language use in (any) society. Curricularization of a 'language' necessarily involves some form of selective and delimited representation of a focal construct through abstraction, particularly for large-scale standardized language testing. The widely observed guiding principle tells us that the construct of any test is the nature of the knowledge and ability we want to measure, by defining it abstractly (see Bachman & Palmer, 1996: 89; also Bachman & Palmer, 2010; Green, 2021: Part 3). This is consistent with a wider paradigmatic assumption in psychometric measure: 'A construct is always an ideal; we use it because it suits our theoretical approach' (Wilson, 2005, p. 28). The abstracted constructs that Bachman and others have alluded to are undoubtedly strongly associated with the theory/ies of communicative competence.

Given that it is not possible to include 'everything' in language teaching and assessment, the abstracted descriptive-cum-evaluative terms such as 'coherently' and 'appropriate' are referenced to a particular variety or type of language use. The seemingly neutral and descriptive terms 'idiomatic language' and 'native speaker speech' are signposts to knowledge and practices associated with that of a specific (and preferred) group of language users in the target language community/ies, again in abstracted terms. In a sense a construct is a characteriszation of the 'essences' of a particular kind of language use. Once the abstracted 'essences' have been identified, they can serve as the target language knowledge and skills that can be specified in language test task development. This movement from abstraction back to real-life test items is succinctly summarised by Cumming, et al:

> ' … our conceptualization of academic writing can be … presented in terms of … task stimuli, rhetorical functions, topic characteristics, and evaluative criteria.' (Cumming et al, 2000, p. 5)
>
> '[test domain] … writing assessment should require students to produce and sustain in writing coherent, appropriate, and purposeful texts in response to assigned tasks.' (Cumming et al, 2000, p. 7)

It is a matter of common observation that curricularized versions of English tend to be built on a sampling of the kinds of language use by the 'educated' language user in semi-formal public situations such as classroom, office/work or service encounters. The sustained portrayal of English language within a delimited range of topics and domains

in teaching and assessment materials over time has contributed to a particular representation of the English language as comprising a set of stable features in terms of topics and domains in the carrier contents, lexicogrammar and conventions of use (Gray, 2007, 2010a&b, 2013, 2016; Noori & Mirhosseini, 2021). The stability of this pedagogically delimiting view of English has been disrupted by studies in the fields of applied language studies and applied linguistics including the works in English as a lingua Franca, flexible multilingualism, translanguaging and World Englishes. It is evident that where social activities and social interactions involve participants from diverse ethnolinguistic backgrounds, increasingly an everyday phenomenon in many world locations including putatively English-speaking communities, the somewhat delimited portrayal of English language and the ways in which it is used do not reflect contemporary linguistic sensibilities and practices. (For a detailed discussion of this development see Leung, 2022, 2023; Leung & Jenkins, 2020). The 'mediation' component within the expanded concept of language proficiency presented in the revised iteration of the Common European Framework of Reference for Languages (CEFR, Companion Volume, Council of Europe, 2020) captures this contemporary complexity well. Mediation is understood to be part of social interaction with others; it is directly about a language user's conduct. 'The term "mediation" is … used to describe a social and cultural process of creating conditions for communication and co-operation, facing and hopefully defusing any delicate situations and tensions that may arise.' (Council of Europe, 2020, p. 91)

Mediation can take place in three ways: mediating a text, mediating concepts and medicating communication. Mediating a text involves 'passing on to another person the content of a text to which they do not have access, often because of linguistic, cultural, semantic or technical barriers.' Mediating concepts 'refers to the process of facilitating access to knowledge and concepts for others, particularly if they may be unable to access this directly on their own.' (Council of Europe, 2020, p. 91). For reasons of direct thematic relevance, we pay particular attention to mediating communication in this discussion.

'"Mediating communication" aims to facilitate understanding and shape successful communication between users/learners who may have individual, sociocultural, sociolinguistic or intellectual differences in standpoint. The mediator tries to have a positive influence on aspects of the dynamic relationship between all the participants, including the relationship with themselves. Often, the context of the mediation will be an activity in which participants have shared communicative objectives, but this need not necessarily be the case. The skills involved are relevant to diplomacy, negotiation, pedagogy and dispute resolution, but also to everyday social and/or workplace interactions. Mediating communication is thus primarily concerned with personal encounters. **This is not a closed list – users may well be able to think of other types of mediation activities not included here**.' (Council of Europe, 2020, p. 91emphasis added)

Furthermore, mediation can involve the use of more than one language and emotional intelligence:

'Cross-linguistic and cross-modal mediation, in particular, inevitably involve social and cultural competence as well as plurilingual competence.'

'A person who engages in mediation activity needs to have a well-developed emotional intelligence, or an openness to develop it, in order to have sufficient empathy for the viewpoints and emotional states of other participants in the communicative situation.' (Council of Europe, 2020, p. 91.)

Mediating communication is thus conceptualised as a multi-faceted act of communication involving:

- agentive and volitional decisions to create 'conditions for communication and co-operation, facing and … defusing any delicate situations and tensions that may arise'
- activation of plurilingual repertoires, multi/trans-cultural knowledge and empathetic understanding of (possible) communication challenges experienced by interlocutors (emotional intelligence)
- possible enactment in variety of social and communication settings ('… relevant to diplomacy, negotiation, pedagogy and dispute resolution, but also to everyday social and/or workplace interactions').

The construct of mediating communication is clearly very different from that depicted in the guidance set out AERA, APA, NCME (2014, p. 23, emphasis added):

'The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be delimited clearly, and the **construct or constructs** that the test is intended to assess should be described clearly.'

There is no clearly defined language user population in the CEFR's description of mediating communication; its enactment can take place in any everyday social and/or workplace interactions ('… not a closed list – users may well be able to think of other types of mediation activities not included here'); the kind/s of language knowledge and skills involved is difficult to pre-identify as it is emergent from actual social interaction as it unfolds – it is not possible to work out in advance what speech functions and lexicogrammatical and pragmatic knowledge and skills are needed; it is not possible to predict what kind of plurilingualism is needed ahead of the actual moment of mediation as much would depend on the participants' linguistic repertoires, subject content and communication needs as they emerge. In addition, emotional intelligence, if understood as an ability or trait of the language user (Mayer et al., 2016), it is difficult to pre-specify as its manifestation in actual conduct can vary across instances of social interaction and

time. The outcome of any mediation effort can only be judged 'in the moment' of occurrence; there is no guarantee that the efforts of a language user to mediate for others would lead to any particular predictable outcome as interlocutors' responses cannot be predicted, as all mediation activities are interactionally brought about 'in the moment'. So, all communicative ingredients in mediating communication is contingently situated and in flux.

## Construct for More Arguments

So, what happens to argument-based validation when a significant component of proficiency is unstable and unpredictable? If test outcomes are fluidly contingent on participant/interlocutor co-construction, is it possible to assign meaning to the test score (even if we could design some sort of appropriate measurement)? Alternatively, as there are unstable elements within the construct upon which to build the measurements, would we need a different kind of argument-based approach to validation, one that is not necessarily based on test scores?

It would seem that the argument-based approach to test validation is only operationally defensible for those aspects of test-taker performance that are closely related to an underlying construct (subscribed to by the test providers), where the construct comprises of more or less stable components (e.g. lexicogrammar in functional use), whereas agentive and contingent language use is situated in social interaction. This would suggest that only those components of the construct that can be tapped into through psychometrically oriented scaled measurement can be represented by the test score. The meaning of the test score would, therefore, need to be clearly stated. Furthermore, in terms of representing test-takers 'global' performance, it would be necessary to combine test scores with outcome-based evaluation of language use. For instance, where mediating communication is part of the test performance, it would be necessary to adjudge the outcome of the interaction on a case-by-case basis, that in turn would necessitate establishing situationally sensitive evaluation criteria in order to avoid arbitrary decision-making. Yet, how can we know if we are dealing with the same phenomenon over time if we do not know what we set out to assess in the first place? Perhaps it would be a good idea to try to empirically identify patterns or attempts by interlocutors to mediate in situated interactions inductively. And, if we were to do this, what might be the implications for language testing research and theorozing?

## ORCID

 https://orcid.org/0000-0001-8533-2837
 https://orcid.org/0000-0003-2426-1443

## Publisher's Note

the authors' affiliated institutions, or EUROKD as the publisher, the editors and the reviewers of the article.

**Acknowledgements**
Not applicable

**Funding**
Not applicable

**CRediT Authorship Contribution Statement**
Constant Leung: Conceptualization, Writing - Original Draft, Writing - Review & Editing
Jo Lewkowicz: Conceptualization, Writing - Review & Editing

**Generative AI Use Disclosure Statement**
No AI use.

**Ethics Declarations**
**World Medical Association (WMA) Declaration of Helsinki–Ethical Principles for Medical Research Involving Human Participants**
No medical data involved or included in the article.

**Competing Interests**
No competing interests.

**Data Availability**
This is a conceptual paper, no empirical data was used.

**References**

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing* [Text]. AERA. https://doi.org/10.1111/emip.12045

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly*, *25*(4), 671-704. https://doi.org/10.2307/3587082

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

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

Canale, M., & Swain, M. (1980). A domain description for core FSL: Communication skills. In *Ontario Assessment Instrument Pool: French as a second language, junior and intermediate divisions* (pp. 27–39). Ministry of Education (Ontario).

Chapelle, C. A. (2011). Validity argument for language assessment: The framework is simple… *Language Testing*, *29*(1), 19-27. https://doi.org/10.1177/0265532211417211

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

Chapelle, C.A., Enright, M.K. & Jamieson, J. (eds.). (2011) *Building a validity argument for the test of English as a foreign language.* Routledge. https://doi.org/10.4324/9780203937891

Chapelle, C.A., Enright, M.K. & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *9*(1), 3-13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Chapelle, C.A. & Lee, H. (2021). Conceptions of validity. In Fulcher, G. & Harding, L. (eds.) *The Routledge handbook of language testing* (pp. 17-31). Routledge. https://doi.org/10.4324/9781003220756

Chapelle, C.A. & Voss, E. (2013). Evaluation of language tests through validation research. In Kunan, A. (ed.) *The companion to language assessment* (1079-97). Wiley. https://DOI: 10.1002/9781118411360

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment - Companion Volume*. Council of Europe.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper*. Educational Testing Service.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing* (Studies in Language Testing, Vol. 7). Cambridge University Press.

Gray, J. (2007). *A study of the cultural content in the British ELT global coursebook: A cultural studies approach. (Unpublished PhD thesis)* Institute of Education]. London.

Gray, J. (2010a). The branding of English and the culture of the new capitalism: Representations of the world of work in English Language textbooks. *Applied Linguistics*, *31*(5), 714-733. https://doi.org/10.1093/applin/amq009

Gray, J. (2010b). *The construction of English-Culture, consumerism and promotion in the global ELT coursebook*. Palgrave McMillan. http:// doi 10.1057/9780230283084

Gray, J. (2016). ELT materials: Claims, critiques and controversies. In G. Hall (Ed.), *The Routledge handbook of English language teaching* (pp. 95-108). Routledge.

Green, A. (2021). *Exploring language testing and assessment (2nd edition)*. Routledge. *Exploring language testing and assessment (2nd edition)*. Routledge.

Gumperz, J. (1964). Linguistic and social interaction in two communities. *The American Anthropologist, 6*, 137-153.

Hymes, D. (1964/1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Penguin.

Hymes, D. (1974). Ways of speaking. In R. Bauman & J. Sherzer (Eds.). *Explorations in the ethnography of speaking* (pp. 433-451). Cambridge University Press.

Kane, M. T. (1992).Validating high-stakes testing programs. *Educational Measurement: Issues and Practices*, *21*(1), 31-41.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Praeger. ISBN 978-0-275981-25-9

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

Leung, C. (2022). Language proficiency: From description to prescription and back? *Educational Linguistics*, *1*(1), 56-81.

Leung, C. (2023). English language proficiencies – recasting disciplinary and pedagogic sensibilities. *Critical Inquiry in Language Studies*, *20*(4), 426–447. https://doi.org/10.1080/15427587.2023.2292185

Leung, C., & Jenkins, J. (2020). Mediating communication – ELF and flexible multilingualism perspectives on the Common European Framework of Reference for Languages. *Australian Review of Applied Linguistics*, *3*(1), 26-41. https://doi.org/10.29140/ajal.v3n1.285

Mayer, J. D., Caruso, D. R., & Salovey, P. (2016). The ability model of emotional intelligence: Principles and updates. *Emotion Review*, *8*(4), 290–300. https://doi.org/10.1177/1754073916650505

Messick, S. (1989). Validity. In R.N.N. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Macmillan.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Sage.

Noori, M., & Mirhosseini, S.-A. (2021). Testing language, but what?: Examining the carrier content of IELTS preparation materials from a critical perspective. *Language Assessment Quarterly*, *18(4), 382-397.* https://doi.org/10.1080/15434303.2021.1883618

Toulmin, S. E. (1958/2003). *The uses of argument*. Cambridge University Press. https://doi.org/10.1017/CBO9780511840005

Wilson, M. (2005). *Constructing measures: An item response modelling approach*. Lawrence Erlbaum Associates, Publishers. https://doi.org/10.4324/9781003286929