Title	The vexing problem of validity and the future of second language assessment
Author(s)	Vahid Aryadoust

This is the accepted author's manuscript of the following article:

Aryadoust, V. (2023). The vexing problem of validity and the future of second language assessment. *Language Testing, 40*(1), 8–14. https://doi.org/10.1177/02655322221125204

The vexing problem of validity and the future of second language assessment

Vahid Aryadoust, PhD

National Institute of Education

Nanyang Technological University

Singapore

Vahid.aryadoust@nie.edu.sg

Citation:

Aryadoust, V. (2023). The vexing problem of validity and the future of second language assessment. *Language Testing, 40*(1), 8-14.

https://doi.org/10.1177/02655322221125204

Abstract

Construct validity and building validity arguments are some of the main challenges facing the language assessment (LA) community. The notion of construct validity and validity arguments arose from research in psychological assessment and developed into the gold standard of validation/validity research in LA. At a theoretical level, construct validity and validity arguments conflate the scientific reasoning in assessment and policy matters of ethics. Thus, a test validator is expected to simultaneously serve the role of

conducting scientific research and examining the consequential basis of assessments. I contend that validity investigations should be decoupled from the ethical/social aspects of assessments. In addition, the near-exclusive focus of empirical construct validity research on cognitive processing has not resulted in sufficient accuracy and replicability in predicting test takers' performance in real language use domains. Accordingly, I underscore the significance of prediction in validation, in contrast with explanation, and propose that the question to ask might not so much be about what a test measures as what type of methods and tools can better generate non-test language use profiles. Finally, interdisciplinary alliances with cognitive and computational neuroscience and artificial intelligence (AI) fields should be forged to meet the demands of LA in the 21st century.

Keywords: artificial intelligence (AI); authenticity; language assessment; interdisciplinary research; neuroscience; validity; validity arguments

The vexing problem of validity and the future of second language assessment

The year 2023 marks the 40th anniversary of the launch of *Language Testing*. In the past 40 years, the field of language assessment (LA) has become increasingly more diverse in scope and focus, and it has experienced several transitional phases. This trend is evident in the research that has hitherto appeared in field journals (notably, *Language Testing*), books, conferences, and webinars around the world (Aryadoust et al., 2020). While a number of research problems have challenged LA researchers and practitioners in the past 40 years, in my view, validity remains the main challenge of the field. I offer two reasons for this chronic challenge: First, construct validity research has been unable to achieve its main goal, which is to unpack the constituents of language tests and yield reproducible and accurate language models. Thus, conducting research by using the currently available tools for construct validity to unravel test-taking processes is insufficient. Second, the formulation of validity as an argumentation process conflates science and social/ethical aspects of assessment. To address the present challenge, I propose that the focus of LA should shift from the construct-oriented explanation of language behavior to an authenticity-based prediction of *language use profiles* in target language use (TLU) domains. I conclude by stating that the rapid growth of technology and artificial intelligence (AI) affords the field opportunities for innovative interdisciplinary research and enhances current assessment practices in LA.

Emergence of Construct Validity

The earliest mentions of validity date back to the late 19th century. At the time, validity was conceptualized as a statistical property of tests. Nowadays, however, many, if not most, contemporary scholars view validity as the appropriateness and relevance of the

inferences drawn from test scores rather than a property of tests (e.g., Chapelle & Voss, 2021; Chalhoub-Deville & O'Sullivan, 2020; Newton & Shaw, 2014). The concept of construct validity was introduced by Meehl and Challman in the midst of the 20th century, and the first in-depth treatment of construct validity was presented by Cronbach and Meehl (1955), who differentiated between criterion, content, and construct validity. At the heart of Cronbach and Meehl's framework was the nomological network that assumed the meaning of psychological traits is necessarily entailed in the laws by which the traits work. In the absence of appropriate methods to investigate construct validity, several scholars proposed empirical methods for validation (investigation of validity) such as the multitrait-multimethod matrix of validation (Campbell & Fiske, 1959), the nomothetic span (Whitely, 1983), the content, structural, and substantive aspects of construct validity (e.g., Loevinger 1957; Messick, 1975, 1989). It may be said that these methods drew upon a combination of reductionism (that complex phenomena are explained in terms of the nature of their fundamental units) and the complex systems approach. The latter approach posits that due to the intrinsic complexity in the dependencies of the units, it is difficult to explain the properties and behavior of the phenomena in terms of their basic units.

At any rate, construct validity was soon recognized as the key requirement in test development. In Messick's (1989) unified framework, it was hailed as the key validity component under which other components were unified. Messick's (1975, 1989) formulation of construct validity marked the beginning of an era that was characterized by the significance of the accumulation and judgment of validity "evidence," inter alia. Unlike most previous scholars, however, Messick underscored the significance of both

evidential and consequential basis of interpretations and uses of test scores in establishing construct validity, thereby assigning a prominent role to non-epistemic virtues in the otherwise value-neutral test validity research and practice.

Limitations of Empirical Construct Validity Research

As discussed above, Messick's (1989) philosophical definition of construct validity entails evidence to buttress both the interpretations and consequences of uses of test scores. However, and in my view fortunately, the majority of empirical second language studies examining the construct validity of measurement tools are focused only on the epistemic properties of tools, seeking to reverse-engineer the causes of observed variance in the test data (Aryadoust et al., 2020). There has been a considerable advancement on this front with regards to the definition and operationalization of a number of constructs. Interestingly, this value-neutral approach to validity has also been adopted by Kane and associates in a less-cited study of validity and reliability (Kane & Case, 2004).

Despite this progress, it would be reasonable to say that it is not yet possible to accurately pin down the constructs elicited by language tests. This is a well-known limitation in construct validity research that has been noted by prominent scholars such as Buck (2001, p. 106), who stated "[u]nfortunately, determining the competencies that underlie performance on a set of test tasks is a complex and indirect process, and we have no way of knowing for certain which components are required by any particular task." Even if one were able to generate a model parameterizing all the causal, mediating, and moderating variables explaining the variance in test scores, the model would not

necessarily be reflective and/or predictive of real-world domains (Yarkoni & Westfall, 2017); rather, it would be limited to the test data that it was fed[1].

In my view, there are four major reasons for the limitation of the construct validity research: (i) little emphasis on tightly controlled experimental studies, while, in contrast, cross-sectional studies dominate the construct validity research in LA and second language; (ii) minimal or no attention to the replicability, reproducibility, and "accuracy" of research findings particularly in evidence-gathering validation in LA (see Ioannidis, 2005, for a definition of accuracy); (iii) minimal attention to neurophysiological and neurocognitive processes of test takers under test and non-test conditions; and (iv) the restriction of assessments as primarily being "snapshots" of language ability rather than well-rounded tools for authentic representations of language use.

Overall, it is compelling that, in contrast with the virtue-laden formulation of construct validity (Messick, 1989), empirical construct validity research mostly has not committed to mixing the current science of assessment with ethical aspects of uses of test scores. In the meantime, it seems to me that this research strand has oversimplified test-taking processes by paying far less heed to the operationalization of the *actual* language use in TLU domains.

Emergence of Argument-Based Validation

Since 1992, Kane has introduced and extended an argument-based validity framework as an upgrade of Messick's (1989) framework. In recent publications, Kane and influential

[1] This is so despite the philosophical arguments intended to link constructs and TLU domains such as domain specification and construct definition in LA. For example, the widely used θ_j (ability) in IRT is not necessarily analogous to the language ability in real TLU domains.

LA scholars, like Chapelle and Voss (2021) and Bachman and Palmer (2010), proposed that the process of collecting and interpreting data for validation purposes should follow Toulmin's (2003) informal logic which includes backing, evidence, warrants, counterevidence, and qualifiers. They further indicated that validation is a perpetual process which consists of building inference chains, refuting counterclaims, and advancing arguments in support of the test designer's claims about the interpretations and use of test scores.

With the exception of Toulmin's (2003) informal logic, the underlying ideas in argument-based validity were mostly discussed by Messick (1975, 1989). Drawing on, for example, system model and adversary model, Messick (1975) proposed that every piece of evidence produced in support of test validity should be assessed against counterproposals that are advanced by "adversaries." He stressed that counterhypotheses should be proposed to "direct attention to vulnerabilities" in construct validation and counterproposal or counterplans should be advanced to highlight vulnerabilities in test use. Thus, Messick (1975) urged validators to evaluate the supporting validity evidence against counterhypotheses and rebuttals. He stressed that, in advancing a counterproposal, the claimant or test designer should propose a "quite different assessment approach" (p. 21) in construct validation. He further emphasized that "counterproposals or counterplans" should be advanced to highlight vulnerabilities in test use, so that the social consequences of using or not using the test scores will be examined. Thus, Messick is not as forgiving as Kane in relation to evaluating rebuttals and counterevidence, whereas Kane's framework and its adaptions in LA give the test

developer more leeway for judging the validity evidence favorably in the presence of rebuttals.

Limitations of Argument-Based Validation

I believe that there are two major limitations with argument-based validity. It is not clear (i) how the evidence accumulated in each step should be weighted, aggregated, and evaluated; and (ii) under what circumstances the test should be regarded as valid for the purposes it was developed. A lack of a weighting system is due to the fact that the framework resembles a litigious rather than a scientific process of the attainment of knowledge. This assimilates validation to discretionary domains wherein a "decision-maker has the freedom to select one interpretation or outcome from a number of permissible options" (Zeleznikow, 2006, p. 290). According to Zeleznikow (2006, p. 290), in such domains, the onus of evidence-weighting falls on the decision-maker. In several systematic reviews of the published L2 research, my colleagues and I have found that the "evidence" backing possible validity arguments is almost always evaluated favorably in the published research, while the significance of attenuating evidence, whatever it might be, is downplayed, if reported at all (e.g., Aryadoust et al., 2020).

Toulmin's informal logic framework, which undergirds the argument-based validity frameworks, has been critiqued on similar grounds. A major omission in Toulmin's framework is the lack of a clear discussion of the nature of rebuttals (Verheij, 2006, p. 196) and procedures for treating them. According to Zeleznikow (2006, p. 290), "it is possible and sometimes probable for two decision makers [validator and claimant in this context] to arrive at a different decision based on the same facts." Such potential discrepancies in decision-making are attributed to factors such as "induction and

intuition" of decisionmakers as well as their perceptions of "the social impact of decisions" (Zeleznikow, 2006, p. 290). Overall, while evidence-weighting is perhaps suitable for judicial systems, in the absence of any rigorously researched guidelines, I believe it is not a scientific endeavor and does not value-add to LA research. Thus, the challenge is to convince the field to move on from and supplant this value-laden approach with a robust value-neutral science.

With regards to the second limitation, there has been a split camp in LA with regards to argument-based validity, with the critics forming a minority group. Alan Davies, one of the critics, contended that the ambiguity surrounding Kane's (and similar) frameworks causes them to be "philosophical standpoints," which are "admirable but […] leave unanswered the question of how you demonstrate that validity has or has not been achieved" (Davies, 2011, p. 39). Relatedly, Davies and Elder (2005) indicated that conflating science and ethical concerns, which is Messick's (1989) proposal, would not be useful for LA research. Davies and Elder (2005, p. 799) wrote that such a practice "adds to the problem of validity by extending its scope into the social and the ethical." In another commentary, Davies (2011) lamented that in Kane's framework, "[t]here is no discussion, no description, no examination of the kinds of evidence and methodologies that validation necessarily requires […] Because Kane is discussing at a meta-linguistic level, it is unclear what evidence to draw on in his search for validity" (Davies, 2011, p. 40).

Finally, a strong objection to the merging of science and the ethical in validity research was voiced by Borsboom and associates from the psychology field. For example, Borsboom and Wijsen (2016, pp. 281-282) noted that "[a] lack of ethical

justification cannot be made up for by better psychometric properties or vice versa; Nazi war crimes are not mitigated by the fact that the Nazis used really good measurement procedures […]."

In sum, building argument-based validity s and investigating their strength and weaknesses tend to render the examination of validity an onerous litigious endeavor. The main challenge of the language assessment field is that ethics of assessment and the science of assessment development and validation should be divorced, as they are not additive or "scalable" (Borsboom & Wijsen, 2016) and cannot compensate for each other's limitations or give one another any legitimacy. There is no denial that, as a community, the LA field should have firm convictions on justice and non-epistemic values in LA and that it should not become cavalier about these matters. Nevertheless, such an effort ought to be distinguished from the epistemic goals of scientific research in the field.

A Way Forward

In view of the limitation of theory and practice in construct validation and argument-based validation, I suggest that researchers in the field shift their attention to authentic assessment more than ever before. In addition, language assessment researchers will do well if they embrace affordances of AI. In this proposal, support for authenticity should be sought from features of natural language input and individuals' neurocognitive interactions with the input, rather than from performance on inauthentic tests that dominate the extant assessment regimes (Hasrol et al., 2022). Specifically, modern information and computer technology (ICT) and AI technology, like webcam-based eye-trackers, keyloggers, neuroscanners etc., afford the field the possibility of aggregating

and analyzing domain-specific language use and engendering "language use profiles" (Gruba, 2019, p. 229) per language learners, in contrast with proficiency scores that primarily represent test conditions. The longitudinal mining of language use data will enable the field to conduct "testing" without using tests.

While I am well-aware of the limitations of technology, I believe it is unlikely that many of the frameworks in LA will remain relevant in the age of AI. I concur with the prediction by Gruba (2019, p. 229) that "[l]anguage testing, as we know it, will cease to exist within a generation." The main contributors to this envisioned fate would be the interdisciplinary elements of AI, computational linguistics (Gruba, 2019), as well as neuroscience, which are capable of generating efficient means to produce longitudinal language profiles.

In sum, to move on from artificial construct development and testing, there is a need to extend interdisciplinary engagements. Importantly, the new generations of graduate programs should venture to cross the boundaries traditionally assumed to exist between LA, cognitive neuroscience, and computer sciences, and take a cross-disciplinary perspective on language to build new theories and finetune the methods for language "assessment."

Acknowledgement

I wish to thank Albert Weideman, Guangwei Hu, and Rie Koizumi for their insightful comments on an earlier and longer draft; and Michelle Raquel and Brad Blackstone for reading a more recent draft of this article. The views expressed are mine and do not necessarily reflect the views of these colleagues.

References

Aryadoust, V., Zakaria, A., Lim, M. H., & Chen, C. (2020). An extensive knowledge mapping review of measurement and validity in language assessment and SLA research. *Frontiers in Psychology, 11*, 1941. doi: 10.3389/fpsyg.2020.01941

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

Borsboom, D., & Wijsen, L. D. (2016). Frankenstein's validity monster: The value of keeping politics and science separated. *Assessment in Education: Principles, Policy & Practice, 23*(2), 281-283. https://doi.org/10.1080/0969594X.2016.1141750

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105. https://doi.org/10.1037/h0046016

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical development and integrated arguments*. British Council Monographs.

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

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. https://psycnet.apa.org/doi/10.1037/h0040957

Davies, A. (2011). Kane, validity and soundness. *Language Testing, 29*(1), 37-42. https://doi.org/10.1177/0265532211417213

Davies, A., & Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 795-813). Lawrence Erlbaum Associates.

Hasrol, B. S., Zakaria, A., & Aryadoust, V. (2022). A systematic review of authenticity in language assessment. *Research Methods in Applied Linguistics, 1*(3), 100023. https://doi.org/10.1016/j.rmal.2022.100023

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0040168

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*(3), 221-240. https://doi.org/10.1207/s15324818ame1703_1

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). American Council on Education and Macmillan.

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

Newton, P. E., & Shaw, S. (2014). *Validity in educational and psychological assessment* (M. Lagrange Ed.). SAGE.

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

Verheij, B. (2006). Evaluating arguments based on Toulmin's scheme. In D. Hitchcock & B. Verheij (Eds.), *Arguing on the Toulmin model: New essays in argument analysis*. Springer.

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197. https://doi.org/10.1037/0033-2909.93.1.179

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zeleznikow, J. (2006). Using Toulmin argumentation to support dispute settlement in discretionary domains. In D. Hitchcock & B. Verheij (Eds.), *Arguing on the Toulmin model: New essays in argument analysis and evaluation* (pp. 289-301). Springer.