Received: 11 January 2024 | Revised: 29 April 2024 | Accepted: 11 May 2024

DOI: 10.1111/medu.15452

CROSS-CUTTING EDGE

Expanding Kane's argument-based validity framework: What can validation practices in language assessment offer health professions education?

David Wei Dai[1] | Thao Vu[2] | Ute Knoch[3] | Angelina S. Lim[2] | Daniel Thomas Malone[2] | Vivienne Mak[2,4]

[1]UCL Institute of Education, University College London, London, UK

[2]Faculty of Pharmacy and Pharmaceutical Sciences, Monash University, Parkville, Melbourne, Victoria, Australia

[3]Language Testing Research Centre, The University of Melbourne, Parkville, Melbourne, Victoria, Australia

[4]Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Melbourne, Australia

Correspondence
David Wei Dai, UCL Institute of Education, University College London, 20 Bedford Way, London, WC1H 0AL, UK.
Email: david.dai@ucl.ac.uk

Abstract

Context: One central consideration in health professions education (HPE) is to ensure we are making sound and justifiable decisions based on the assessment instruments we use on health professionals. To achieve this goal, HPE assessment researchers have drawn on Kane's argument-based framework to ascertain the validity of their assessment tools. However, the original four-inference model proposed by Kane – frequently used in HPE validation research – has its limitations in terms of what each inference entails and what claims and sources of backing are housed in each inference. The under-specification in the four-inference model has led to inconsistent practices in HPE validation research, posing challenges for (i) researchers who want to evaluate the validity of different HPE assessment tools and/or (ii) researchers who are new to test validation and need to establish a coherent understanding of argument-based validation.

Methods: To address these identified concerns, this article introduces the expanded seven-inference argument-based validation framework that is established practice in the field of language testing and assessment (LTA). We explicate (i) why LTA researchers experienced the need to further specify the original four Kanean inferences; (ii) how LTA validation research defines each of their seven inferences and (iii) what claims, assumptions and sources of backing are associated with each inference. Sampling six representative validation studies in HPE, we demonstrate why an expanded model and a shared disciplinary validation framework can facilitate the examination of the validity evidence in diverse HPE validation contexts.

Conclusions: We invite HPE validation researchers to experiment with the seven-inference argument-based framework from LTA to evaluate its usefulness to HPE. We also call for greater interdisciplinary dialogue between HPE and LTA since both disciplines share many fundamental concerns about language use, communication skills, assessment practices and validity in assessment instruments.

1 | BACKGROUND

A critical part of health professions education (HPE) is making necessary and defensible educational decisions related to students' learning, progression and readiness for practice. These educational decisions can directly affect students' learning experiences and outcomes in various degrees. Consequences of these decisions also impact a range of HPE stakeholders such as the patients, colleagues and communities with which health professional students interact.

Education research literature, including HPE literature, has established the importance of ensuring and evaluating the quality of such educational decisions.[1,2] More than a century of work on assessment validation – centring on the process of collecting and interpreting evidence from assessment to argue for the validity of our judgments and decisions about learners[3] – has resulted in an evolution in how we understand and undertake assessment validation. Moving from focusing on *types of validity*, the field has increasingly oriented to *sources of evidence* to support construct validity as a single type of validity in a unified framework.[1] Most recently, validation literature has also turned to promote a two-step process: (i) beginning with statements or arguments of what we want to argue in the inferences we make from assessment scores, and (ii) then evaluating the defensibility of these arguments by interpreting relevant types of evidence that can support or refute each of the arguments.[2,3] This argument-based validation framework has received strong support in medical education research as it is applicable to both quantitative and qualitative assessment tools, as well as assessment programmes that utilise multiple assessment data points and forms of assessments.[3,4]

While there is an intensifying interest in applying Kane's argument-based validity framework to HPE assessment validation, there has been little cross-disciplinary conversation in HPE and other related disciplines in terms of how to conduct Kanean validation studies.[5] Scholars in language testing and assessment (LTA), e.g. have been drawing on, and expanding, the Kanean framework for nearly two decades.[6] This includes specifying additional inferences not originally proposed by Kane, and further elucidating possible assumptions and sources of backing that may strengthen an argument in a specific context.

The purpose of this Cross-Cutting Edge article is to introduce a body of work from the discipline of LTA and to suggest how a cross-disciplinary perspective can help enhance assessment validation in HPE. Using the expanded Kane's framework in LTA, we – a team of LTA and HPE researchers – revisit the application of argument-based validity in selected publications from the HPE literature and identify areas for improvement. Given the notable lack of interdisciplinary discourse on methodologies for expanding Kane's framework, we invite HPE assessment researchers to experiment with the expanded validity framework from LTA and evaluate its usefulness to HPE.

2 | ARGUMENT-BASED VALIDATION IN LTA

LTA is a research field that examines whether language users have the ability to communicate effectively in everyday, academic or professional contexts. Since language test scores are routinely used for high-stakes decisions such as university admission and professional registrations, many language testing companies spend substantial resources to produce sound, stakeholder-accountable validity arguments for their test products. As a side-product, researchers involved in LTA have grappled with integrating research questions and practical concerns into validation arguments originating from educational assessment.[7] This has prompted further specification of validation arguments.[8–10]

With this long history of engagement with Kane's argument-based validation in the language testing community, researchers have noted the value of increasing elaboration of Kane's original four-inference structure. On top of the four inferences in Kane's original model (scoring or evaluation, generalisation, extrapolation, implications), current argument-based validation endeavours in LTA tend to also include a *domain description* inference before the scoring inference, an *explanation* inference between the generalisation and extrapolation inferences and a *decisions* inference and a *consequence* inference after the extrapolation inference (see grey boxes in Figure 1). In what follows, we first describe these 'new' inferences as they are used in LTA, before unpacking what is typically included in the other four inferences that largely overlap with those proposed by Kane. For each of these, we explicate the types of assumptions and typical sources of backing included in each inference, followed by a holistic presentation of the seven-inference framework and their associated claims, assumptions and backing in Table 1.

The reasoning behind a separate *domain description* inference is that we need to first ascertain if the selection, design and delivery of the assessment tasks take the relevant target domain, or in LTA parlance, target language use (TLU) domain, into account. This inference was first introduced by Chapelle[6] in their work building a validity

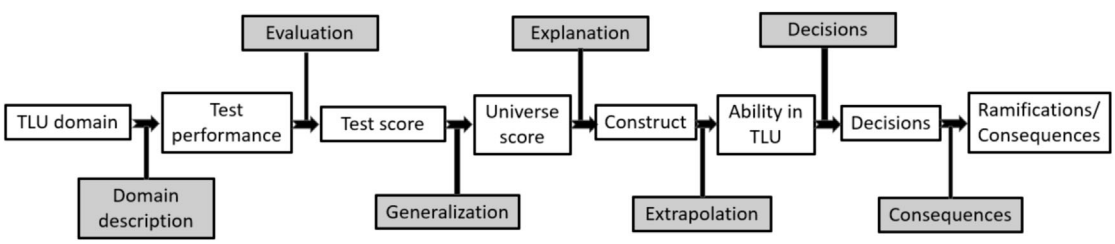

FIGURE 1 Inferences in LTA validation argument (inferences in grey).

argument for the Test of English as a Foreign Language (TOEFL) test, an English language proficiency test developed to make university entry decisions. A key aspect of the domain description inference is to justify to key stakeholders that the test tasks are likely to elicit skills test-takers are required to perform in the target domain. For this reason, test tasks should correspond closely to tasks in the target domain. To achieve this, the development of assessments should usually start with a careful examination of the domain (e.g. by conducting a domain or needs analysis) and attempt to mirror, as far as possible, these tasks in the assessment (see Dai[11] and Youn,[12] e.g. in LTA).

Similarly, LTA scholars have added an *explanation* inference to expand on Kane's framework. Although not stated explicitly, Kane implicated the importance of explanation when he contended that we cannot "assign trait interpretations to observable attributes by assuming the existence of an underlying trait or construct that accounts for the observed regularities in performance".[13] Because of the complexity of measuring intangible, amorphous constructs such as language abilities, LTA researchers have added the explanation inference to ensure that the test scores from language tests truly reflect and explain the underlying construct or trait that is being measured. The claim of the explanation inference is that expected scores are attributable to the theoretical construct the assessment is designed to measure. Common sources of evidence to back up the explanation inference in language test validation include (i) investigating the test-taking process and relating it to theories of language proficiency; (ii) examining the relationship between test scores obtained from this test in question with tests that claim to measure similar constructs (convergent validity) and different constructs (divergent validity) and (iii) explorations of the internal structures of sectional test scores in the test to show that they conform to theory.

LTA researchers have also split the original Kanean *implications inference* into two inferences: *utilisation* (also called *decisions*) and *consequences*, each focussing on different issues. The claim underlying the *utilisation or decisions* inference is that decisions made based on the estimates of the quality of the performance collected from test takers are appropriate and well communicated.[14] Backing for this inference includes statistical analyses and the collection of qualitative data to establish whether the estimates resulting from the performance on the assessment are useful for decision-making, and a review of the standards set on the assessment to ensure that test stakeholders are using appropriate and fair cut-scores to classify test takers into decision-making levels.

The *consequences* (also sometimes called *ramifications*) inference in the LTA literature has the underlying claim that test consequences are beneficial to test users in the domain. Backing to support this inference investigates whether test takers who have achieved the required score have sufficient language proficiency to function in the relevant domain. A further concern of the consequences inference is whether the format of the assessment has a positive influence (also called 'washback') on teaching, learning and test preparation. Backing to support this inference would involve interviews with test stakeholders or end-users such as employers, teachers, test preparation providers and test takers.

The inferences described above – namely domain description, explanation, utilisation or decisions and consequences or ramifications – are those that have been added by researchers in the field of LTA in recent years. In the next paragraphs, we present how LTA researchers interpret the three standard Kanean inferences – *scoring or evaluation*, *generalisation* and *extrapolation*, and the types of warrants, assumptions and sources of backing that are typically included in LTA validation research for these three inferences.

The *scoring or evaluation* inference claims that test-taker observations are evaluated using procedures that result in observed scores with intended characteristics. The focus of this inference is on ensuring that the methods that lead from the performance to the score are based on current best practices, including evaluating whether the test item properties are acceptable, and that the test tasks spread test takers into the desired levels for decision-making. The inference is also concerned with the internal consistency of rater-mediated scoring and the appropriateness of the test administration procedures.

The underlying claim of the *generalisation* inference is that the observed scores are estimates of the expected scores over comparable tasks, test forms and raters. Backing for this inference examines whether parallel versions of the test are equivalent, whether sufficient tasks, items and scorers are included and whether scoring across test administrations is consistent.

Finally, the *extrapolation* inference examines whether the operationalised construct of the assessment sufficiently accounts for the quality of the performance in the target domain. This inference is different from the domain description inference in that it examines whether an already established assessment sufficiently accounts for the quality of performances required in the domain. More specifically, backing for the domain description inference is usually collected prior to test development while backing for the extrapolation inference occurs when the assessment is operational and hence allows for the comparison between test performance and real-world behaviour.[15]

Table 1 (based on Chapelle[6]; Dai[16]; Knoch & Macqueen[14]) summarises the seven inferences described above as well as the location of typical assumptions and sources of backing across the various inferences, as they are usually used in the LTA validation literature. Importantly, the framework is not fixed but allows researchers to select or create new assumptions and sources of backing appropriate to their context and purpose of test use. The match between inference and assumptions and related backing, however, is relatively stable in LTA research, which facilitates comparison between and evaluation of different validation studies. Here, we focus on assumptions and backings commonly used for large-scale, summative, high-stakes, standardised assessments. We acknowledge that local, formative classroom assessments can equally adopt the argument-based validation framework although their backings will be more qualitative and less measurement-centred (see Kane and Wools[17] for a discussion on argument-based validation in classroom assessments and see Gu,[18] e.g. of such validation projects in LTA).

TABLE 1 Key inferences, sample assumptions and sources of backing in LTA validity arguments.

Inference	Sample assumptions	Possible sources of backing
Domain description Claim: the selection, design and delivery of the test tasks take the relevant target domain into account.	Assessment tasks mirror the domain	Domain analysis, interviews and surveys with domain insiders
	Assessment tasks sufficiently capture the domain	Domain analysis, interviews and surveys with domain insiders
Evaluation or Scoring Claim: Test-taker observations are evaluated using procedures that result in observed scores with intended characteristics.	Test administration procedures are appropriate	Review of test conditions, interviews with test takers
	Statistical characteristics of items are acceptable	Statistical analyses, including item analyses or Rasch analyses
	Test tasks are able to spread test takers into appropriate levels for decision-making	Statistical analyses, including Rasch analyses
	Test tasks provide sufficient opportunity for test-takers to display their ability	Interviews with stakeholders, review of test materials, including response characteristics
	Scoring of performances is reliable.	Statistical analyses
Generalisation Claim: The observed scores are estimates of the expected scores over the relevant parallel versions of the tasks, test forms and across raters.	The number of tasks, items and scorers included is sufficient to arrive at a reliable score	G-theory studies
	Appropriate equating procedures for test scores are used	Review of test development documentation, evaluation of scaling methods
	No construct-irrelevant variance is introduced owing to administration conditions	Monitoring of statistical results and administration conditions
	No construct-irrelevant variance is introduced owing to candidate characteristics, such as gender, first language etc.	DIF studies
	Scoring is consistent across test administration	Statistical analyses
Explanation Claim: Expected scores are attributed to the theoretical construct the assessment is designed to measure.	The knowledge, processes and strategies required to complete test tasks vary in keeping with theoretical expectations	Verbal protocols of test takers; discourse studies
	The internal structure of the test scores is consistent with a theoretical view of the skills required in the context	Factor analyses
	Performances on tasks are related to performances on other, related tasks	Correlational analyses
Extrapolation Claim: The operationalised construct of the assessment sufficiently accounts for the quality of the performance in the target domain.	Performance on the assessment is related to performance in the domain	Interviews with domain insiders
	Test taker's interaction with the test tasks is similar to performance in the domain or on similar tasks deemed good proxies of the domain	Interviews with domain insiders; discourse analysis of discourse from test and real-world domain
Decisions or Utilisation Claim: Decisions made based on the estimates of the quality of the performance collected from the test takers are appropriate and well communicated.	The assessment differentiates test takers into appropriate levels for decision-making	Statistical analyses
	Cut-scores are defensible	Review of standard-setting methodology; interviews with test users
Consequences Claim: Test consequences are beneficial to test users in the domain and not harmful to test takers.	Test takers who have passed the assessment have sufficient ability to function in the domain	Interviews with test takers, employers
	The format of the assessment provides an appropriate model of communication in the domain	Interviews with test takers
	The test scores and/or feedback have a positive influence on teaching, learning and motivation	Interviews with stakeholders

3 | COMPARING ARGUMENT-BASED VALIDATION RESEARCH IN HPE AND LTA

Having set out argument-based validation in LTA, we selected six studies in the HPE literature that drew on the argument-based approach to validation.[3,19–23] We are cognizant of the fact that the studies chosen are more measurement-driven, summative assessments than functional-driven, formative assessments.[17] Our rationale for this is that we aimed to examine HPE studies that collected both quantitative and qualitative backings to articulate their validation arguments. Since formative, classroom-based assessments allow for mostly qualitative backings,[24] we believe the studies chosen in this paper offer a more comprehensive presentation of the range of backings that can be used to substantiate a validity argument, which is one of the aims of this methodology demonstration paper. Our second consideration in choosing these six studies is that they provide a wide spectrum of HPE assessment contexts, including laboratory-based,[3] simulation-based[19,22] and residence-based[23] assessment. Thirdly, we purposefully chose these six studies as they assessed different constructs, such as procedural skills,[3] telehealth communication skills[21] and interprofessional collaborative skills.[20] To sum up, these three considerations shaped our purposive sampling, with a view to making this methodology display paper more useful and relevant to HPE researchers working with diverse assessment contexts and constructs.

Table S1 presents our analysis of the six sample studies. We note that all six HPE studies included all of Kane's original four inferences (scoring or evaluation, generalisation, extrapolation and implications). We have, however, observed inconsistency in where various aspects of validation endeavours are housed in relation to the four inferences across the selected six studies. To make this clearer, Dai and Knoch, who have a background in LTA, independently coded each backing in the six studies based on which inference language assessors would associate it with, using the seven-inference coding scheme routinely employed in LTA validation: **DOM** (Domain description), **EVA** (Evaluation or scoring), **GEN** (Generalisation), **EXP** (Explanation), **EXT** (Extrapolation), **DEC** (Decisions or Utilisation) and **CON** (Consequences). Vu, Lim, Malone and Mak, who have a background in HPE, reviewed Dai's and Knoch's coding. Differences in coding were resolved through discussion and all six authors agreed with the final coding as presented in Table S1. Throughout our interactive coding and discussions, we observed that the unique composition of our team, consisting of both HPE and LTA researchers, allowed for greater reflexivity and more robust interdisciplinary conversations.

Zooming in on Table S1, we can see various examples of inconsistency in these select HPE studies as to where different aspects of validation endeavours are housed in relation to the inferences. For example, aspects of inquiry relating to the domain, which LTA researchers generally locate in the domain description inference, could be found in generalisation (e.g. in Daniels and Pugh[22] who described drawing on blueprints to ensure sampling from the domain) or in extrapolation (e.g. in Cook et al[3] who described using a needs analysis to define the domain).

Similarly, questions relating to score generalisation across raters, parallel forms of an assessment and different assessment conditions are generally located in the generalisation inference in LTA research. Such aspects were variously identified in the HPE literature in the scoring inference (e.g. Fraser et al[20]), generalisation inference (various studies) or implications inference.[3,20] The scoring or evaluation inference in the HPE literature focussed on the types of aspects listed in Table 1 but also encompassed the equivalence of test forms (Cook et al[3]) and G-theory studies (Fraser et al[20]), which in the LTA literature fall under the generalisation inference.

Investigations that LTA researchers would group into the explanation inference, such as examinations of the internal structure of an assessment (e.g. by drawing on structural equation modelling, confirmatory factor analysis, MTMM), were found to be variously located in the extrapolation inference (e.g. in Cook et al[3]; Hess et al[21]) or the generalisation inference (e.g. Fraser et al[20]).

Considerations around setting and reviewing the appropriateness of pass-fail standards are generally placed into the decision inference in LTA. In the HPE literature, we reviewed, we found these mentioned in the scoring or evaluation inference[3,23] or the implications inference.[20–22]

In view of these inconsistencies, we reorganised the backings in each study by presenting them using the LTA seven-inference structure in Table S2. We can see in Table S2 that the expanded Kanean framework allows for easier comparison and evaluation between studies in terms of the inferences investigated, the sources of backing gathered and whether any inference needs more substantiation.

More specially, we argue the seven-inference framework in LTA can contribute to HPE assessment by further improving the rigour, justifiability and positive washback of HPE assessments. The need to establish evidence for the additional *domain description* inference encourages and reinforces the well-recognised practice in HPE of creating an assessment blueprint and defining the scope of assessment through, e.g. interviews with expert content leads and practitioners. The *explanation* inference necessitates more intentional use of evidence-informed conceptualisation of the assessed competencies (e.g. clinical communication skills, empathy, etc.) to ascertain that the task requirement and the assessment criteria in the marking rubric reflect how the competencies are theoretically validated in the literature. Finally, splitting the original Kanean *implications inference* into *utilisation* or *decisions* and *consequences* inferences provides a more guided structure, encouraging HPE assessors to methodically consider concurrently (i) how the evaluation of the student performance is used and (ii) what effects their assessment generates. HPE has long regarded it crucial to make necessary and defensible assessment decisions related to students' learning, progression and readiness for practice.[3,25] Defining a specific utilisation or decision inference can allow HPE to formalise this central consideration in their assessment validation practice. Similarly, the consensus framework for good assessments in HPE[26] contends that good assessment needs to have a catalytic effect that "motivates all stakeholders to create, enhance, and support education" and "drives future learning forward and improves overall program quality". By having a specific consequences

inference, HPE researchers can now more explicitly and proactively evaluate if their assessment has resulted in a positive washback on students and stakeholders.

Having explicated the advantages of a seven-inference structure and its potential contributions to HPE validation, we emphasise that the current seven-inference structure is not without its limitations. LTA researchers have raised, e.g. how the current Kanean argument framework does not account for test practicality.[16,27] An impractical test that is difficult or resource-intensive to deliver, no matter how valid it is, is unlikely to gain traction in the real world. Another related consideration is stakeholder assessment literacy. We can gather sufficient backings to support the validity argument of an instrument but if no effort is put into developing end-users' appreciation of the value or benefit of this instrument, the instrument again is not likely going to be taken up. Dai[16] offers a further discussion on these considerations, taking into account the practical, logistic and commercial dimensions of assessment. Future research in HPE and LTA validation can explore, potentially collaboratively, opportunities in refining the argument-based validation framework to account for these issues.

4 | CONCLUSION

As Kane argues in his 2013 paper, the purpose of an argument-based approach to validation is to "provide a framework for the evaluation of the claims based on the test scores".[13] It is, therefore, crucial that HPE establishes a shared disciplinary framework to approach validation, especially when test-score claims from HPE assessments are often used for high-stakes decisions such as health professional registration and practice. To promote this undertaking, this Cross-Cutting Edge paper introduced the expanded seven-inference Kanean validation framework that has been tried and tested over decades in the field of LTA validation. We presented the conceptual development behind the expansion of the original four-inference structure to its current seven-inference framework in LTA, explicating how each of the seven inferences is defined along with their related claims and sources of backing. We argue that an expansion of the number of inferences typically used in HPE validation arguments and a more standard approach to the location of various aspects of validation may help scholars and practitioners in HPE to create clearer, and more tightly specified validation arguments. This may help newcomers to the discipline gain clarity about the types of questions they would pose about their own assessment and score use contexts. A clearer, more tightly specified structure of validation research will also make it easier for outsiders to evaluate the robustness of validation activities. We conclude this paper with a call for greater cross-disciplinary dialogue and collaboration between HPE and LTA, as both disciplines share central concerns in effective communication and justifiable test use.

AUTHOR CONTRIBUTIONS

David Wei Dai: Conceptualization; methodology; data curation; investigation; formal analysis; visualization; project administration; writing—original draft; writing—review and editing. **Thao Vu:** Conceptualization; investigation; writing—original draft; methodology; visualization; writing—review and editing; formal analysis; project administration; data curation. **Ute Knoch:** Conceptualization; investigation; writing—original draft; visualization; writing—review and editing; methodology; formal analysis; project administration; data curation. **Angelina S. Lim:** Writing—original draft; writing—review and editing; visualization. **Daniel Thomas Malone:** Writing—review and editing; writing—original draft; visualization. **Vivienne Mak:** Visualization; writing—original draft; writing—review and editing.

ACKNOWLEDGMENTS

We are grateful for the detailed and constructive feedback from anonymous reviewers, Prof John Norcini and Prof Kevin Eva. We would like to mention that Dai and Vu are joint first authors for this paper.

CONFLICTS OF INTEREST STATEMENT

The authors declare that they have no competing interests.

ETHICAL STANDARDS

The work reported in this paper did not involve primary data collection so ethical approval was not required.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

David Wei Dai https://orcid.org/0000-0002-3575-131X
Thao Vu https://orcid.org/0000-0001-8275-1083
Ute Knoch https://orcid.org/0000-0002-3306-337X
Angelina S. Lim https://orcid.org/0000-0002-8219-1191
Daniel Thomas Malone https://orcid.org/0000-0002-4838-8441
Vivienne Mak https://orcid.org/0000-0002-1325-5809

TWITTER

David Wei Dai *drdavidweidai*

REFERENCES

1. Messick S. Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher*. 1989;18(2):5-11. doi:10.2307/1175249
2. Kane M. Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy & Practice*. 2016;23(2):309-311. doi:10.1080/0969594X.2016.1156645
3. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49(6):560-575. doi:10.1111/medu.12678
4. Schuwirth LW, van der Vleuten CP. Programmatic assessment and Kane's validity perspective. *Med Educ*. 2012;46(1):38-48. doi:10.1111/j.1365-2923.2011.04098.x
5. Kinnear B, Schumacher DJ, Driessen EW, Varpio L. How argumentation theory can inform assessment validity: a critical review. *Med Educ*. 2022;56(11):1064-1075. doi:10.1111/medu.14882
6. Chapelle CA. The TOEFL validity argument. In: Chapelle CA, Enright MK, Jamieson JM, eds. *Building a Validity Argument for the*

Test of English as a Foreign Language. Routledge; 2011:319-352. doi:10.4324/9780203937891-15
7. Kane MT. Validation. *In: Brennan R Educational Measurement American Council on Education and Praeger*. 2006:17-64.
8. Bachman LF, Palmer AS. *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford University Press; 1996.
9. Chapelle CA. *Argument-Based Validation in Testing and Assessment*. Sage Publications; 2020.
10. Chapelle CA, Voss E (Eds). *Validity Argument in Language Testing: Case Studies of Validation Research*. Cambridge University Press; 2021. doi:10.1017/9781108669849
11. Dai DW. What do second language speakers really need for real-world interaction? A needs analysis of L2 Chinese interactional competence. *Lang Teach Res*. 2023;1-38. doi:10.1177/13621688221144836
12. Youn SJ. Task-based needs analysis of L2 pragmatics in an EAP context. *Journal of English for Academic Purposes*. 2018;1(36): 86-98. doi:10.1016/j.jeap.2018.10.005
13. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013;50(1):1-73. doi:10.1111/jedm.12000
14. Knoch U, Macqueen S. *Assessing English for professional purposes*. Routledge; 2019.
15. Lim A, Krishnan S, Singh H, et al. Linking assessment to real life practice–comparing work based assessments and objective structured clinical examinations using mystery shopping. *Adv Health Sci Educ*. 2023;1-20. doi:10.1007/s10459-023-10284-1
16. Dai DW. *Assessing interactional competence: principles, test development and validation through an L2 Chinese IC test*. Peter Lang; 2024.
17. Kane MT, Wools S. Perspectives on the validity of classroom assessments. In: Brookhart SM, McMillan JH, eds. *Classroom Assessment and Educational Measurement*. Routledge; 2019:11-26. doi:10.4324/9780429507533-2
18. Gu PY. An argument-based framework for validating formative assessment in the classroom. *Frontiers in Education*. 2021;6:1-10. doi:10.3389/feduc.2021.605999
19. Tavares W, Brydges R, Myre P, et al. Applying Kane's validity framework to a simulation based assessment of clinical competence. *Adv Health Sci Educ Theory Pract*. 2018;23(2):323-338. doi:10.1007/s10459-017-9800-3
20. Fraser KL, Charania I, Hecker KG, et al. Summative assessment of interprofessional "collaborative practice" skills in graduating medical students: a validity argument. *Acad Med*. 2020;95(11):1763-1769. doi:10.1097/acm.0000000000003176
21. Hess BJ, Kvern B. Using Kane's framework to build a validity argument supporting (or not) virtual OSCEs. *Med Teach*. 2021;43(9):999-1004. doi:10.1080/0142159X.2021.1910641
22. Daniels VJ, Pugh D. Twelve tips for developing an OSCE that measures what you want. *Med Teach*. 2018;40(12):1208-1213. doi:10.1080/0142159x.2017.1390214
23. Poudeh MD, Mohammadi A, Mojtahedzadeh R, Yamani N, Delavar A. Providing a model for validation of the assessment system of internal medicine residents based on Kane's framework. *J Educ Health Promot*. 2021;10(1):386. doi:10.4103/jehp.jehp_1500_20
24. Hopster-den Otter D, Wools S, Eggen TJ, Veldkamp BP. A general framework for the validation of embedded formative assessment. *J Educ Meas*. 2019;56(4):715-732. doi:10.1111/jedm.12234
25. Bowe CM, Armstrong E. Assessment for systems learning: a holistic assessment framework to support decision making across the medical education continuum. *Acad Med*. 2017;92(5):585-592. doi:10.1097/ACM.0000000000001321
26. Norcini J, Anderson MB, Bollela V, et al. Consensus framework for good assessment. *Med Teach*. 2018;40(11):1102-1109. doi:10.1080/0142159X.2018.1500016
27. Roever C, Fraser C, Elder C. *Testing ESL Sociopragmatics: Development and Validation of a Web-Based Test Battery*. Peter Lang; 2014. doi:10.3726/978-3-653-04598-7

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Dai DW, Vu T, Knoch U, Lim AS, Malone DT, Mak V. Expanding Kane's argument-based validity framework: What can validation practices in language assessment offer health professions education? *Med Educ*. 2024;1-7. doi:10.1111/medu.15452