Language testing in changing times: An interview with Professor Daniel Isbell

Edward Schaefer[1] and Jeffrey Martin[2]
[1]edwards16@me.com
[2]jeffmjp@gmail.com
Ochanomizu University
Momoyama Gakuin University

Keywords: language testing, validity, validity argument, ethics, questionable research practices, feedback, diagnostic assessment

Daniel R. Isbell is an Assistant Professor in the Department of Second Language Studies at the University of Hawai‘i at Mānoa, where he teaches courses and supervises MA and PhD students in language assessment. He serves on the the editorial boards of the journals *Language Testing* and *TESOL Quarterly*. His primary research interest is language assessment, and he has conducted research on diagnostic assessment, self-assessment, rater effects, repeated test taking and proficiency development, test equating, specific purposes language testing, and vocabulary in context. His interests also include instructed second language acquisition, ranging from L2 pronunciation to language learning apps. It is the pleasure of the Testing and Evaluation SIG of JALT to sponsor Dr. Isbell as one of the plenary speakers at JALT PanSIG 2024.

The following interview was conducted via email correspondence to allow for a thoughtful exchange of ideas and in-depth responses to each question.

Professor Isbell, thank you for agreeing to do this interview. We'd like to start by asking you about your keynote address at PanSIG 2024. Can you give us an idea of the things you'll be discussing?

It's my pleasure to do this interview and deliver a keynote at PanSIG 2024! I have greatly enjoyed engaging with JALT's Vocabulary SIG recently, and I am very much looking forward to the PanSIG next year. I'd like to thank TEVAL for inviting me.

Language testing has really changed a lot over the last 5 to 10 years, with technology playing a much bigger role. Artificial intelligence (AI), at-home testing, automated scoring, computerized delivery of multimodal test tasks, and remote proctoring have been increasingly integrated into a wide range of language assessments, including most notably large-scale standardized tests, but we also see classroom teachers using more of these technologies, too. It is all quite exciting and also overwhelming! The conference theme of PanSIG 2024 is "Getting Back to Basics", which I plan to address by talking about how assessment basics, like validity, fairness, and practicality, can help us grapple with and appropriately evaluate major, technology-driven developments in language testing. There's a lot of hype around things like AI, but as language professionals we shouldn't lose focus of basic principles when making decisions about developing or using assessments.

One of your interests is the teaching and assessment of pronunciation, and you developed the Korean Pronunciation Diagnostic, or KPD, for your dissertation. Although it's designed as a low-stakes diagnostic assessment, you based it on a rigorous validity argument, following Kane and Chapelle. Many teachers are responsible for creating classroom assessments, but for readers who may not be familiar with the concept of validity as argument, can you say a few words about the desirability of basing even low-stakes tests on a structured validity argument?

So first, a bit about validity. Validity is something most people understand as a property of a test; people will say things like "TOEIC is not valid" or "EIKEN is valid." If you ask most people to define validity, they'd probably say things like "how accurate the test is" or "whether the test is a good measurement of something." But in language testing, and educational assessment more broadly, validity has a broader scope and includes the *use* of test scores. Our updated, more specialized definition of validity is something like *the degree to which test scores reflect the targeted knowledge, skill, or ability and are relevant and useful for making a specific decision*. So, while a test might provide a good measure of some ability, we have to think beyond that and evaluate whether that test score is appropriate for making a specific decision about a learner (e.g., what class they should be placed in, whether they have sufficient language ability to handle studying full-time in L2-medium classes).

This expanded conceptualization of validity can be hard to grapple with in practice – there's so much ground to cover. This is where validity arguments come in. Validity arguments are a framework for evaluating whether the interpretations and uses of test scores are justified. They consist of a series of inferences that we have to make when interpreting and using test scores, and each inference requires some evidence, or backing, for us to be able to accept it and advance the argument. So for example, one inference is *generalization*, which has to do with the consistency of test scores across possible conditions (e.g., different test forms, taking a test on two different days). Reliability analyses provide backing to support the inference

that *test scores are consistent*. If test scores are not consistent, it makes it hard to move the argument forward to start talking about the meaning of test scores – if you could get very different test scores on different forms of the same test, inferring something about what your test score means in terms of your English ability is premature. Validity arguments extend to test use, too, and consider inferences like *extrapolation*, the idea that test scores are predictive of performance in the real world, and *consequence implication*, the idea that using test scores to make decisions leads to desirable outcomes (e.g., student success in university, positive washback on language teaching).

Some large-scale test developers create validity arguments (Educational Testing Service, for example, but admittedly validity arguments aren't necessarily things that teachers need to create for every low-stakes test they create or use in the classroom. Kane (2013) pointed out that lower stakes demand less rigorous support. I think it's also fair to say that for most lower stakes tests, and especially those that will not be used very widely, it's okay for validity arguments to be more or less informal – most likely not even written down.

For diagnostic tools that are developed by specialists and intended to be used more broadly, however, I do think a validity argument is useful even though the stakes may be low. Validity arguments help clarify the reasoning behind test score interpretation (i.e., what a test score means about a person) and test score use (i.e., how we use a test score to make decisions and how well those decisions pan out). Validity arguments also help identify information necessary to support interpretations and uses of test scores, and in that way they double as a research agenda, helping test developers and researchers identify areas worthy of investigation. For diagnostic tools in particular, I think it's quite important for prospective users to know whether there is potential for using the scores/diagnostic feedback to support learning, which relates to inferences that assessment researchers refer to as *utilization* and *consequence implication* (or sometimes *impact*). Otherwise, there's little point in diagnosing strengths and weaknesses in the first place.

Due to the pandemic, some U.S. universities started accepting the online Duolingo English Test (DET) as an admissions test for L2 students. You did a study of the relationship between DET speaking scores and university stakeholders' evaluations of DET speaking performances. Can you tell us a little about how you applied a validity argument in order to extrapolate from DET test scores to the Target Language Use Domain (TLU), that is, success at an English medium university?

In that study (Isbell et al., 2023), we collected data relevant to what is called an *extrapolation* inference in a validity argument. This inference connects test performance to performance in the 'real world' TLU Domain. For us, we wanted to see whether DET scores and speaking performances aligned with the expectations of listeners in the real-world context of an English medium university. So, we recruited faculty, administrative staff, graduate students, and undergraduate students at our university to judge speaking task performances elicited by the DET. This approach draws heavily on the idea that 'linguistic laypersons', or non-linguists, make intuitive judgments about language performance based on their knowledge of community expectations and their own life experiences (Sato & McNamara, 2019). What we found was that test takers with higher DET scores were generally perceived by our listeners as being more comprehensible and more capable of handling increasingly demanding roles in a university – people with higher DET scores were perceived as being more capable of handling graduate studies or teaching courses.

This study does have some important limitations, though. Our approach to accessing the TLU domain was indirect, with listeners judging test-based speaking performances rather than authentic TLU performances (e.g., a real office hours interaction or class presentation). But it does lend some support to the inference that DET scores indicate differences in speaking ability that are relevant to the expectations of academic study – a preliminary but important step given that the automated scoring and simple nature of DET tasks has led to some very reasonable skepticism (see Wagner's 2020 review of the test).

Another one of your interests is ethics in language testing. In 2023, for example, you published a study reviewing developer involvement in high-stakes English proficiency tests (HSEPTs), noting that many published studies did not provide conflict of interest statements (COIs), such as ETS financing studies on TOEIC or TOEFL. You argue that the absence of COIs in published studies is an ethical lapse in the field of language testing. Can you explain what COI statements are and why they are important?

We generally expect research to provide an objective, disinterested account of some phenomenon. Conflict of interest statements are brief statements of a researcher's associations that could otherwise influence their scholarly work. Associations can be personal (e.g., a researcher's spouse or family member works for a test developer), professional (e.g., a researcher works at the university which develops and uses a placement test), or financial – it is typically the latter that comes up most frequently. Financial relationships can include regular employment (e.g., a British Council employee researching IELTS), ownership/investment (e.g., a researcher owning a part of a test developer), and other kinds of financial benefits (e.g., receiving honoraria or consulting fees). One important thing to note about COIs is that we are really talking

about *potential* COIs. We can't prove or disprove that someone's associations caused them to do anything dishonest or otherwise influenced how they conducted research and reported the results, but transparently disclosing potential COIs helps readers digest research with appropriate skepticism/criticality.

The statement for a given article is typically limited to entities addressed in the article. I've included an example disclosure statement at the end of this interview. As I have mentioned several tests and test developers, I have disclosed my relationships with relevant parties.

I believe COIs matter in language testing because test developers generally have an interest in research results that support their tests. The commercial incentives are obvious; research that strongly supports the use of an HSEPT could influence governments, universities, and other test users to accept the test, which in turn could lead to more sales. Vice versa, research findings that cast doubt on a test could have negative commercial impacts. Although research funding programs are intended to generate independently obtained evidence related to the quality or use of a test, test developers are ultimately the ones deciding which studies get funded and sometimes do (very much necessary) behind the scenes work like supply data for researchers to analyze.

In my view, it's not just commercial interests at play. Frankly, it's hard for even non-commercial test developers, such as people who develop and research placement tests for university language programs, to really be able to claim total objectivity. Test developers take pride in their work, which is a good thing, and may also want to be seen as competent professionals by their peers; research showing that their tests work well can demonstrate their skill as developers.

Some people feel that COIs in language testing research are not really needed because the relationships are obvious. But I don't feel the same way; while the connection between Duolingo and the Duolingo English Test is obvious, the connection between the Center for Applied Linguistics and the WIDA ACCESS for ELLs (a test used in U.S. public schools) is less so, especially to 'outsiders' who do not know the language testing industry very well. Peer-reviewed language testing research should strive to be something that can inform real-world test use where relevant, and to that end non-specialists should be made clearly aware of potential COIs.

In a related area, you conducted a survey of applied linguistics researchers asking them about questionable research practices (QRP). You note that QRPs exist in an ethical gray area and the label of "QRP" should not necessarily be seen as a marker of individual malintent, and you contrast this with deliberate research misconduct. You found that virtually all applied linguists who completed your survey admitted to at least one QRP (94%) and that 17% admitted to research misconduct at least once. These are disturbing findings and don't reflect well on our profession. Can you tell us the difference between QRPs and research misconduct and comment on your findings?

Research misconduct is clearly unethical, fraudulent behavior that comprises practices such as fabrication (making up data or results entirely), falsification (altering research materials, including data, with the intent to deceive), and plagiarism.

In contrast, we hope that researchers engage in research responsibly by conducting research carefully, in accordance with best practices, and transparently reporting findings. Questionable research practices are the gray area between best practices and outright misconduct. QRPs are not inherently unethical. Sometimes we engage in QRPs because there is no consensus around best practices, leading reasonable people to do things that their peers might not agree with. Sometimes QRPs happen due to carelessness or ignorance, too – it's important to keep in mind that researchers are only human. However, QRPs can also be deliberately exploited to specific ends, like obtaining a result that is statistically significant, which might cross the line into falsification in some cases (if we could somehow know the researcher's true intent, that is).

So why do misconduct and QRPs matter? Published research that is based on misconduct like fabrication of data or falsification of results is misinformation, and it can distort our understanding of language learning and teaching. This kind of fraudulent research seems fairly rare, thankfully, but it does show up every once in a while, so I think we do need to be on guard. QRPs are much more common, as our study (Isbell et al., 2022) showed. Like misconduct, QRPs can potentially distort our understanding. One common QRP is excluding non-significant findings from a study. Let's say you conduct a study on a hot topic like ChatGPT in L2 writing instruction, and you want to examine grammatical accuracy, lexical diversity, and fluency (length). Compared to a control condition, you might find that students who are allowed to use ChatGPT produce texts with significantly fewer grammatical errors. So you report those findings, which are quite exciting, and maybe you can get it published. But let's say you also examined fluency, in terms of text length, and lexical diversity, and found that there weren't any significant differences between the ChatGPT and the control condition. If you don't report those findings, too, you end up with an unbalanced picture.

In the context of research in Diagnostic Language Assessment (DLA), you highlighted the importance of actionable feedback, emphasizing its adaptability in both quantitative and qualitative formats. Considering the diverse spectrum of L2 learners in terms of proficiency, identifiable strengths and weaknesses, age, language learning goals, and motivation,

individualized feedback becomes crucial. Could you provide insights into which types of learners stand to benefit the most from diagnostic feedback, given these varying factors?

This is a difficult question, and one that, to my knowledge, there's not a clear, universal answer from empirical research. So, my answer should be taken with a grain of salt.

One thing that seems to be important is the ability to understand and 'digest' the feedback. Feedback that is rather granular and detailed can be quite helpful, but it might be inaccessible if it is delivered in the target language rather than a learner's L1 (or other proficient language). Also important is the motivation of the learner – not so much in capital-M motivation associated with big-picture theories, but specific motivation to really engage with feedback provided by a diagnostic procedure and do the work to address any weak areas. It is often the case that learners will look at some feedback (whether it is diagnostic results or any other kind of feedback) but only pay attention to overall results and quickly move on, as they might lack a specific drive or desire to really work on specific language features.

Harding et al., (2015) and Isbell (2021) propose that for a test to be diagnostic, its results need to be in the form of feedback that is relevant and actionable in subsequent L2 learning. Concerning the question of which language skills or competencies are able to be diagnosed, first, suppose a framework for language use was applied, such as Bachman's model of Communicative Competence, what skills or competencies do you think are currently best diagnosable? What are examples of such tests that currently exist? Second, to what extent could diagnostic tests be developed to assess and provide feedback for the other components that are not yet well tested, according to language models?

I am very much sympathetic to Harding et al.'s (2015, and Alderson et al., 2015) argument that subcomponents of communicative language skills, particularly those that can be assessed in a discrete, granular manner, are most readily diagnosable. Their view clearly influenced my design for the KPD, which drills down into the perception and production of individual phonemes. While L2 pronunciation (and speaking ability more broadly) depends on more than just segmental aspects of pronunciation, phonemes do matter in every spoken utterance. Ideally, there would be other diagnostic tools to dig deeper into specific weaknesses of suprasegmental pronunciation features, but I had to start somewhere.

Clark and Endres (2021) is a nice example of a diagnostic assessment of English grammar targeted at the A2 proficiency level. Grammar, by itself, is not really so important as a communicative language learning outcome – except that it is something we draw on *constantly* when using language to communicate. Hence the inclusion of Grammatical Competence in Bachman's model. So Clark and Endres' grammar diagnostic, which is all presented in the written modality, is something that might help teachers and learners understand in greater detail why they have difficulty understanding when reading or have difficulty expressing some ideas clearly when writing.

I do think there's room to expand on diagnostic assessment practices through more formalized screening or observation procedures. Alderson and colleagues do discuss this and allude to some ways that less formal observations can motivate more detailed diagnostic tools being used, but in my own research and experience (including supervising student projects related to DLA), the first step of figuring out who might benefit from additional diagnostic procedures is really key. So in this area, coarser grained diagnostic screening/observation of more communicative, even integrated skills seems necessary and, I think, is quite possible.

Thank you for taking the time to answer these questions. We're looking forward to hearing your keynote address at the JALT PanSIG in May 2024.

Declaration of competing interests:
D. Isbell has received research funding from British Council, Duolingo, Educational Testing Service, and Pearson, honoraria from Educational Testing Service, and has consulted for IELTS UK and Duolingo.

References

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a Theory of Diagnosis in Second and Foreign Language Assessment: Insights from Professional Practice Across Diverse Fields. *Applied Linguistics*, *36*(2), 236–260. https://doi.org/10.1093/applin/amt046

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press.

Clark, T., & Endres, H. (2021). Computer-based diagnostic assessment of high school students' grammar skills with automated feedback – an international trial. *Assessment in Education: Principles, Policy & Practice*, *28*(5–6), 602–632. https://doi.org/10.1080/0969594X.2021.1970513

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*(3), 317–336. https://doi.org/10.1177/0265532214564505

Isbell, D. R. (2021). Can the Test Support Student Learning? Validating the Use of a Second Language Pronunciation Diagnostic. *Language Assessment Quarterly*, 1–26. https://doi.org/10.1080/15434303.2021.1874382

Isbell, D. R., Crowther, D., & Nishizawa, H. (2023). Speaking performances, stakeholder perceptions, and test scores: Extrapolating from the Duolingo English test to the university. *Language Testing*. Online advance publication. https://doi.org/10.1177/02655322231165984

Isbell, D. R., Brown, D., Chen, M., Derrick, D. J., Ghanem, R., Arvizu, M. N. G., Schnur, E., Zhang, M., & Plonsky, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in applied linguistics. *The Modern Language Journal*, 24. https://doi.org/10.1111/modl.12760

Isbell, D. R., & Kim, J. (2023). Developer involvement and COI disclosure in high-stakes English proficiency test validation research: A systematic review. *Research Methods in Applied Linguistics*. https://doi.org/10.1016/j.rmal.2023.100060

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores: Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Sato, T., & McNamara, T. (2019). What Counts in Second Language Oral Communication Ability? The Perspective of Linguistic Laypersons. *Applied Linguistics*, *40*(6), 894–916. https://doi.org/10.1093/applin/amy032

Wagner, E. (2020). Duolingo English Test, Revised Version July 2019. *Language Assessment Quarterly*, 1–16. https://doi.org/10.1080/15434303.2020.1771343