Future challenges and opportunities in language testing and assessment: Basic questions and principles at the forefront

Tineke Brunfaut, Lancaster University

Abstract

In this invited Viewpoint on the occasion of the 40th anniversary of the journal Language Testing, I argue that at the core of future challenges and opportunities for the field – both in scholarly and operational respects – remain basic questions and principles in language testing and assessment. Despite the high levels of sophistication of issues looked into, and methodological and operational solutions found, outstanding concerns still amount to: what are we testing, how are we testing, and why are we testing? Guided by these questions, I call for more thorough and adequate language use domain definitions (and a suitable broadening of research and testing methodologies to determine these), more comprehensive operationalisations of these domain definitions (especially in the context of technology in language testing), and deeper considerations of test purposes/uses and of their connections with domain definitions. To achieve this, I maintain that the field needs to continue investing in the topics of validation, ethics, and language assessment literacy, and engaging with broader fields of enquiry such as (applied) linguistics. I also encourage a more synthetic look at the existing knowledge base in order to build on this, and further diversification of voices in language testing and assessment research and practice.

Title and abstract in Dutch

Toekomstige uitdagingen en mogelijkheden voor taaltoetsing: Basisvragen en -beginsels voorop

In dit opiniestuk ter gelegenheid van het veertigjarig bestaan van het wetenschappelijk tijdschrift Language Testing beargumenteer ik dat de basisvragen en -beginsels van taaltoetsing nog steeds de kern vormen van de toekomstige uitdagingen en mogelijkheden voor het vakgebied, op vlak van zowel onderzoek als testpraktijk. Ondanks het feit dat vele complexe en gedetailleerde onderwerpen reeds onderzocht zijn en dat methodologische en praktische oplossingen gevonden zijn, komen de nog bestaande vraagstukken neer op: wat toetsen we, hoe toetsen we en waarom toetsen we? In het licht van deze vragen roep ik op tot volledigere en adequatere definities van taalgebruiksdomeinen (en een verbreding van onderzoeks- en toetsmethodes om dit te kunnen doen), meer omvattende praktijkvertalingen van deze domeindefinities (vooral met betrekking tot technologie in taaltoetsing), en diepgaandere overwegingen over toetsdoeleinden en toetsgebruik en over hun verband met domeindefinities. Om dit te kunnen verwezenlijken pleit ik ervoor dat het vakgebied blijft investeren in de onderwerpen van validering, ethiek, and taaltoetsgeletterdheid en op de hoogte blijft van ontwikkelingen in bredere onderzoeksvelden zoals (toegepaste) taalkunde. Ik roep onderzoekers ook op tot het verwerven van een meer omvattend overzicht van bestaande onderzoeksresultaten en tot een grotere verscheidenheid aan stemmen binnen het taaltoetsonderzoek en de taaltoetspraktijk.

Briefly looking back

By now, language testing and assessment has firmly developed into a mature, well-established and recognized scholarly and operational field, with a historical track record. Without doubt, the journal Language Testing has played an important role in the development of the field over the past 40 years, not in the least by publishing the latest theoretical and empirical work, and enabling scholarly exchange. Just over 20 years ago, Lyle Bachman (2000) took stock in the journal of how the field had developed by the turn of the century and what had been achieved by then (as had other prominent language testers done in the preceding decades). Bachman also put forward two key priorities, as he saw them, for the next decade(s). While the aim of the present article is not to make an inventory of past accomplishments, it is interesting to briefly look back at Bachman's 'prophecy' and see what has materialized two decades later, as this provides some indication of what we can build on in the coming decade(s). Another reason why Bachman's (2000) state-of-the-art review is interesting, is that it was written around the time when many of those who are now considered mid-career language testers were probably only just becoming familiar with the field. While they may or may not have read Bachman's article at the time, they have increasingly helped shape the field over the past 20 years and are likely to continue for several more years (as will early and late career colleagues, of course).

A first priority identified by Bachman (2000) was professionalisation of the field, which he defined as "1) the training of language testing professionals; and 2) the development of standards of practice and mechanisms for their implementation and enforcement" (p. 19). It is reassuring that the training offer has indeed exponentially increased since 2000, in terms of number of training opportunities, content focus of training, and format of training – ranging from specialist degree programmes, summer/winter schools, workshops, research and development projects, conferences, to various types of online webinars. At the scholarly level, this has been paralleled with theoretical and empirical research interest in, and conceptualizations of, language assessment literacy. The majority of training has focused on language testing professionals – the term Bachman used, which presumably was intended to mainly comprise researchers, operational language testers, and practitioners (teachers). In recent years, we have, however, also seen calls for, and reaching out to, broader stakeholder groups (e.g. learners, parents, score users, policy makers, etc.) and the field is starting to go beyond Bachman's main training scope. It is fair to say that at least part of the training has been enabled by, and in some cases offered by or funded by, the many regional subject associations that have been set up over the last two decades (e.g. Asian Association for Language Assessment (AALA), Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ), European Association for Language Testing and Assessment (EALTA), Latin American

Association for Language Testing and Assessment (LAALTA), various country-based associations, etc.). Additionally, some language testing associations (e.g. International Language Testing Association (ILTA), EALTA) developed a code of ethics (setting out principles of ethical standards in language testing and assessment; e.g. https://www.iltaonline.com/page/CodeofEthics) and/or guidelines for good practice (outlining how to achieve minimum standards of ethical practice; e.g. https://www.iltaonline.com/page/ILTAGuidelinesforPractice and https://www.ealta.eu.org/guidelines.htm). Other associations have joined them in promoting the implementation of these codes and guidelines (e.g. United Kingdom Association for Language Testing and Assessment (UKALTA)), and a number of tests have been evaluated as attaining to minimum standards (e.g. those with an ALTE Q-mark or ICAO recognised status). Enforcement of standards, as Bachman phrased it, however, remains a debated issue.

The second priority stated by Bachman was validation research. He observed that "validation has become the de facto paradigm for language testing research and development" (Bachman, 2000, p.23) and accordingly predicted a continuation of validation work, especially framed by Messick's (1989) view of validity. Indeed, in the last two decades, we have seen the publication of seminal work to (further) clarify and structure, in practice, empirical investigations of test use and consequences, including by Lyle Bachman (assessment use argument, see e.g. Bachman, 2005), Michael Kane (interpretation/use and validity argument, see e.g. Kane, 2013), and Cyril Weir (socio-cognitive approach to test validation, see e.g. Weir, 2005). Numerous language test validation studies have been conducted, adopting one of these validation approaches and demonstrating their implementation feasibility, even while being ambitious and resource-demanding undertakings. Also, as had been hinted at by Bachman, validation research has become characterised by mixed-methods research, integrating findings resulting from multiple quantitative and qualitative data collection and analysis methods, and providing a more comprehensive and triangulated validation picture. Importantly, Bachman emphasized that these priorities – professionalization of the field and validation studies – are inextricably linked and go hand-in-hand. Namely, language assessment literacy allows for and can lead to the formulation of ethical standards, ethical standards encourage wider language assessment literacy development and validation research, and validation research requires and encourages language assessment literacy and insights into ethical standards.

Of course, apart from work directly related to these topics, a larger body of language testing work has been conducted over the past two decades. The fact that the field has been thriving, in terms of both research and development, is obvious from the expansion of the journal Language Testing, the introduction of additional language testing-focussed journals, renown textbook series, encyclopaedia editions, and the launch of many new language tests for various purposes and in various contexts – to provide just a few examples.

Looking ahead

When invited to write the present viewpoint article about challenges and opportunities in language testing and assessment for the upcoming decade(s), I found myself reflecting on a wide range of issues, but each time concluding that what was underlying them were concerns regarding the three fundamental questions: what we are testing, how we are testing, and why we are testing? Despite the continuous and many valuable innovations and improvements in language testing and assessment research and practice, the answers to these questions remain the key challenges in all the work we do. Bachman's (2000) priorities can essentially also be boiled down to these, as will become clearer below. Consequently, I have structured my views on where we may (need to) be heading as a field along these core questions. By definition, as a viewpoint article, these are just one set of views among many other possible ones, informed by my personal journey in conducting research, teaching, test development and consultancy work in the field, and selective due to article length stipulations.

A focus on 'what'

Understanding what our assessment instruments are actually testing has been regarded of utmost importance for a long time now – especially as the field moved beyond overemphasis on reliability and almost exclusive reliance on quantitative methodologies and expert judgements – not the least to detect and help avoid construct-irrelevant variance. As Alderson (2000) wrote: "what matters is not what the test constructors believe an item to be testing, but which responses are considered correct, and what process underlies them" (p. 97). From a theoretical perspective, this focus on what our tests test has been captured and further stimulated by, for example, the dimension of cognitive validity as put forward in the socio-cognitive approach to language test validation. From a methodological perspective, this was originally stimulated by an opening up to qualitative methods such as think-alouds and retrospective interviews with test-takers and raters, who voice their test completion and rating thought processing, respectively. About 10-15 years ago, language testing scholars started to employ, and triangulate this with, methods already established in cognitive sciences (e.g. psychology), such as eye-tracking and keystroke logging. Most recently, neuroimaging techniques such as functional near-infrared spectroscopy (fNIRS) have been adopted in studies on cognitive processing during language test completion. This diversification of research methods used in the field helps to increase confidence in conclusions drawn on what tests test, and to gain insight into aspects of cognitive processing that were previously difficult to access and observe. However, as we are discovering these methods and introducing them in our field, it is imperative that we do not just learn from the existing experience and expertise other fields have with these methods, but also that we critically reflect on and evaluate whether and how they may serve the questions that are key to our field and that can advance our field. The fact, for example, that cognitive research is typically conducted in highly-controlled lab settings with very restricted types of stimuli, whereas language testing typically strives for authenticity (e.g. lengthy, varied, original task input) and also involves both input and tasks (and their interaction), means that the focus of our research questions is different, and that data collection procedures and data analysis measures are not necessarily or only partially transferable to our field. A challenge therefore is to adapt, and sometimes also create new procedures, measures, and analyses that are suitable for our research questions, needs and priorities.

Cognitive validity is also not just about cognitive processing during test completion, but equally about linking this back to a language use domain. It is about verifying the appropriateness of the cognitive processes needed to successfully complete a test and the extent to which the processes replicate those employed in the target language use domain. Similarly, the explanation and extrapolation inferences in argument-based approaches concern the extent to which test scores (and their underlying processes) reflect the intended construct and performance in the target domain, respectively. Consequently, I propose that in the coming years more effort goes into researching and establishing: What should we be testing? Or put differently, what really is the nature of the target language use domain and thus the target construct? How is language used, by people, in context, today? In validation theory terms, this is about domain definitions. While our field comprises a body of work in this area already, for example needs analyses, and can also rely on knowledge established by other areas of (applied) linguistics, the speed with which language use domains and forms of communication are evolving urges for greater currency in domain definitions. To give one example, multimodal language use descriptions and test constructs are quite limited in the field at present. Implications of this are also that the field will need to find more and better ways to go beyond the still predominant paradigm, at least in testing practice, of a language 'standard' and 'language separability'. While challenges around testing English as a lingua franca and World Englishes have been on the agenda for quite some time, and mediation has been adopted in the Common European Framework of Reference (Council of Europe, 2020), our field would benefit further from greater insights into and debates around the realities of, for example, heritage language use, translanguaging, intercultural communication, language in highly specific domains, testing in languages other than English, and so on. It will thus continue to be vital that we engage with knowledge and insights as they become available in other areas of (applied) linguistics. From a methodological point of view, domain definitions are also likely to benefit from engaging more extensively and systematically with advances and applications in, for example, corpus linguistics as well as ethnography – to name two 'ends' of the spectrum. As we do this, again it will be important to evaluate these interfaces with findings and methods from other fields through the lens of key questions, principles and issues in language testing and assessment.

A focus on 'how'

The operationalization of language constructs constitutes crucial activity in language testing practice, while often being a balancing act between research recommendations, available resources, educational policy priorities, marketing and commercial drives, etc. From a validation perspective, considerations on how to test are the focus of, for example, the dimensions of context and scoring validity in the socio-cognitive approach (i.e., fair test task characteristics and administration, and dependable and meaningful test scores, respectively). While scholarly and operational work on how to test is well-established, continuous and increasingly fast developments in this area require it to remain in the spotlight. A case in point at present is the role of technology in language testing and assessment. Undeniably, technological innovations have created many useful opportunities and advances (for an overview, see Schmidgall & Powers, 2017), with key ones including: 1) facilitating and optimizing administrative procedures in test development and rating (e.g., item development tracking, item banking and archiving, performance storage and distribution to raters), 2) generating items and compiling tests, 3) delivering tests (e.g., online administration, accessibility of tests, frequency of testing), 4) enhancing security in technical respects (e.g., monitoring, authentication), 5) reflecting technology-mediated language use, and 6) automated scoring and feedback generation. One important concern, however, in which I argue a significant risk lies, is that language testing becomes technology-driven rather than technology-enhanced, with 'what we test' being overrun by 'how we test'. Essentially, the challenge is not to lose sight of the construct and to ensure that, when using technological solutions, we have a clear understanding of the type of language (use) we aim to test, we are actually testing, and to what extent it reflects the target language use domain adequately and comprehensively. This issue is, for example, illustrated by Huawei and Aryadoust's (2022) systematic review of automated writing evaluation (AWE) systems, which concluded that domain definition inferences are underrepresented in AWE research. Similarly, even the most sophisticated automatic item generation (AIG) systems (e.g. Attali et al., 2022), while making great advances, are currently still trained on more 'traditional', 'conventional' language use, and restricted in the kind of tasks (input and questions) they can produce (as well as continuing to require considerable human reviewing). Thus, a challenge for future work on technology in language testing and assessment will be to reflect the progress made in (applied) linguistics regarding the nature of language and language use in society, and to prioritise construct operationalisation insights, in order to avoid restricting domain representation through technology-mediated testing. Another challenge for technology-enhanced language assessment is that, to date, it is primarily associated with large-scale (English) proficiency testing. Future opportunities lie therefore in feasibility of, for example, small-scale, classroom testing, and testing in a variety of world languages and for different purposes and needs. At the same time, the possibility that technology 'drops' a test-taker into the target language use context might soon no longer be just blue sky thinking.

A focus on 'why'

It is often reasonably straightforward to classify tests according to 'traditional' purposes such as progress and achievement to evaluate language learning success in instructed contexts, proficiency to establish level of language ability, diagnosis to evaluate weaknesses in language learning, etc. However, this is often only the surface level of the answer to the question of why we are testing. Behind a seemingly low-stakes progress-test purpose, for example, may lie not just learning gain checks or pedagogic adaptations, but higher-stakes external teacher evaluation checks and job retainment or promotion. Or, underlying proficiency testing for societal participation might be particular political ideologies and strategies. A transparent, deep answer to the why-question is, in fact, indispensable from a validation perspective. For instance, clearly articulating why the test exists is necessary as part of the utilization inference in argument-based approaches, as this requires an evaluation of the usefulness of scores on a test for an explicit, stated purpose. In that sense, an enduring challenge can also be found in denouncing test re-purposing or retrofitting practices such as the use of general proficiency tests for high-stakes decision-making on professional functioning in very specific language use domains (and vice versa). If a test is re-purposed, the why question shifts, necessitating a new utilization inference and new or expanded validation.

A related question that is perhaps not posed as much, nor made explicit, is: Should we test? This question is really fundamental, and equally central to considerations of test use and consequences. To illustrate this point with a recent unsettling example, an airline introduced a factual knowledge test administered solely through the medium of one language spoken in a country with 11 official languages – effectively rendering it a second language reading and writing proficiency test for the majority of the country's nationals. The stated purpose was verification of the legitimacy of passports from the country, and thus the airline essentially introduced a new, random border control check and entry requirement with this test. Fortunately, after public outcry (including from representatives of language boards from the country concerned), the test was withdrawn, but it seemed the airline had never in the first place asked the question of whether they should test. A less extreme example may be the introduction of, for example, national, standardized tests, or of weekly progress tests, in regular language education. The first question that ought to be asked is: Does it address an issue or need in the language education system, and is a language test a meaningful, useful, and justifiable tool for addressing the issue? Sometimes, despite an identifiable reason, the answer may be that we should not test (e.g. language testing purely as an accountability check, language testing for ideological reasons rather than language use reasons). From a validation perspective, the question of 'should we be testing' feeds into the consequence inference in argument-based approaches or consequential validity in the socio-cognitive approach, i.e. will the intended use of the test scores (and thus of the test itself in the first place) result in positive consequences.

An important task for our field is therefore to ensure that both parts of the why-question (what do we test for, and should we test) continue to be asked and evaluated systematically, openly and critically. This requires extensive language assessment literacy and ethical language testing of us as scholars and practitioners. It also necessitates continued and expanding efforts in language assessment literacy development of test commissioners and score users.

Two more pleas

Finally, I would like to put forward two further opportunities for our field. First, I think it is fair to say that our field has reached firm maturity by now, with a significant body of work we can rely and build on. This includes empirical work conducted on similar topics, often with similar methodologies, even when in different contexts. Therefore, our field is in a position now to not only continue conducting new empirical studies, but also to increasingly take stock through synthetic approaches such as systematic reviews or meta-analyses. As Luke Plonsky's bibliography of meta-analysis in applied linguistics demonstrates, these are proportionally still limited in language testing and assessment (https://lukeplonsky.wordpress.com/bibliographies/meta-analysis/). Such synthetic approaches can provide us, however, with greater insights into the scalability and generalizability of research findings and issues, as well as help shape our future research and development agenda.

Secondly, and perhaps most crucially, I would like to plea for further diversification of voices in our field. This is about ensuring the growing representation of the views and lived experiences of stakeholders such as test-takers, parents, classroom teachers, and test score users, as well as representation of the diversity within each of these stakeholder groups (e.g. cultural, economic, educational, geographical, linguistic, etc.). It is also about further awareness of and increases in language assessment research and practice of a diversity of languages and of diversity within language and language use. Within the scholarly community, it is about much stronger representation of perspectives and work from, for example, the Global South and other underrepresented regions, as well as underrepresented voices within well-represented regions. And finally, it might also involve looking into the representation of ourselves as stakeholders, how we and our work are being constructed, for example, in the media or by policy makers, and the pathways we (are (not) prepared to) walk in language testing and assessment.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., and von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2022.903077

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42. https://doi.org/10.1177/026553220001700101

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34. https://doi.org/10.1207/s15434311laq0201_1

Council of Europe (2020). *Common European framework of reference for languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. https://www.coe.int/lang-cefr

Huawei, S., & Aryadoust, V. (2022). A systematic review of automated writing evaluation systems. *Education and Information Technologies*. https://doi.org/10.1007/s10639-022-11200-7

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

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

Schmidgall, J. E., & Powers, D. E. (2017). Technology and high stakes language testing. In C. A. Chapelle & S. Sauro (Eds.), *The handbook of technology and second language teaching and learning*. Wiley.

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