Bridging assessment and learning: A view from second and foreign language assessment

J. Charles Alderson, Tineke Brunfaut & Luke Harding

Lancaster University, UK

Abstract

This paper considers issues around the relationship between assessment and learning, as put forward by Baird, Andrich, Hopfenbeck and Stobart (2017), from the perspective of the field of second and foreign language assessment. In our response, we describe shared observations on the nature of research and practice in general educational assessment and in language assessment (including with respect to linking assessment with theories of learning, managing impact, and enhancing assessment literacy). At the same time, we also identify areas where language assessment seems to diverge from current research and practice in general educational assessment (for example in the areas of assessment purposes, construct definitions, and validation theory and practice). As a consequence, we believe that close monitoring of advances in both fields is likely to be mutually beneficial.

Keywords

Language testing, second and foreign language assessment, validity, constructs, language assessment literacy, consequences

1. Introduction

As researchers working within the field of second and foreign language assessment we welcome the chance to respond to the position paper by Baird, Andrich, Hopfenbeck and Stobart (2017). Language assessment is conducted across a range of educational and professional contexts (e.g., large-scale international tests, curriculum-based progress tests and school-leaving examinations, certification assessments for employment purposes) and is often considered – at least by those who work within the field – to be a related, but distinct, disciplinary area from ‘general’ educational assessment. Language assessment has its own intellectual traditions, with connections to education and psychometrics, but also, fundamentally, to (applied) linguistics and second language acquisition. As a result, we recognise many common themes and areas of overlap in the paper. Language assessment is also increasingly concerned with the need to connect assessment – particularly large-scale high-stakes testing – with theories of learning, with understanding and managing the effects of washback, and with enhancing assessment literacy (particularly among policy-makers). At the same time, however, we also recognise areas where language assessment does not fit the characterisation of educational assessment that is presented in the position paper. Language assessment's intellectual roots have steered the field in a different direction to other forms of educational assessment such that, for example, there is a very strong tradition of research into the language constructs which underlie language assessments (even if gaps in knowledge still remain), and there is an equally strong body of research work which validates language assessment (indeed, some advances in validity theory have been spurred by scholars in the area of language assessment). In this paper we will elaborate on both these commonalities and these divergences with a view to charting what language assessment can learn from the critique put forth by Baird et al., and what language assessment might contribute to the broader field of educational assessment.

2. Commonalities

2.1 Learning theories and assessment

The main impetus for Baird et al.'s position paper is their observation of 'a missing link' between learning theory and assessment theory. Although they identify three types of learning theories and provide brief examples of assessment approaches typically associated with each, Baird et al. concede that 'tracing the connections between learning theories and assessment design is difficult' (p.3). At the same time, they recognise the impact of assessment on teaching and learning as documented in the research literature. While being aware of the 'intuitive' expectation of impact in high-stakes assessment contexts, Baird et al. purposefully explore two contrasting examples of assessments (Case A – international tests, Case B – Assessment for Learning) to extend the discussion of impact beyond the concept of stakes. Essentially, they describe how, in principle, both of the examples have enormous potential to mutually inform learning and assessment, but how theoretical, political, commercial, and methodological issues so far have prevented realising this potential.

Language assessment has also been grappling with ways of harmonising learning theories with assessment practices. To a certain extent language assessment has long had connections with, and benefitted from the insights of, second language acquisition (SLA) research into language development. The importance of dialogue between these two fields was noted almost thirty years ago in a paper by Bachman (1989), and elaborated in a later edited volume (Bachman & Cohen, 1998). The annual international conference for language testing and assessment (the Language Testing Research Colloquium) provides one forum where second language acquisition theories are addressed alongside more practical matters of assessment design and analysis. Similarly, EUROSLA (the European Second Language Association) was founded in 1989 and has contributed many books, research articles and conference papers on the interface between SLA and language assessment (e.g., Bartning, Martin & Vedder, 2010).

More recently, however, there have been several new paradigms proposed for bringing assessment practice more into line with theories of *learning* (that is, with what should ideally happen in the language classroom or in other domains of language instruction). Relatively new and vibrant areas of research have been proposed, such as diagnostic assessment (Alderson, 2005; Alderson, Haapakangas, Huhta, Nieminen, & Ullakonoja, 2015a; Alderson, Brunfaut, & Harding, 2015b; Harding, Alderson, & Brunfaut, 2015), learning-oriented assessment (Turner & Purpura, 2015; Jones & Saville, 2016), and dynamic assessment (Lantolf & Poehner, 2008; 2011) (the latter two derived from educational research outside of language assessment). While these approaches differ in the nature of the assessment procedures recommended (e.g., the role of self- and peer-assessment), and the extent to which assessment tasks need to resemble authentic real-world language tasks, they share a common vision with respect to the central place of the individual learner, the need for ‘diagnostic competence’ among teachers, and, perhaps most importantly, the provision of meaningful feedback which promotes further learning. Within these paradigms, learning and assessment have become not only intertwined, but mutually dependent. As Lantolf and Poehner (2008) have written of dynamic assessment, assessment and instruction “become as tightly conjoined as two sides of the same coin – and there are no one-sided coins” (p. 274).

We concur, therefore, that the relationship between language assessment and language learning is indeed important, and that the need to bridge the ‘missing link’ identified in the position paper is a shared mission. In the case of language assessment, the interface with SLA and the learning “turn” in research and practice have proved productive. However we would also note that while diagnostic assessment, learning-oriented assessment and dynamic assessment might be considered “hot topics” in terms of research, they are still on the fringe with regard to wide-scale use and practice.

2.2 Washback

A key motivation for Baird et al.’s plea for a clearer connection between theories of learning and educational assessment is the effect assessments have on teaching and learning – otherwise termed ‘washback’, which may be positive, negative or mixed. Within the field of second and foreign language assessment, for several decades now, there has been a strong research interest in identifying the washback of, in particular, large-scale language examinations (see e.g., Cheng (2014) for an overview). What has become clear from the empirical findings in this area is that washback is a very complex phenomenon and typically a mixed bag of intended *and* unintended effects. Another important observation from this research stream is that the power of assessments to “change [learners and teachers’] behaviours” (Baird et al., 2017, p.4) is primarily found in *what* is being learned and taught (the content), and less so in *how* language is learned or taught (the methodology), and that the effects are not consistent in intensity nor direction for all those involved (the stakeholders).

At the same time, there have been a number of conceptual advances which prioritise the need to take an effect-driven approach to test design (Davidson & Fulcher, 2007). Within the field of second and foreign language assessment, impact and washback have become theorised as key principles of test usefulness (see e.g., Bachman and Palmer, 1996); Messick (1989) captured these as consequential aspects of construct validity to the extent that they can be traced back to construct underrepresentation and construct-irrelevant variance. Other scholars have conceptualised consequences beyond the quality of the test (see e.g. Chalhoub-Deville, 2012) and some have labelled this as ‘consequential validity’ (see e.g., Weir, 2005). Also, language test validation theory and practice have evolved in such a manner that evidence is being required even before the test event (see e.g., Weir’s 2005 *a priori* validation). In this way, validation theory and practice seem to have steered an increasing number of language test developers to design their instruments with test uses and consequences at the forefront. We have seen evidence of this sort of “washback by design” approach in Europe with the reform of the language examinations in the Austrian Matura (see Spöttl, Kremmel, Holzknecht, & Alderson, 2016) and the planned reform of the secondary school-leaving English exam in Luxembourg (Brunfaut & Harding, in press). For example, the intended washback effect of the reform of the Austrian Matura’s foreign language exams represented a switch from knowledge-based to competence-oriented language teaching and assessment. This objective fundamentally shaped decisions on the exams’ construct, test development approach, stakeholder involvement, test formats, etc. Some evidence for the intended washback is, for instance, provided in Froetscher (2017) who investigated the washback of the reading section of the exam. She found an increased use of reading tasks by classroom teachers, with the tasks also being of higher quality (targeting a range of reading behaviours underlying the Matura exam and adhering to a wide range of principles of good task design). Interviews with teachers also suggested increased teacher awareness in the area of task and test design, and more opportunities for language learners to demonstrate their ability in a wider range of language skills (Froetscher, 2017).

2.3 Language assessment literacy

A common motif running through the position paper is the challenge of aligning a learning-based approach to assessment with the edicts of policy-makers. This is made more complicated by the observation that “policy-makers who take decisions on the basis of educational assessment data rarely understand the content of the tests or the effects upon learning of changing them” (p.26). This last point reflects a topic which is currently receiving considerable attention in our field: language assessment literacy. The focus of much current language assessment literacy research and commentary has, to date, been on classroom teachers: what they need to know, or be able to do, in order to conduct assessment in the classroom (e.g., Inbar-Lourie, 2008; Scarino, 2013), or on the breadth of knowledge and abilities required of test developers and other assessment professionals (e.g., Fulcher, 2012). However, language assessment shares a pressing need to engage more with policy-makers, and to find ways in which to enhance knowledge across this important stakeholder border (i.e., policy-makers becoming more assessment literate, and assessment professionals becoming more policy-literate).

The research which has been conducted in language assessment suggests that the need for engagement is profound and urgent. For example, Pill and Harding (2013) investigated evidence of assessment literacy among policy-makers in an Australian parliamentary inquiry into the certification of overseas-trained health professionals (part of the process of which involves a language proficiency test). Policy-makers, and other attendant witnesses to the inquiry, were found to display some very basic misunderstandings of assessment procedures, and to rely for their understanding of the relevant language constructs on the advice of others who were not assessment or language specialists.

Assessment literacy problems do not always result in "passive" misconceptions, however. Pižorn and Nagy (2009) report on language exam reform projects in Slovenia and Hungary which were designed to shift language assessment in those countries to a more communicative, four-skills approach (e.g., to develop assessments based on more authentic and useful constructs of language ability). In both cases, the reforms were actively thwarted by decision-makers who lacked expertise in assessment, but who made arbitrary and, ultimately, detrimental decisions based on political rather than pedagogical goals. An example is the instalment, withdrawal and re-instalment of subject test teams, in the course of one year, for primary school national assessment in Slovenia, according to the specific political party that was in power and ultimately leading the English team to resign due to the unfeasibility of developing a good quality exam in the remaining time. Another example is situated in the context of the abandonment of the Hungarian Examinations Reform Project: the deletion of paragraphs from a test specifications document by a Ministry official and the denial of the value of standard-setting as opposed to the simple imposition of a cut-score by this same official (Pižorn and Nagy, 2009). It is clear that further research is required into the influence of policy-makers on language assessment decisions more generally. However, direct action is also required to remedy these observed problems.

3. Divergences

3.1 Assessment purpose

Baird et al. open their position paper with the observation that "[a]ssessment plays a central role in education. Assessments are used to investigate what people know and can do and to make decisions regarding whether they have learned what was expected." (p.1). From the outset it is clear that assessment is understood chiefly as an activity that takes place within educational systems. From this perspective, assessment – whether it be in the domain of mathematics, science, geography, and so on – should have a strong connection with state-of-the-art learning theories, an argument which the paper advances throughout. There is nothing controversial in this view, and indeed Baird et al. make the case that while this connection between assessment and learning might be understood as natural, it is unfortunately not common in practice across many educational systems.

From the perspective of language assessment, however, views about the strength of connection between assessment and learning theories have often been tempered by considerations of test purpose. Language assessment is perhaps unique as a form of assessment in that it is practised just as commonly outside of educational systems as it is within them. For example, a language assessment might also be used to screen the language proficiency of nurses, or to test the language proficiency of aviators – pilots and air traffic controllers – for professional purposes. Such forms of test use are independent of any particular theory of learning, based, as they are, on the skills and abilities deemed necessary to function adequately in a given target-language use domain. It is, of course, ideal if in these test contexts there is also a strong connection with learning through preparation practices that are relevant (i.e., positive washback), or through feedback systems which provide fine-grained information to test-takers about how they have performed. However, in such cases the test developer or score user is primarily interested in what the person's proficiency is in the target language use domain, not necessarily in how they have acquired it or learned the language. Indeed, theories of validity in language assessment emphasise the centrality of test or assessment purpose. Thus, an analysis of the purpose of a language test or an assessment procedure is essential, but the primary purpose need not be educational.

3.2 Investigating constructs

A fundamental shortcoming of educational assessment at present, as identified by Baird et al., is the omission of construct definitions as a trade-off for the large emphasis on *how* to assess. However, without clear views on *what* is being assessed, Baird et al. argue, test validity has largely remained a theoretical exercise. The field of language assessment, however, may form an exception in this regard. More specifically, a long tradition of enquiry into constructs exists. There is a tendency for scholars and practitioners to specialise in both the nature and assessment of certain aspects of language and language use, and to work at the intersection of language assessment and second language acquisition. This can be witnessed in many publications in the field, which preface assessment research findings and discussions with a section defining the construct. Each of the volumes in the *Cambridge Language Assessment Series*, for example, start by providing the reader with an overview chapter on 'The nature of … [reading, writing, etc.]'. A considerable amount of such research into constructs has in fact been funded by exam boards (see, for example, the grant programmes by the British Council, Cambridge English, ETS, Pearson, etc.) and published in these organisations' research report series (e.g., the IELTS Research Reports, the TOEFL Research Reports and Monograph Series, the British Council's Assessment Research Awards and Grants Reports and their Validation Series). Consequently, construct definitions have directly informed the development of language assessment instruments, and, in turn, formed the basis of empirical validation research. A recent example of this can be found in the development of the Aptis reading test and related validation research into this test for English second language speakers (see O'Sullivan, 2015; Brunfaut & McCray, 2015; Brunfaut, 2016). Language testing researchers have thereby been assisted by developments in research methodology, which have resulted in most evidence now being based on findings generated by means of multiple methods – some being collected prior to test administration (for example using expert judgements), some during test administration (for example using eye-tracking, key-stroke logging), and some after test administration (for example through stimulated recalls, test performance statistics, corpus linguistic analyses of performances).

Despite this considerable amount of work on language constructs in the area of second and foreign language assessment, we acknowledge that there remains much to discover about language constructs, with debates continuing around the (multi-)dimensionality of language skills, the boundaries of language ability (in connection with, say, professional competence or personality factors), the immense variability in language use attributed to contextual variables, and the role of values in determining what counts within construct definition in the first place. We also recognise that, to date, in particular in smaller-scale and more classroom-based language assessment contexts, the connections between constructs and assessment instruments (and their validation) may be less extensively conceptualised, documented, and researched.

3.3 Validity and validation in language assessment

One further area where it could be argued that language assessment shows a divergent path from the picture painted of general education in Baird et al. would be the application of validity theory, and validation processes, to testing practice. Baird et al. (p.6) make the point that validation is sorely under-researched in the context of many educational assessments. In language assessment, validity and validation have been central to the research efforts of many large-scale assessments (e.g., the Cambridge Exams and TOEFL), and validation research forms the backbone of the two major journals in our field: *Language Testing* and *Language Assessment Quarterly*. To take an example, the development of a revised version of the TOEFL test resulted in a very detailed application of Kane's interpretive argument approach to validation, a process which itself acted as a useful testing ground for the application some of Kane's theoretical principles (as described in Chapelle, Enright, & Jamieson, 2008). Further, as some scholars within the field have embraced consequences within a broader validity framework (e.g. Weir, 2005), the connection between assessment and learning has come to the fore in validation research.

This is not to say that gaps do not exist. For example, a concerning problem for the field arose when it was noted that a number of tests designed to assess pilots and air traffic controllers against the language proficiency requirements developed by the International Civil Aviation Organisation did not demonstrate sufficient evidence of validation for such a high-stakes decision. In such cases, strong professional associations are important in encouraging and, if necessary, lobbying for evidence which will support decisions based on test scores. An important role is played by the various professional associations for language testing and assessment – e.g., the International Language Testing Association (ILTA), the European Association for Language Testing and Assessment (EALTA), the United Kingdom Association for Language Testing and Assessment (UKALTA), the Asian Association for Language Assessment (AALA), the Japan Language Testing Association (JLTA) and the Assessment and Language Testing Association of Australia and New Zealand (ALTAANZ) – in encouraging and upholding codes of practice and ethical values, all of which have validity at the forefront.

4. Conclusion

In this response, we have set out some areas where we see overlaps between the Baird et al. position paper and current discussion around theory and practice in language assessment. These include the turn towards learning theories in developing relevant and useful assessment, the key consideration of washback, and the need for a greater awareness of, and engagement with, the assessment literacy of policy-makers and other key decision-makers in educational contexts and beyond. We also noted some aspects where the state-of-the-art depiction of educational assessment by Baird et al. was not as clearly applicable to our field: the notion that a good deal of language assessment takes place for certification or screening purposes, where tasks are more appropriately related to modelling the demands of future language *use*; the deep tradition and interest in theorising and researching language constructs, and the central place of validity and validation in the field.

We would conclude, therefore, that our fields seem to be pointing in the same direction, although the unique intellectual and disciplinary traditions of language assessment – its key connection with applied linguistics and language studies more generally, and the resulting sense of a separate disciplinary identity – suggest we will continue to travel somewhat different paths. Nevertheless, language assessment will stand to benefit from closely observing the way in which educational assessment more generally grapples with the challenges of integrating learning theories into assessment design. Cross-pollination of ideas is likely to be fruitful, provided that new approaches can be adapted for the assessment of language constructs. Conversely, engagement with the research literature on language assessment might provide those in other areas of educational assessment with some useful insights, particularly with respect to the tradition of research on construct definition and validation.

References

Alderson , J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C., Haapakangas, E-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. New York: Routledge

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.

Bachman, L. F. (1989). Language testing-SLA research interfaces. *Annual Review of Applied Linguistics, 9*, 193-209.

Bachman, L. F., & Cohen, A. D. (1998). *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bartning, I., Martin, M., & Vedder, I. (2010). *Communicative proficiency and linguistic development: intersections between SLA and language testing research*. Amsterdam: Eurosla.

Brunfaut, T. (2016). *Looking into reading II: A follow-up study on test-takers' cognitive processes while completing Aptis B1 reading tasks* (British Council Validation Series, VS/2016/001). London: The British Council.

Brunfaut, T., & Harding, L. (in press). Teachers setting the assessment (literacy) agenda: a case study of a teacher-led national test development project in Luxembourg. In D. Xerri, & P. Vella Briffa (Eds.), *Teacher involvement in high stakes language testing*. Springer.

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes whilst completing reading tasks: a mixed-method eye-tracking and stimulated recall study* (ARAGs Research Reports Online, AR-G/2015/001). London: The British Council.

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing, 33*(4), 453-472.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York and Oxford: Routledge.

Cheng, L. (2014). Consequences, impact, and washback. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp.1130-1146). Hoboken, NJ: Wiley-Blackwell.

Davidson, F., & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching, 40*(03), 231-241.

Enright, M. K., & Tyson, E. (2011). Validity evidence supporting the interpretation and use of TOEFL iBT Scores. *TOEFL iBT Research Series 1, Volume 4*. http://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v4.pdf

Froetscher, D. M. (2017). *An investigation into the washback of a standardized national exam on the classroom testing of reading* (Unpublished doctoral dissertation). Lancaster University, UK.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113-132.

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.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*(3), 385–402.

Jones, N. & Saville, N. (2016). *Learning-oriented assessment: A systemic approach*. Cambridge: Cambridge University Press.

Lantolf, J. P., & Poehner, M. E. (2008). *Sociocultural theory and the teaching of second languages*. London: Equinox.

Lantolf, J. P., & Poehner, M. E. (2011). Dynamic assessment in the classroom: Vygotskian praxis for second language development. *Language Teaching Research, 15*(1), 11-33.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp.13-23). New York: American Council on Education.

O'Sullivan, B. (2015). *Aptis test development approach* (Technical Report TR/2015/001). London: The British Council. https://www.britishcouncil.org/exam/aptis/research/publications/test-development-approach

Pill, J., & Harding, L. (2013). Defining the language assessment literacy 'gap': Evidence from a parliamentary inquiry. *Language Testing, 30*(3), 381-402.

Pižorn, K., & Nagy, E. (2009). The politics of examination reform in Central Europe. In J. C. Alderson (Ed.), *The politics of language education*. Bristol: Multilingual Matters.

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing, 30*(3), 309-327.

Spöttl, C., Kremmel, B., Holzknecht, F., & Alderson, J. C. (2016). Evaluating the achievements and challenges in reforming a national language exam: The reform team's perspective. *Papers in Language Testing and Assessment, 5*(1), 1-22.

Turner, C. E., & Purpura, J. E. (2015). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari and J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 255-272). Berlin: DeGruyter.

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