A most engaging scholar: Tim McNamara and the role of language testing expertise

John Pill & Luke Harding

Lancaster University

Abstract

Research in language testing often highlights a mismatch of expectations between test developers, who seek to create and maintain the best possible tests for different contexts, and other stakeholders, whose more pragmatic use of tests and test scores may conflict with established testing principles. Discussion of research findings may then point to the importance of improving stakeholders' language assessment literacy. However, where does responsibility lie for such a project and by what methods is it likely to be achieved? How can the appropriate approach be found between simply asserting that the testing experts know best and underselling the contribution that this field can make to finding real-world solutions?

During his career, Tim McNamara has theorised on and provided critique of policy and practice in language and citizenship testing, and language analysis for the determination of origin of asylum seekers, among many areas of study. He has consistently challenged the field of language testing to look for new ideas beyond its self-imposed borders. In this chapter, we reflect on the need for and implications of active engagement between academic expertise and decision-making processes that have possibly life-changing consequences for those involved.

Introduction

There are few academic scholars working in the field of language assessment as engaging as Tim McNamara. We use the word "engaging" here to signify both being interesting (drawing others into a topic) and being interested (taking an interest in others' affairs). Those who know Tim – as we have had the good fortune to as students and, later, as colleagues – know that his engaging nature is first and foremost a personal characteristic. Tim is interested in the world, and in the lives of others. This is why, if you speak to the many students Tim has taught, supervised and mentored at the University of Melbourne over the years, they will often talk not only of the experience in intellectual terms, but of the warmth and personal connection that Tim fostered. Beyond the personal domain, however, Tim’s intellectual journey can also be well described as engaging – drawing others into the field of language testing (by introducing them to the intellectual richness and breadth of the issues assessment touches upon), and engaging with scholarly work outside the field. This outside engagement involves both broadening the range of connections made between language testing and other fields of academic enquiry – for example, Tim was an avid, long-term member of an interdisciplinary Derrida reading group at the University of Melbourne – and engaging more practically with the real-world impact and consequences of language tests – for example, Tim’s work on language analysis for the determination of origin of asylum seekers.

Our own work has included, in the last few years, a focus on language assessment literacy (LAL) – the consideration of what different stakeholders involved in language assessment need to know in order to make sound decisions, and how this can best be developed. The central theme of this area of research is also engagement – encouraging stakeholders to engage with expertise in language assessment, and encouraging language assessment practitioners to engage both with a wide range of academic thought (to expand their repertoire of knowledge) and to engage with real-world applications of language testing by becoming more policy-literate.

Engagement thus serves as an organising principle for this chapter. In the sections that follow, we first problematise the nature of engagement in language testing and assessment, discussing both the advantages and the limitations of existing conceptualisations of language assessment literacy. We then go on to discuss a specific example where LAL is found wanting, and where the requirements of different stakeholders are difficult to reconcile. In the final section, we discuss a potentially fruitful approach that combines elements of a McNamara-inspired engagement: exploring the utility of frame theory in language assessment communication.

Engagement in language testing

A criticism that could be levelled at academics involved in the study of language testing is that we are too often inward-looking and conservative in our endeavours (Jenkins & Leung, 2014; McNamara, 2014). Our research typically deals with topics that we believe to be important in the field: investigating, for example, the reliability of a scoring method, the cognitive processes that test-takers employ in performing an assessment task, or the empirical support for a particular cut-score. Nevertheless, to a non-practitioner the value of this work might not be immediately recognisable. Rather, test score users may be more interested in knowing the answers to questions such as “how do test scores on different tests relate to each other?”, “which test is more susceptible to cheating?”, and “which test can be delivered most widely?”. The concerns of researchers and practitioners, on the one hand, and score users, on the other, may be quite different – and in some cases, at odds.

This gap in what is considered important knowledge has implications for the fair and ethical use of language tests within society. The role of testing, including language testing, has grown in recent years. For example, in many countries children now take standardised tests that measure their progress through the compulsory education system, providing statistics used to describe the performance of teachers, schools and regions. Language test requirements are imposed by governments to regulate the flow of people seeking to work or study in their jurisdictions. Tests are also employed to indicate individuals' readiness to be given permanent residence or citizenship in a new country and to take on the rights and responsibilities attributed to such a status. However, tests appear to be taken for granted – as an element of the infrastructure of modern society – and consequently the risk arises that a test and its purpose might come to be understood in a way which is different from that intended by the test designer and supported through validation studies. In turn, this misconstruing of a test may lead to inappropriate use of that test, with unfair outcomes for individuals due to poorly informed decision-making. This apparent gap in knowledge and skills around testing has been recognised in our field, prompting the emergence of the concept of language assessment literacy (e.g., Taylor, 2009).

The now common usage of *literacy* beyond its original sense of having the ability to read and write – for example, in terms like "health literacy" or "information literacy" – seems to indicate that the area of concern is new or changing (becoming more complicated) and requires individual knowledge and skills across a larger group of people than was previously expected to possess them. New literacies therefore emerge as society changes and people need to learn how to read (interpret) new materials. Concepts that were once understood only by a minority are found to be required more widely. Through the coining of a new term ("X literacy") in academia and the media, the topic ("X") is presented as something that members of society need to engage with and learn about. Becoming literate is generally viewed positively, seen as the development of skills necessary for a productive life that can benefit from the new technologies and services available. For example, the promotion of computer literacy was prominent in the 1990s, when, at least in more economically developed countries, personal computers became accessible to larger numbers of people in the workplace and at home, and computer skills were recognised as vital to employment and economic progress. To some extent, this focus has since decreased and such literacy often now seems to be assumed. Personal finance is an area in which literacy has also been promoted; financial literacy relates to an individual's knowledge and skills to make effective decisions relating to their financial resources (see, e.g., Organisation for Economic Co-operation and Development, 2017, for an international study of schoolchildren's financial literacy). Professional experts remain available to guide financial planning and decision-making, but governments and financial institutions promote the development of greater awareness of this topic among individuals across society.

The focus on language assessment literacy has arguably come about because of the increased use of standardised tests (taken more frequently and in more contexts) and the perceived need for comparability between tests and test scores from different countries, due to the increasingly global nature of education, employment and migration. More people are involved in taking tests, and in using test scores for decision-making purposes. Expectations in many societies today include respecting the rights of individuals as well as holding people and organisations to account. These beliefs are likely, too, to affect how people view tests and the impact they have on their opportunities in life. Similar to financial literacy above, therefore, there is a belief that test processes and decisions should be transparent and comprehensible to the individuals directly affected by them. Test takers should understand the purpose of the test and its suitability to achieve that purpose effectively. Likewise, test users should inform themselves to be able to make defensible decisions about, for example, whether to implement a test, which test (existing or newly designed) best meets their needs, and which test scores are appropriate to sort test takers into the categories required. Test users can be held accountable for such decisions both by individual test takers who believe they have been treated unfairly and by representative groups who may argue that the process as a whole is flawed and discriminatory. The promotion of language assessment literacy may therefore also be seen as part of an ongoing negotiation, between testing experts and those who are affected by tests – test takers, test users and other stakeholders, to define what is feasible and appropriate in terms of how tests are applied.

Language teachers are clearly an important group of test users. Research has shown that, as well as recognising their need for increased language assessment literacy, teachers are also willing to be active in seeking the knowledge and skills to meet this need (Harding & Kremmel, 2016; Hill, 2017). Language teaching is a good example of a field where knowledge about testing that was perhaps not so necessary in the past has now become a central requirement in many contexts. Large-scale assessment literacy projects are underway in some areas of teaching, indicating a healthy level of engagement, although willingness and capacity to participate must vary given the wide range of education situations in which language teaching and assessment occur (e.g., public–private, compulsory–voluntary, child–adult, formal–informal).

The level of engagement with language assessment literacy is often less clear for other test users. Policy makers at different levels of governmental or organisational structure – public service administrators, government advisers, chief executives and operations managers, for example – are likely to be required to consider the use of language tests and how scores may affect the outcomes they are seeking. In many cases, these outcomes are not directly related to the purpose of the test. For instance, the language test score required for a particular type of work visa could be raised, not because the language demands made of those workers were originally underestimated and the workers are not performing satisfactorily, but because too many workers have been obtaining visas and the market for their employment is now saturated. However, there is little evidence that policy makers seek expert advice from the field of language testing to inform their decision-making processes (Pill & Macqueen, 2017). It may be that the policy in force is based on decisions which current post-holders assume to have been made by those previously responsible. The decisions may be undocumented and therefore unsupported by evidence or argument; they may nevertheless have acquired great status and be incontestable. Amendments to policy, to accommodate changes such as a revised test scoring system or a test promoted as equivalent by a new provider, are likely to be made on an ad hoc basis, commonly paralleling the action taken by a comparable institution or jurisdiction that has faced the same decision. The effects of any changes to policy are often not monitored, reducing the possible impact of policy review. Experience indicates that policy makers even in educational contexts where expertise is available within the institution rarely seek out language testing researchers to participate in test-related decision-making. The published admissions requirements for international students will include complex regulations about which tests are recognised and the scores demanded by the institution, but a review of these regulations that invites the participation of representatives from pre-sessional language programmes or staff with direct knowledge of the language tests involved is likely to be an exception rather than the rule.

The definition of literacy attempted above included the sense that the goal was to improve knowledge and skills widely in society. It seems that this characteristic of literacy does not fit particularly well for policy makers. While test takers and teachers do appear amenable to and, in some instances, are actively pressing for opportunities to improve their understanding of language assessment, the same claim is not so easily made for policy makers. Instead, three scenarios can be imagined: (a) policy makers may believe that there is no need for them to develop their literacy in issues around language assessment; (b) they may not recognise their own lack of literacy or their need for increased literacy (see Pill & Harding, 2013); or (c) they may recognise this need but not know what to do about it. The field of language testing is not large and does not have a high profile, so it could easily be overlooked as a source of expertise. All three scenarios are, ultimately, problems of engagement.

Regarding the first possibility (a), it is not so clear in the case of policy makers that possessing greater knowledge of language testing would in fact help them carry out their professional duties more effectively. In most contexts, policy-making involves pragmatism and compromise; it has been argued that testing expertise is not sought because it is viewed as too “pure” and inflexible to be of help in the give-and-take of the real world (e.g., Deygers & Malone, 2019). There is some truth in this. The assumptions of academic research are different from those in an environment that may be strongly outcome- and market-driven. Nevertheless, there is surely value in promoting, even in conflicting circumstances, what is commonly recognised as good practice in the discipline, and in maintaining academic and professional standards. In any case, experience indicates that it is wrong to claim that testing experts are not deeply aware of practicalities in their field. Developing language tests is by its nature a practical challenge, constrained by the resources available and many other operational factors which have to be considered in the process. Perhaps the task for policy makers and language testers is to understand each other’s assumptions as a way to find common ground to build on. We come back to this point below. (Scenarios (b) and (c) can be seen as lagging behind scenario (a) in terms of awareness or knowledge; engagement with issues in language assessment is the likely catalyst to move test users on from these positions.)

Another reason for test users’ apparent lack of interest in developing literacy in language testing could be that this group generally holds the power. While test takers may disagree with the implementation of a test or the interpretation given to test scores, their longer-term goal is to obtain the status or access that passing the test allows. In this position, test takers will be unwilling or unable to complain strongly about the situation, however unfair it might seem. Tests and the power relationships they create are accepted as part of the fabric of society. Test users benefit from this acceptance and may therefore see no reason for change. This challenges the position described above regarding the general assumption of personal and organisational accountability. Moreover, it may suit test users to keep their involvement in the technicalities of language testing to a minimum. Outsourcing the responsibility for decisions about test methods and scoring to a language test provider allows test users to redirect complaints and to attribute blame elsewhere if problems arise. Test users might well argue, “If this field is as arcane and impenetrable as language testers suggest, why shouldn’t we just leave it to them?”

An engagement problem

An example is given here to illustrate the mismatch between what testing experts (researchers and test developers) hope to offer and what stakeholders in tests (test takers and test users) believe they need. The example may seem obvious; the issues involved are nevertheless pertinent to the discussion, as they illustrate the inherent tensions when empirical evidence is lacking but the need persists for an “answer” to a practical problem.

Test score currency

The currency or shelf life of test scores (also often referred to as their validity period), obtained on large-scale standardised tests of a foreign/second language in particular, is a topic that illustrates some of the difficulties of engagement between different test stakeholders. The question concerns the appropriate period for which test takers' scores should be viewed as an accurate representation of their true proficiency. (Some of the perspectives set out here are drawn from a discussion of this topic in June 2018 on the electronic mailing list LTEST-L, http://lists.psu.edu/archives/ltest-l.html.) Testing experts and test developers start to address the question by considering, for example,

- test construct and design: the extent to which the test captures aspects of language knowledge or language skills (or other evidence of linguistic ability) and how it does this
- test quality: the reliability of scores as a representation of test-taker abilities (score meaning)
- up-to-date understandings of language development and attrition in research on second language acquisition.

Test takers might prepare intensively for a test focusing on language knowledge and achieve a high score but very quickly forget what they studied. Attrition is likely to vary depending on test takers' age, motivation and opportunity to use the language subsequent to taking the test, and level of proficiency attained. Testing experts will seek to base their response to the question on empirical research, to the extent that it exists.

On the other hand, test users are likely to be interested in practical issues, for example,

- whether the period of currency should start on the date of the test or the date when results are released
- whether language test results should be current when the application requiring them is submitted, when it is initially reviewed, or when the final decision is made (given that the results are most likely one element of a larger set of information to be collected and checked)
- how information provided on test certificates (test date, scores) is to be corroborated.

Test providers are often concerned with exposure to risk. If their stated period of currency for test results is long (or unlimited), eventual test-taker performance may not be of the standard expected based on the satisfactory results obtained much earlier. In this case, test users will complain about the reliability of the test. However, if the period of currency is short, test takers may have to retake the test in order to keep their language certification up to date during a drawn-out application process, incurring additional expense and running the risk on each occasion of not obtaining the results required. In this situation, test takers will complain that test providers exploit the situation for financial gain.

Nevertheless, test providers are generally expected to have a clear response to the question of the shelf life of test results because test users "in the real world" require it. The response may be as much based on the practicalities of having to maintain accurate databases of test-taker results as on the (at present rather scant) research on attrition and maintenance of language proficiency. This inability to base a decision on firmer ground might therefore encourage researchers and test providers to suggest that test users decide themselves for their own circumstances, taking into account the risks involved, similar to carrying out a local standard-setting exercise to determine the required level of test-taker performance. However, this practice is not common, perhaps because test users are unwilling to be held responsible for such decisions, as noted above. It is not an unreasonable stance to take given the imprecision and apparent dispute even among experts. Similarly, in operational terms, administrators receiving test results discover that test takers may obtain different scores if they retake the same test. They may remain unimpressed by testing experts' explanations of measurement error given that they still have to deal with the challenge of interpreting results fairly when an applicant presents several certificates containing a mix of satisfactory and unsatisfactory scores for different test components, all of which are within the accepted shelf life.

Frames of language assessment

Having used an example to foreground the tensions between testing expertise and operational practice, we now offer a possible means of improving communication with stakeholders. In an earlier paper (Pill & Harding, 2013), we made the observation that the necessary next step for language assessment literacy – in moving beyond theoretical and descriptive accounts of LAL issues – is to develop a research focus on language assessment communication. This trajectory has a precedent in the field of science, where science literacy was the precursor to an emerging subfield of science communication, a flourishing area of research in its own right, with its own dedicated journal, *Science Communication*. A focus on language assessment communication would involve developing a research base for effective techniques of engagement, drawing on wider scholarly work related to the effectiveness of narrative, metaphor, and so on. Language assessment communication would also focus on developing an evidence base to assess the outcomes of such engagement and campaigns, establishing a clearer understanding of "what works" when we communicate with different audiences.

In particular, there is great potential to explore "frames" for language assessment literacy (Goffman, 1974; Lakoff, 2014). Frame theory has been widely applied in science communication as well as in broader fields such as political science. However, it has its roots in linguistics. In this sense, drawing on frame theory to extend the notion of language assessment communication would be an approach in line with the type of McNamara-style engagement we described at the beginning of this chapter. Frames are understood to be "interpretive schemata" through which people make sense of concepts or ideas (see Goffman, 1974). Different framings of the same content may result in different interpretations; framing is therefore a key concept in understanding how messages may be tailored for different audiences in effective ways.

An important first step in taking a framing approach is to understand what frames exist around a given topic. For example, Nisbet (2010) describes a consultation in the United States which sought to understand which frames of interpretation helped best to explain why alternatives to evolution were not suitable to be taught in school science classes. As Nisbet explains:

> Although the committee had expected to find the most convincing storyline to be the authority of past legal decisions and the constitutional separation of Church and state, the data revealed that audiences were not as persuaded by this framing of the issue. Instead, somewhat surprisingly, the committee discovered that emphasizing evolutionary science as the modern building block for advances in medicine was the most effective frame for translating the importance of teaching evolution. The research also pointed to the effectiveness of reassuring the public that there was no conflict between teaching evolution and the beliefs of many religious traditions. (2010, p. 40)

A consideration of frames leads to the question of which frames currently exist with respect to language testing. One useful example is the frame of pass/fail – that is, the common interpretation of a test as something that has an inherent pass mark. As an instance of this frame in action, we cite an article published in *The Observer* newspaper in 2017 concerning the role of IELTS, an English language test, in the nurse recruitment "crisis" in the UK National Health Service (NHS). The following quotes demonstrate the centrality of the pass/fail frame in the newspaper's presentation of the issue.

> Even native English speakers with degrees struggle to pass exams, as number of applicants from EU falls to 46 in April from 1,304 last July

> Growing nursing shortages mean that the NHS has major gaps in its workforce, but this is being added to by Australians and other English-speaking nurses being turned down because they cannot pass the English tests.

> The high language requirements are reflected in a sharp drop in the number of nurses registering in the UK, according to medical recruiters, who believe that many British nurses would also fail the International English Language Testing System test (IELTS). (Tapper, 2017, p. 9)

This framing appears to relate to a conventional conceptual link between the notion of a test and of passing/failing. However, in framing the issue in this way, the newspaper report erases the important role of a key stakeholder group – in this case the UK Nursing and Midwifery Council (NMC) – in setting particular scores as proficiency standards for accreditation purposes. As a result, important issues concerning the evidence base for standard setting are largely ignored in the article until the final paragraph; the report is framed as "native speakers fail language test", rather than "NMC sets proficiency standards that some native speakers do not achieve", for instance. This example demonstrates the limiting effect of a simplistic frame. Re-framing issues to make clear the agency of the professional body would have great benefits for the wider public discussion. While the media may prefer frames that promote tension or controversy in a situation (as a way to attract bigger audiences), others, including testing practitioners perhaps, can use framing more subtly to influence their stakeholders. By framing an argument differently, taking what is important in stakeholders' eyes as the starting point, testing researchers might be able to focus and guide the decision-making process more effectively.

Framing needs to be carried out responsibly and ethically. However, we must remember that all researchers are involved in presenting results of research in a way that supports their argument. This is not viewed as manipulating or "spinning" the data, but rather making a strategic representation of findings that appeals to the intended audience. Researchers need to be able to communicate their work effectively, as they are likely to be the experts sought directly for advice, without intermediaries to help fine-tune their message. Inbar-Lourie (2017) calls for collaboration with stakeholders to "[merge] expertise in language assessment with expertise in the local context … [to] create meaningful assessment solutions to the dynamic issues that arise" (p. 267).

Concluding remarks

In this chapter we have explored and discussed some pressing issues concerning the notions of engagement, expertise and communication of knowledge. While we do not know whether Tim supports the positions we have outlined here (and no doubt he will have an insightful critique to make!), we are nonetheless inspired by Tim's approach to language testing in thinking through these issues. Engagement – in the McNamara sense – requires creativity: thinking of ways to draw other groups into our field, and looking beyond our field for new solutions to practical problems. We have developed an argument in this chapter that language testing has an engagement problem, and that stakeholders in assessment may hold competing priorities. To bridge the gap, we have suggested that an approach to communication which draws on frame theory could provide a way of telling stories about language assessment that stakeholders can interpret and process effectively. To develop this project, we need to go outside our comfort zone to consider parallel fields where communication of complexity has become a focus of research in its own right. This sense of engagement as knowledge simultaneously offered and sought has firm roots in the example set for us by Tim McNamara.

References

Deygers. B., & Malone, M. E. (2019). Language assessment literacy in university admissions policies, or the dialogue that isn't. *Language Testing*. Advanced online publication. http://doi.org/10.1177/0265532219826390

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. New York, NY: Harper & Row.

Harding, L., & Kremmel, B. (2016). Teacher assessment literacy and professional development. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 413-427). Boston, MA: De Gruyter.

Hill, K. (Ed.). (2017). Teacher assessment literacy in second and foreign language education [Special issue]. *Papers in Language Testing and Assessment, 6*(1).

Inbar-Lourie, O. (2017). Language assessment literacy. In E. Shohamy, I. G. Or & S. May (Eds.), *Language Testing and Assessment* (*Encyclopedia of Language and Education*, 3rd ed., Vol. 7, pp. 257-270). Cham, Switzerland: Springer.

Jenkins, J., & Leung, C. (2014). English as a lingua franca. In A. J. Kunnan (Ed.), *Companion to language assessment* (pp. 1605-1616). Hoboken, NY: Wiley-Blackwell.

Lakoff, G. (2014). *The all new don't think of an elephant: Know your values and frame the debate*. White River Junction, VT: Chelsea Green.

McNamara, T. (2014). 30 years on: Evolution or revolution? *Language Assessment Quarterly, 11*(2), pp. 226-232. http://doi.org/10.1080/15434303.2014.895830

Nisbet, M. C. (2010). Framing science: A new paradigm in public engagement. In L. Kahlor & P. A. Stout (Eds.), *Communicating science: New agendas in communication* (pp. 40-67). New York, NY: Routledge.

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 results (volume IV): Students' financial literacy.* Paris, France: PISA, OECD Publishing. http://doi.org/10.1787/9789264270282-en

Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing, 30*(3), 381-402. http://doi.org/10.1177/0265532213480337

Pill, J., & Macqueen, S. (2017, March). *Test construct in policy and public perception: Score-user perspectives on occupation-related language skills*. Paper presented at the American Association for Applied Linguistics conference, Portland, OR, USA.

Tapper, J. (2017, June 24). Difficulty of NHS language test "worsens nurse crisis", say recruiters. *The Observer (London, UK),* p. 9.Retrieved from https://www.theguardian.com/society/2017/jun/24/english-speaking-ovserseas-nurses-fail-nhs-too-tough-language-test

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics 29*, 21-36. http://doi.org/10.1017/S0267190509090035