Defining the language assessment literacy gap: evidence from a parliamentary inquiry

John Pill
The University of Melbourne, Australia

Luke Harding
Lancaster University, UK

Published as:

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

Available online: http://ltj.sagepub.com/content/30/3/381.short

Abstract

This study identifies a unique context for exploring lay understandings of language testing, and by extension for characterizing the nature of language assessment literacy among non-practitioners, stemming from data in an inquiry into the registration processes and support for overseas trained doctors by the Australian House of Representatives Standing Committee on Health and Ageing. The data come from Hansard transcripts of public hearings of the inquiry. Sections of the data related to language and language testing (as part of the current registration process for doctors seeking employment in Australia) were identified and coded using a thematic analysis. Findings reveal misconceptions about who is responsible for tests and for decisions based on scores in this context, as well as misconceptions about language testing procedures. Issues also emerge concerning the location of expertise in language and language testing. Discussion of these findings contributes to current debate within the language testing community (e.g. Taylor, 2009) about where responsibility lies for increasing language assessment literacy among non-practitioner stakeholders, and how this might best be achieved.

KEYWORDS: advocacy, International English Language Testing System (IELTS), language assessment literacy, language policy, Occupational English Test (OET), overseas trained doctors (OTDs)

The development and practical implementation of a language test is a complex process. The work can be technical and might be perceived by non-practitioners (i.e. those without direct experience of developing or administering language tests) as arcane and lacking transparency. The impact of a test

is, nevertheless, real and possibly far-reaching, not only for test takers but for a wide range of stakeholder groups. Those groups who use language test scores as the bases for decisions, but who are not actively involved in the construction of test materials, may make assumptions about tests, testing processes and outcomes that are at odds with what is intended or can be endorsed by the language testing community. Such misconceptions may have serious consequences for decision-making based on test scores, but have seldom been explored and indeed are rarely available for scrutiny.

It is important to understand the nature of such misconceptions in order to gain insight into those gaps in knowledge of language assessment practices which may be most usefully addressed through professional engagement. However, there has been little research to date on the level of 'language assessment literacy' displayed by non-practitioners in their conceptualizations of language assessment, making it difficult to establish how best to raise awareness of assessment practice and processes with these stakeholder groups.

Language assessment literacy

'Assessment literacy' has become a widely accepted term in educational research, and this has recently extended to the field of language testing through the term 'language assessment literacy' (LAL) (see Inbar-Lourie, 2008). As with many other 'new' literacies – computer literacy, media literacy, scientific literacy, health literacy, statistical literacy – LAL represents a broadening of the traditional association of literacy with reading and writing skills. While definitions vary depending on the context of use, language assessment literacy may be understood as indicating a repertoire of competences that enable an individual to understand, evaluate and, in some cases, create language tests and analyse test data.

To date, assessment literacy in a broad sense, and LAL more narrowly, have mostly been discussed and researched within the context of teacher or practitioner knowledge (see, e.g. Brindley,

2001; Inbar-Lourie, 2008; Stoynoff & Chapelle, 2005). However, Taylor (2009) has pointed out that LAL must extend beyond the teaching professions, as the stakeholders in language testing are diverse, and include many groups not traditionally associated directly with a need for knowledge of assessment, including:

> personnel in both existing and newly established and emerging national examination boards, academics and students engaged in language testing research, language teachers or instructors, advisors and decision makers in language planning and education policy, parents, politicians and the general public. (p. 25)

One problem, though, in attempting to increase the assessment literacy of such diverse groups is knowing what sort of information would meet the needs of different stakeholders, and allow good decisions to be made about tests and test scores. It is to be expected, for example, that the assessment literacy needs of practitioners (that is, teachers, academics and students engaged in language testing research, test designers, school principals, etc.) might be quite different from those of test takers themselves, policy makers, and the greater public. Different levels of expertise or specialization will require different levels of literacy, and different needs will dictate the type of knowledge most useful for stakeholders (see also Brindley, 2001, pp. 128–129; Taylor, 2009, p. 27).

Regarding what might be expected of non-practitioner stakeholders, Bracey's (2000) booklet, *Thinking about tests and testing: A short primer in 'assessment literacy'*, provides an interesting example. Written for a general audience, this short publication covers a range of testing terminology, including essential statistical terms, and some important issues in testing (such as 'who develops tests' and 'what agencies oversee the proper use of tests'). Others, such as Newton (2005), focus on the need for policy makers, in particular, to understand specific psychometric phenomena, such as measurement error, in order to understand better how test scores can be used and misused. In contrast to the literature on assessment literacy for educators, however, there is a distinct paucity of information concerning precisely what level of assessment literacy 'non-practitioners' should be

encouraged to reach, whether this can be a 'reduced' form of the basic training that practitioners should receive, or something else entirely, and how improving language assessment literacy (LAL) for these particular stakeholders should proceed.

One potentially fruitful area is the application of conceptualizations of literacy from other fields into LAL. An approach adopted in the fields of scientific and mathematical literacy education is the rejection of a dichotomy of 'literacy' or 'illiteracy' in preference for viewing literacy as a continuum. A seminal book is Bybee's (1997) *Achieving scientific literacy: From purposes to practices*, in which five 'stages' of literacy are identified and described: illiteracy, nominal literacy, functional literacy, procedural and conceptual literacy and multidimensional literacy. A brief description of each – transformed to relate to language assessment – is given below, based on Bybee's categories as expanded and applied by Kaiser and Willander (2005):

Illiteracy:	Ignorance of language assessment concepts and methods
Nominal literacy:	Understanding that a specific term relates to assessment, but may indicate a misconception
Functional literacy:	Sound understanding of basic terms and concepts
Procedural and conceptual literacy:	Understanding central concepts of the field, and using knowledge in practice
Multidimensional literacy:	Knowledge extending beyond ordinary concepts including philosophical, historical and social dimensions of assessment

While the exact meaning of the descriptions of each level can be contested, it may be reasonable to expect in general terms that, for example, policy makers do not require assessment literacy at the 'multidimensional level' or even the 'procedural level', but that a 'functional level' might be sufficient for engaging with, and drawing sensible conclusions from, information about language tests. However, before suitable levels of LAL for non-practitioners can be recommended and elaborated, it is necessary first to collect empirical evidence of the specific gaps in language testing

knowledge displayed by these key stakeholder groups. Through this diagnostic, needs-based approach, recommendations for LAL, and for methods of professional engagement, will have a firm basis.

Aim

The aim of this exploratory study, then, was to investigate the types of misconceptions about language testing evident in the discourse of policy makers and other concerned parties in their discussions of language proficiency assessment. It was anticipated this would permit a clearer definition of what an individual non-practitioner requires to be literate in the area of language assessment. The context, which provided a rich site for investigating the nature of non-practitioner LAL, was a broader Australian parliamentary inquiry into the registration processes and support for overseas trained doctors.

Context

The Standing Committee on Health and Ageing – one of 15 House of Representatives standing committees (43rd Parliament) at the Federal level of Australian Government – currently consists of seven members of the House of Representatives of the Australian Federal Parliament: four members from the Australian Labor Party (the governing party), one member from the National Party and two Liberal Party members (opposition parties); these members represent constituencies in four of the eight Australian states and territories. The committee's remit is to carry out inquiries which have been referred by either the House of Representatives, or any minister of the Commonwealth Government. One such inquiry, the *Inquiry into registration processes and support for overseas trained doctors,* was instigated by the Minister for Health and Ageing in late 2010. Following receipt of initial submissions, public hearings were held around Australia. The current study is based on 13

hearings held from late February to mid August 2011. Further hearings took place subsequently and the final report from the inquiry has been tabled (Standing Committee on Health and Ageing, 2012).

Overseas trained doctors (OTDs) make up almost 40 per cent of the medical workforce in Australia (Department of Health and Ageing, 2011, p. 4) and the demand for doctors remains. (This article refers to OTDs, as it is the term used by the inquiry; many involved in this field prefer to use 'international medical graduates'.) The terms of reference of the inquiry indicated that it was to investigate impediments to initial and specialist registration for OTDs who wish to practise in Australia. The inquiry aimed to find ways of clarifying and speeding up registration pathways and improving support for OTDs without compromising current standards of healthcare and patient safety (Standing Committee on Health and Ageing, 2012). Regulation of OTDs was previously a state/territory responsibility; national registration under the Australian Health Practitioner Regulation Agency and its medical arm, the Medical Board of Australia, began in July 2010. This was a major change involving new legislation and the harmonization of standards and practice across Australia. There was a great deal of confusion and complaint as new processes were implemented, in preparation for and at the time of this shift of responsibility. Figure 1 gives a simplified outline of the current process for OTD general registration following the Standard Pathway (Australian Medical Council, 2012; Medical Board of Australia, 2010).

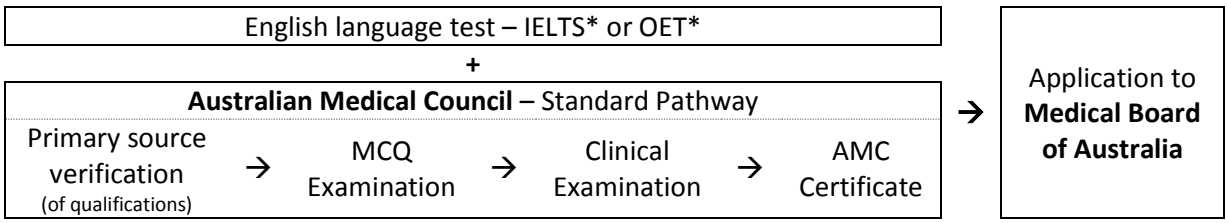

Figure 1. Simplified outline of registration process.

*Results must have been obtained within two years prior to applying for registration.

Language assessment was not the principal focus of the inquiry although meeting an English language skills standard is part of the registration process for OTDs. As with other aspects of the registration process, the English language proficiency requirements for OTDs had been criticized in a number of the written submissions to the inquiry, and thus emerged as a discussion point in the public hearings. Currently, OTDs are required to obtain four Band 7.0 scores in the International English Language Testing System (IELTS) or four B grades on the Occupational English Test (OET). Criticisms of the tests in written submissions included issues of test difficulty and relevance to the Australian context and workplace, the two-year period of currency of test results leading to some OTDs sitting a test more than once while undertaking registration, the requirement to pass all four modules/sub-tests at the same test administration, as well as particular concerns with test reliability and administration.

Knowledge of the fields of applied linguistics and language testing was certainly not expected of the committee; however, its members do have detailed knowledge of aspects of the Australian healthcare system. The inquiry was a learning process for the committee members and there is no intention in this study to represent them or any of the witnesses as unsuited to their role or deficient in any way. In extracts from the transcripts, participants' roles have been given rather than their names, as the aim is to draw general conclusions about language assessment literacy, not to appear critical of any individual.

Submissions to the inquiry and its hearings include the presentation and discussion of different models for the assessment of professional knowledge and skills for initial and specialist registration as a medical practitioner (i.e. to ascertain the comparability of medical training and qualifications). This study restricts itself to content of the public hearings related to language assessment. Examples of comments made about non-language-related assessment practices (e.g. as used by specialist colleges) are introduced in the Findings section to provide a contrast to comments made about language assessment practices.

Methods

Data set

The data set comprises the Hansard transcripts of 13 public hearings of the House Standing Committee on Health and Ageing *Inquiry into registration processes and support for overseas trained doctors*. The transcripts of these hearings were made available at http://www.aph.gov.au/Parliamentary_Business/Committees/House_of_Representatives_Committees?url=haa/overseasdoctors/hearings.htm in accordance with parliamentary protocol. Transcripts were initially presented as 'proof' copies, which were later replaced with final versions. Details of the schedule of the hearings are shown in Table 1.

Table 1. Schedule of inquiry hearings.

Place	Date	Duration (including breaks)
Canberra	25 February 2011	4 hours
Brisbane	10 March 2011	6 hours
Melbourne	18 March 2011	5 hours 15 minutes
Sydney	31 March 2011	5 hours 30 minutes
Canberra	24 May 2011	30 minutes
Canberra	31 May 2011	30 minutes
Perth	28 June 2011	4 hours 30 minutes
Canberra	5 July 2011	30 minutes
Cairns	11 August 2011	1 hour 45 minutes
Townsville	12 August 2011	2 hours
Canberra	16 August 2011	1 hour
Canberra	19 August 2011	2 hours 30 minutes
Canberra	23 August 2011	30 minutes

This collection of transcripts constitutes a corpus of approximately 27,000 words. For this research, proof copies of transcripts of later hearings were used, as final versions were not available; comparison of the two versions for earlier hearings indicates that few changes were made to the documents.

At the public hearings, a range of invited witnesses gave evidence. Although this was not under oath, witnesses were informed that providing misleading information could be considered

contempt of parliament. A total of 74 witnesses appeared over the 13 hearings in the data set, ranging from those in a 'private capacity' (often OTDs with experience of the registration process) to representatives of major stakeholder groups, such as the Medical Board of Australia (the registration body) and the Royal Australasian College of Surgeons.

A fourteenth hearing, held at the end of August 2011, involved representatives of language test providers and an 'English language academics and teachers forum'. As it was clear that the hearing would involve practitioners directly, a decision was made to deal in this paper with the initial 13 hearings only, in order to maintain the focus on the knowledge and assumptions of non-practitioners. Further hearings took place subsequently. Their transcripts do not form part of the data set for this study; a brief review of contents indicates that language testing was not a main focus of discussion.

Data analysis

A qualitative, thematic analysis of the data was undertaken by the two authors in collaboration. It involved three steps:

a. Isolating parts of the transcripts which concerned language assessment;

b. Analysing language assessment-related episodes to identify instances of misconceptions, inaccuracies or confusion about language testing concepts or the practice of language testing;

c. Categorizing misconceptions or inaccuracies by type to enable discussion.

The authors recognize that the Hansard transcripts are not made for the purposes of linguistic analysis; however, they are the official verbatim record of proceedings and are suited to thematic

analysis. All stages of coding were performed manually. This was necessary because the researchers were not present at the hearings which form the data set (with one exception), so, lacking field notes, careful reading of the transcripts and hands-on data management were required. Therefore, in the first step, all the transcripts were read through several times, and sections about language testing were highlighted. This approach was checked through a keyword search for around 20 terms which had emerged through initial read-throughs (e.g. 'English', 'language', 'test', 'assess', 'proficiency', 'exam', 'IELTS', 'OET', etc.). Through this triangulated approach, saturation in identifying instances related to language and language testing was achieved. In keeping with the tenets of thematic analysis, since a theme may be evident in a single word or constructed over a series of spoken turns between more than one participant, rigid segmentation of the data was not applied (see Fereday & Muir-Cochrane, 2006).

In the second step, the identification of instances of misconception (defined as conceptual misunderstandings or erroneous beliefs) or confusion was primarily inductive. Participants often did not overtly express misunderstanding or confusion around points relating to language testing, so the researchers relied on their own expert knowledge of language testing terminology and procedures as language testing practitioners/researchers, and on knowledge of the context itself. The researchers' claims of expertise are of central importance in judging the credibility of the interpretations made in the analysis (see Lincoln & Guba, 1985), so further details are provided here for the sake of transparency. In terms of general knowledge of language testing, one researcher has completed a PhD in the field, and the other is nearing completion of a PhD thesis. Both have taught language testing at the postgraduate level and have conducted workshops on language testing topics. With respect to the specific context of the study, both authors have worked on, or are currently working on, materials development for the OET. One author was previously the assessment manager at the OET Centre; one author has been an IELTS examiner. The first author is also at present involved in a research project funded by the Australian Research Council titled *Towards improved healthcare communication*; this is concerned with the spoken language and

communication skills required by health professionals, particularly those from non-English speaking backgrounds and with qualifications from outside Australia. This expertise formed the basis for judgements of misconceptions during the coding stage. The researchers' judgements were triangulated with reference to information in the public domain from regulatory bodies and test providers (such as test handbooks) to confirm any intuitive identification of misconceptions regarding testing procedures, test content, or other facts verifiable through documentary evidence. In addition, the availability of the full Hansard transcripts in the public domain increases the confirmability of the coding, as the raw data are open to scrutiny (Lincoln & Guba, 1985). In practical terms, the authors each reviewed half of the hearing transcripts adding annotations to indicate perceived instances of misconception. They subsequently swapped transcripts, read through this second set of transcripts, and read and responded to each other's notes. Over a series of iterations of this process, the researchers agreed on the location of instances of misconception.

Finally, in the third stage, the misconceptions were categorized by theme. Through collaborative discussion and revision of category labels, the authors reached agreement on data classification. This was achieved by following the phases of thematic analysis outlined by Braun and Clarke (2006, pp. 87–93). Specifically, at this point of the analysis, the authors followed phases 2 to 5 of Braun and Clarke's (2006) guide: generating initial codes, searching for themes, reviewing themes, and defining and naming themes. Details of the themes yielded through this analysis are provided throughout the next section.

Findings

An overall indication of the quantity of discussion of issues specifically relating to language testing in the 13 public hearings in this study is that the topic is mentioned on approximately 20% of the 405 transcript pages; in some cases, this is only a passing reference.

A general feature of the data is that different issues relating to language and language testing are mixed together by participants in the hearings. This makes providing clear headings and examples in this section difficult, as many examples given are relevant to more than one category. The reference in square brackets following each example gives the date of the hearing (yymmdd) then the page on which the example starts in its transcript. Examples are spelled and punctuated as in the transcripts; ellipsis indicates omission of data (words, phrases or turns) in an extract. The three-letter code introducing each speaker indicates his or her role, for example, OTD; explanation is provided in Table 2.

Table 2: Codes indicating speakers' roles

Code	Role
COL	Representative of a specialist college
GOV	Representative of a government department
INQ	Member of the committee of inquiry
MBA	Representative of the Medical Board of Australia
OTD	Overseas trained doctor
RCG	Representative of a recruitment body (government)
RCP	Representative of a recruitment body (private)
REP	Representative of a professional representative body
TRG	Teacher/representative of a training body

Figure 2 summarizes the findings of the thematic analysis: there are three main themes, each with two or three subcomponents. The subsequent text explains and exemplifies the themes (numbered 1 to 3) and their components (1a, 1b, 1c, etc.).

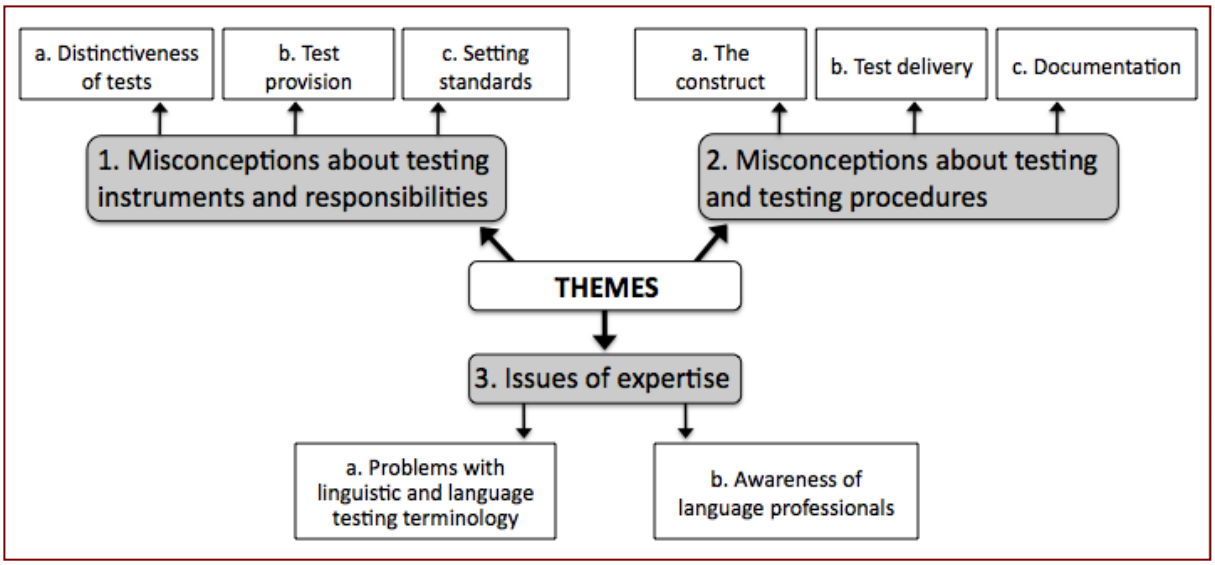

Figure 2. Map of themes.

1. Misconceptions about testing instruments and responsibilities

The first main theme in the data relates to misconceptions about testing instruments and responsibilities. Details of its three subcomponents are given below.

1a. Distinctiveness of tests.

One notable issue in the data was a misconception about the distinctiveness of testing instruments used in the registration process. Two English language proficiency tests are currently recognized in the *English language skills registration standard* set by the Medical Board of Australia (2010): the International English Language Testing System (IELTS) and the Occupational English Test (OET). However, at times these tests were conflated in the inquiry discussion. An exchange in the first hearing provides a case in point. The witness, a representative of the Medical Board of Australia, introduces the two tests and distinguishes between them (see example [1]), but the distinction does not appear to be noted and is not maintained by an inquiry committee member in a follow-up question which makes reference to one unnamed test (see [2]).

[1]

MBA: *The English language tests that the board requires are nothing to do with the board. The occupational English test – that is, the OET – or the IELTS, which is the international English language testing system are completely separate organisations.* [110225-24]

[2]

INQ: *Is there anything in the English test that relates to Australian slang?* [110225-24]

Later in the same exchange, the witness appears to follow the committee member's lead, referring only to one test, and as if it were the only test available (see [3]). This is particularly problematic, however, because the witness here seems unaware that the OET specifically claims to assess *use of language in a medical cultural setting*. The subsequent question from the committee member maintains the misconception that only one test is available. In this way, a particular distinction between the two tests that might have been of relevance to the inquiry is lost: the IELTS is a general or academic-purpose language test and the OET is a specific-purpose language (LSP) test for health professionals.

[3]

MBA: *the English language test is basic competency to speak, to listen, to write and to read. It does not deal with cultural awareness, and it does not deal with issues about the use of language in a medical cultural setting. …*

INQ: *So this test is a general fit that is being used for a particular purpose.*

MBA: *It is the standard English language testing used for all professionals in Australia who come without English as their first language.* [110225-24]

In a later hearing, a witness involved in preparing OTDs for the Occupational English Test explains more about the nature of its listening sub-test, and a committee member appears to treat this account of the OET's LSP construct as new information (see [4]).

[4]

TRG: *There are two components. One is taking notes from a health professional and patient interview. The other is a lecture of perhaps 20 minutes on a health subject that you have to write specific answers to. …*

INQ: *Even though the test is an English test, it does assume you have medical knowledge.*

TRG: *General medical knowledge – no specific medical knowledge.* [110331-28]

1b. Test provision.

A related misconception was that stakeholder groups within the medical community had jurisdiction over the language tests themselves, or a detailed knowledge of the content of the tests. Inquiry committee members appear to address questions about English language testing to whoever is appearing before them. Witnesses indicate this responsibility does not lie with them; for example, the representative of the Medical Board of Australia does so in [5] (although responsibility for language testing *policy* does lie with this body), and a representative of a specialist college does the same in [6]. (The content of example [6] is revisited in [32] below.)

[5]

INQ: *I want to ask a question about the English testing.*

… [22 lines referring to submission extracts and comments]

MBA: *The English language tests that the board requires are nothing to do with the board.* [110225-23]

[6]

INQ: *We have had evidence this morning, feedback on an English test, which said based on the feedback received that that should not have been a fail. …*

COL: *As colleges, we cannot really comment on things like the English test assessment. We have a finite responsibility which is the medical knowledge.* [110331-48]

1c. Setting standards.

Participants also demonstrated misconceptions around who was responsible for setting standards for 'pass' marks. In examples [7] and [8], the use of indefinite *they* by a witness from a professional representative body and by an OTD appears to indicate a lack of clarity about where responsibilities related to the setting of standards lie.

[7]

REP: *The reason she has been out of work is that they have changed the standards of the English test and she cannot meet the new standard.* [110225-43]

[8]

OTD: *The standard of English that they are expecting from IMGs* [international medical graduates] *is that of professorial English* [110310-16]

In [9], taken from an early hearing of the inquiry, a committee member asks a representative from a government department about standards. The phrasing of the questions seems to assume that the tests used and the standards applied are inherently linked.

[9]

INQ: *Does the department have access to the types of tests that overseas trained doctors are subject to? Do you approve of the standards that are being applied?*

GOV: *Again, it is not something that we get involved in.* [110225-12]

The respondent indicates that the responsibility does not lie with her department, illustrating the problematic nature of the inquiry discourse: it is difficult to comprehend the full picture – in this instance about responsibility for the 'passing' standard – because particular bodies take responsibility only for certain aspects of the whole. This tends to obscure rather than clarify. In

reality, while the test providers deliver the final results (as grades or bands), it is the Medical Board of Australia that is responsible for setting the standard, that is, the results regarded as satisfactory. This standard was originally approved by the Australian Health Workforce Ministerial Council (Medical Board of Australia, 2010). However, the information sought remains hidden in this exchange.

2. Misconceptions about testing and testing procedures

The second theme in the data concerns misconceptions about testing and testing procedures. Its three subcomponents are discussed in detail below.

2a. The construct.

The committee members and witnesses may reflect inaccurate or incomplete views of the constructs being tested. In [10], a committee member assumes that the test is entirely written, when in fact the two recognized English tests include speaking components. In [11], extending [7] above, a witness from a professional representative body seems to assume that a language test equates with a grammar test, when in fact both tests have four components: reading, writing, listening and speaking; and neither has a stand-alone 'grammar' paper.

[10]

INQ: *That is a written test that they must pass at a certain level.* [110705-3]

[11]

REP: *they have changed the standards of the English test and she cannot meet the new standard. So this community is being denied a doctor because probably her grammar is not very good.* [110225-43]

A committee member assumes assessment is of the content of a test taker's response rather than of linguistic performance. Example [12] concerns discussion of the OET speaking sub-test, which is audio recorded for later assessment. (The issue of subjectivity is considered in section 2b.)

[12]

INQ: *When the tape is sent back, is there a standardised answer for the questions that removes the subjectivity from it?* [110331-24]

Elsewhere, the transcripts show some witnesses distinguishing carefully between language and communication and the assessment of these two constructs, as in [13] and [14].

[13]

TRG: *There is a difference between the level of English which the college examines them at, the IELTS* [Band] *7 that they have to perform at, and what is required as true communication with the patient. … They may have no trouble in speaking English, but they do have a problem addressing it to local conditions and to the local patient.* [110310-45]

[14]

COL: *the English language component is done by the Australian Medical Council. We are then interested in their use of the language in their practice. … In practice they are assessed on how they interact with patients, peers and other health professionals.* [110225-54]

However, the distinction is not maintained consistently (see [15] and [16]).

[15]

OTD: *In the English speaking test they assess your communication skills, your fluency, a number of things.* [110331-24]

[16]

COL: *On the English language test, I realise there are concerns about people passing it. But communication is incredibly important. … I cannot emphasise enough that communication is very complex. The ability to handover patients and to make that succinct and accurate takes quite a high level of English.* [110331-39]

It should be noted that the nature of the distinction between these two constructs (in the tests in question and more generally) is a matter which has not been resolved in applied linguistics research, so the distinction is unlikely to be evident in this data set. A project at the University of Melbourne, *Towards improved healthcare communication*, is currently investigating the components of performance that medical supervisors focus on in their feedback to trainee health professionals about their interaction with patients; it seeks to understand how language and communication skills are perceived as aspects of health professional–patient interaction (see Elder et al., 2012).

2b. Test delivery.

Inquiry participants are not aware that test versions change at each administration of a test. In [17], the test taker has already described his history of English tests on his registration 'journey', explaining that he was successful with the OET, then with the IELTS, and then was asked to take the IELTS again. The inquiry committee member's question conflates the two tests (as discussed in section 1a), while the OTD's response apparently also indicates lack of awareness of the different formats of the two tests and the different content of test versions.

[17]

INQ: *That is three times that you sat the test. On those three occasions was it exactly the same test or did it vary from time to time?*

OTD: *No, it is exactly the same.* [110628-19]

Comments in the data also reveal misconceptions of how the tests are delivered. For example, in [18], there is an erroneous assumption that the tests are not available outside Australia, when in fact the two recognized English tests and the Australian Medical Council's MCQ Examination referred to are also delivered outside Australia.

[18]

INQ: *Instead of bringing people here to go through those functionary elements of qualifying for Australia such as the English test and the multiple answer test, why aren't we doing that overseas …?* [110628-13]

In [19], an OTD witness demonstrates a general mistrust of the ability of non-native speakers to rate speech. Beyond this, he shows a misconception that the interlocutor conducting the face-to-face interview assesses the OET speaking sub-test; in fact, the audio recording of the interaction is assessed by trained raters subsequently, when all test materials are sent for processing centrally in Melbourne.

[19]

OTD: *many of the people conducting the speaking tests are not Aussie, not native speakers. If they are not native speakers, how will they make the assessment?* [110331-23]

Discussion of subjectivity in the assessment of the OET speaking sub-test in [20] focuses on content (*the answers*) rather than on procedures to reduce variation, while example [21] from a witness representing a government recruitment body appears to indicate that human variation is an inherent administrative problem, using the IELTS as the example. Apart from the use of double rating in the OET (mentioned in [20]), procedures available to manage rater reliability in the IELTS and OET (e.g. through training and certification of raters, ongoing moderation and feedback,

statistical correction) are not considered, and thus the quality of these procedures never comes under explicit scrutiny.

[20]

INQ: *So there can be a degree of subjectivity when the answers are evaluated in Melbourne* [where OET rating is done]*?*

TRG: *There could be. I think that is why they normally have two people listen to it, not one – in order to cut that out.* [110331-25]

[21]

RCG: *the people who actually run IELTS have a different person testing different people all the time, so you are going to have some variation purely and simply because of the human factor*. [110628-12]

There are also concerns raised about transparency and reliability of assessment processes [22], a lack of procedures for independent review [23] and of feedback on individual test taker performance [24].

[22]

REP: *one of the things that concerns me is a lack of transparency. There seems to be no ability to judge how people are being marked. The other thing from the perspective of an educator, because I am a medical educator, is the lack of reliability of these tests.* [110225-45]

[23]

TRG: *There has to be some system whereby an independent person or group can look at that and say, 'Well, you were hard done by,' or 'No, actually the result was quite fair and you need to redo the test.'* [110331-32]

[24]

INQ: *If someone fails an exam, they need the information to know where they have gone wrong.* [110331-23]

The discussion of these issues indicates that inquiry participants are not aware of what is in fact available in the public domain from the test providers, for example, published reliability statistics and reports, or what routine measures are taken by test providers to promote reliability and minimize error, for example, double marking of performances on productive skills tests, administrative checks for clerical error, and appeals or re-marking procedures. The provision of meaningful feedback for test takers is not always possible given that these tests are proficiency measures, not diagnostic instruments (although, given the summative purpose of the tests in this context, the test results are themselves a form of feedback, interpreted through level descriptors and the test specifications). However, no extended questioning of the purpose of the tests in these terms (e.g. whether diagnostic assessments would be more useful measures in this context than proficiency tests) occurred during the surveyed hearings.

2c. Documentation.

In [25], a committee member asks whether it is possible to see copies of language tests. This request suggests a lack of awareness of the availability of exemplar tests, for example, via test providers' websites, no recognition of potential ramifications for test security if such a request were met, and no understanding of the number of different test versions that would have been used during the period concerned.

[25]

INQ: *Could we possibly request copies of the tests that have been used in the last five years by the providing agency* [110225-26]

However, there are no requests made in the first 13 hearings for documents that might have been more relevant to the purposes of the inquiry, including validation reports or other test-related materials, such as those listed as supporting documentation for test audit purposes by Bolton (2010): for example, test specifications, handbooks for candidates, research or examination reports (p. 32). We note that such information was provided by testing organizations prior to the fourteenth hearing, but, importantly, it was not *actively sought* by the members of the committee.

3. Issues of expertise

The third theme relates to issues of expertise. Its two subcomponents are discussed in detail below.

3a. Problems with linguistic and language testing terminology.

The data showed participants using imprecise terminology to discuss language or language testing issues. Examples [26] to [28] are from inquiry committee members. Terms used for linguistic phenomena may be unclear to the intended audience and/or the reader of the transcript although naturally they have meaning for the speaker. (The terms viewed by the authors as potentially unclear or ambiguous are marked with **bold** type.)

[26]

INQ: *Some languages have **a very thick intonation** within the use of English words, particularly with some of the descriptions around medical treatment.* [110225-8]

[27]

INQ: *Would you say you would like to see more emphasis on the **verbal English** – you seemed to be saying that there was a bit of a problem with **verbal English** – rather than on some of the other English testing?* [110823-6]

[28]

INQ: *We heard of one gentleman who has practised for 25 years in a particular state. He was allowed to practice, to teach students, to lecture at university and to write professional papers but he failed the English test on* ***simple things****.* [110628-27]

Example [29] provides an instance where terms that are familiar in the field of applied linguistics need to be clarified for the committee members. In this extract, an English language trainer is explaining changes in the OET to the committee.

[29]

TRG: *They have added a second component to reading, where you have to skim and scan some texts and complete a cloze –*

INQ: *Skim and scan?*

TRG: *Skim and scan means to read very quickly and then scan through to find particular pieces of information. They then complete a cloze, which is like a summary with missing bits, and you have to fill in the missing bits.* [110331-28]

Another problem was that assessment concepts may be interpreted in different ways in different contexts, limiting the shared understanding of testing-related discourse. In [30], a committee member's views on what might constitute a 'pass' grade appear to be related to personal experience with assessment.

[30]

INQ: *You speak about the three Bs and one C and how that would be a fail in the English test. … Our concept is that an A, B or C would be a pass*. [110331-22]

3b. Awareness of language professionals.

As noted at the end of section 1b, witnesses refuse to take responsibility for language testing, but they are still asked to share their views on related topics. For example, despite her initial caveat, the

representative of a government recruitment body in [31] is encouraged to discuss her perspective on the constructs of language and communication.

[31]

INQ: *One of the things that has come up a couple of times in hearings is the notion of English versus communication. I am interested in what you define as 'communication'.*

RCG: *I am not an assessor, I am just the CEO.*

INQ: *I am still interested, because it goes to another critical point that I want to ask about.*

RCG: *Patient centred care is based on effective communication. Part of that fitness for practice assessment* [a separate assessment undertaken by the witness's recruitment agency] *will assess the doctor's communication skills and the ability to engage and ask questions. It is not just about talking. It is also about the ability to listen and the ability to assess. Again, I am not the expert in that. There is a definition* [sic] *between being able to write the English language and being able to engage with a patient and, therefore, getting to the crux of what their needs are.* [110318-41]

This can be contrasted with data from the inquiry regarding views on professional and medical practice. In a discussion of how an ombudsman might assist in an appeals process for doctors seeking registration with specialist medical colleges, a college representative notes that an independent non-specialist would not be able to *rule on knowledge* but only on *process* [110331-47], as he or she would not share the specialist knowledge required. In [32], a committee member responds, seemingly to distinguish such knowledge from knowledge of language:

[32]

INQ: *I understand that at that level but I think a lot of the questions we are having are about the knowledge being disputed. We have had evidence this morning, feedback on an English test, which said based on the feedback received* [by a test taker from a language test provider] *that that should not have been a fail. … no amount of process is going to get to the heart of the question*

unless there is an opportunity for the artefact to be reviewed independently at some point. [110331-48]

The committee member does not trust the assessment of the test provider when it has been questioned by a witness (in this case a teacher involved in preparing test takers for the OET). The view seems to be that medical experts alone can judge the professional competence of other medical specialists, while there is little restriction on who may hold an 'informed' view of someone's linguistic ability.[1] For example, a committee member indirectly indicates his views on the English proficiency of some of the doctors appearing in the inquiry hearings in [33].

[33]

INQ: *we found that a lot of doctors who have come in and given evidence they said they are competent in English, and we would hear them talking to us, yet they had failed that side of it. So there was a lot of misunderstanding, I suppose, about why they had failed.* [110705-2]

On the other hand, in [34] a committee member is careful to defer to (medical) professional expertise when a witness from a specialist college notes the committee member has made a generalization about the comparability of specialist medical training completed overseas.

[34]

INQ: *I know it is very general, and for us being lay people, not medical people, we do not understand the detail, and we need a bit of clarification.* [110318-52]

However, not all witnesses feel able to comment on language testing issues: a representative of a private recruitment agency [RCP] points out that *it is very difficult as a lay person to judge how somebody is going to go in the* [English] *exam* [110318-40].

[1] There are parallels here with the devaluing of English as a Second Language teaching in school systems where K-12 English language learners are 'mainstreamed' (e.g. see Harper & de Jong, 2009).

Summary of findings

The analysis shows that individuals involved in the hearings displayed either some lack of knowledge or a degree of misconception or vagueness about: the differences between testing instruments used in the professional registration process, the bodies responsible for overseeing testing procedures, the bases for the standards upon which decisions are made and who sets the standards, the constructs the tests assess, how the tests are administered and assessed, and the types of documentation which are available in the public domain. In addition, in their discussions around these issues, participants often used imprecise linguistic terminology to discuss relevant language matters, and displayed little awareness of the type of professional expertise that might be of use in their deliberations.

Discussion

The findings which have emerged in this study are notable because many of the points associated with assessment literacy which are often prescribed for teachers or other practitioners (e.g. an understanding of standard error of measurement and other sophisticated psychometric phenomena) are not touched upon in these discussions. Rather, the language assessment literacy (LAL) problems among non-specialists – at least in this context – have been revealed to be of an elementary and 'macro' kind. That is, the fundamental misconceptions or knowledge deficits concern, for example, who is responsible for what aspects of testing, erroneous assumptions about the testing process, and what these particular language tests in fact assess. According to Bybee's (1997) framework, introduced earlier, many of the comments illustrated in the findings would indicate a standard of LAL among participants which corresponds to the first two levels of the scale: 'illiteracy' or 'nominal literacy'.

Illiteracy

'Illiteracy' in the context of LAL was suggested to mean ignorance of language assessment concepts and methods. The data demonstrate gaps in participants' knowledge of test providers, their tests and the scope of their responsibility (theme 1). There was also a lack of awareness of methods of test delivery (subcomponent 2b) and the type of documentation available to stakeholders (subcomponent 2c). Overlaying this is the broader problem of an apparent general lack of awareness that a field of expertise in language testing exists and that reference to its scholarship might be relevant to achieving the inquiry's goals (subcomponent 3b).

Nominal literacy

Bybee (1997) defines nominal literacy as when 'individuals demonstrate a token understanding of phenomena', noting that 'cognitive psychologists would call this *naïve theory* and *misconception*' (p. 83). His definition seems pertinent to the findings of this study, which exemplify problems with terminology for discussing language and language testing (subcomponent 3a), and misconceptions about testing and testing procedures (subcomponents 2a and 2b). *Naïve theory* – closely related to 'folk theory' – is a useful term for understanding the partial conceptualizations of, for example, 'subjectivity', 'reliability' and, indeed, 'language test' in the data set. In some ways, nominal literacy may be more difficult to address than illiteracy, because it requires the dismantling of a misconception before an accurate conception can be achieved. For example, it needs to be demonstrated that language tests do not necessarily take the form of grammar tests before a full justification for a modular four-skill proficiency test can be put forward.

Evidence in the data set of LAL beyond these two initial levels is scarce. One notable finding of this study is that, apart from issues of scoring reliability, the psychometric quality of the two tests was not addressed in the data set, perhaps because there were more fundamental barriers to understanding among this group of participants. This suggests that many of the recommendations

for increasing LAL among practitioners (especially classroom teachers) do not easily apply to the needs of non-practitioners, and only serves to underline the importance of further studies of this kind, as suggested in the Conclusion.

It should be noted that these first two stages of Bybee's (1997) model for the development of literacy are described mainly in negative terms. While the framework provides a more sensitive way of conceptualizing literacy development than a binary distinction alone (illiterate/literate) and is therefore helpful in defining gaps in language assessment literacy among non-practitioners, it remains problematic that the initial stages focus on defining what people lack – a deficit model – rather than on exploring what they are coming to understand. Subsequent research could investigate more specifically the process of becoming literate in this field through an acquisition model which takes a fuller account of knowledge at these first stages.

Implications

It has been argued above that it would be desirable for policy makers and other non-practitioners with some connection with language testing to have a level of LAL which is 'functional'; in other words, to possess a sound understanding of the basic terms and concepts of language assessment. It has also been shown for this data set that non-practitioners demonstrated a level of LAL that could best be characterized as 'illiterate' or 'nominally literate'. The question, then, is how best to address this LAL gap.

Towards assessment-literate non-practitioners

In a broad sense misconceptions might be resolved through greater engagement between language testing professionals and policy makers or other decision-making non-practitioner stakeholders. However, such professional engagement will not be easy. The findings of the current study were troubling as they indicated a 'professional invisibility' in the Australian context, which may be

analogous in other contexts. (It is noted again here that, for a subsequent hearing of the inquiry, representation from language test providers was sought.) For this reason, language testing practitioners must develop strategies to identify relevant policy makers and other key non-practitioner decision-makers and invite them to conferences or seminars which concern the principles and practice of language assessment. Workshops may also be designed specifically for the needs of non-practitioner stakeholders, which take as their starting point knowledge gaps around the practical aspects of language testing that were identified in this study, addressing, for example: what a language proficiency test looks like and what it aims to measure; how to find publicly available documentation on language tests and what to look for in it; how cut scores are determined and who determines them.

As well as outreach of this kind, however, non-practitioners dealing with language tests need accessible resources which provide direction for asking the right questions, particularly about fundamental matters. For example, in this study, participants might have found useful a framework providing a range of questions to ask about a language proficiency test, for example, from 'how many tests are recognized for assessing OTDs?', 'who conducts the tests?', 'what language skills do the tests cover?' and 'who decides the pass mark?' to 'what evidence is there of the test's reliability?' or 'what is a criterion-referenced test?'. Although the particular questions to ask will inevitably depend on the context of the test, the capacity to ask questions of this kind is crucial, as it provides a foundation from which discussion may grow, leading to consideration of the more typical 'universal' criteria against which the quality of language tests might be evaluated. However, it should also be noted that, while resources are already available which would most likely have helped participants in the current study, these were not consulted. Therefore creating the resources is not sufficient: they need to be promoted in order to be utilized, and this can only be achieved through the type of close engagement described above.

Towards policy-literate practitioners

As much as encouraging policy makers to become assessment-literate, there is also a need for language testing practitioners to become 'policy-literate', that is, to understand the processes of inquiries such as the one in this study, and how to engage effectively with policy makers so they are routinely invited to participate in similar strategic discussion involving language assessment.[2] (It is noteworthy that submissions by test providers appear to have been made to this inquiry only at the committee's request.)

Two areas of difficulty in discussing the need for greater policy engagement are: who will be responsible for this? and what role will they play? In the Australian context, the recently formed Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ) includes in its purpose statement an advocacy role on policy formation and advice; this body and similar international bodies, such as the International Language Testing Association (ILTA), may represent the considered view of the profession without being drawn into supporting or criticizing particular test instruments. An advocacy officer or committee might be responsible for monitoring the media and other sources of information for issues involving language testing on behalf of the association and serve as the first point of call if a member wishes to draw attention to a particular development. In the specific case under discussion, the providers of the OET and IELTS might now consider including at their routine stakeholder meetings representatives from the Standing Committee, in an effort to communicate information about their tests more effectively. (The authors do not know whether this is currently the case.) In this sense, test providers ultimately have direct responsibility for ensuring that their tests are understood by those involved in the decision-making process; they therefore also need to keep abreast of public policy discussions which concern the use of their specific instruments.

It must be recognized that language and language testing are predominantly areas in which the constructs are open to all to have an opinion. Consequently, it is likely to be difficult to find a

[2] See Lo Bianco (2001) for a discussion of the implications of 'policy literacy' for educators and researchers in the context of literacy education.

balance between when stakeholder consultation is required and when test practitioner expertise is needed. Such difficulty is likely, in turn, to create confusion about where responsibility for the decision lies. Although the situation may appear relatively easy to address through the provision of accurate information on a case-by-case basis, the line between stakeholder involvement and the need for test practitioner expertise needs to be carefully articulated and explained to promote ongoing improvement in general understanding of this issue. This would seem to be especially relevant in the context of language testing today, where the focus is generally on a test taker demonstrating proficiency in particular skills more than on his or her command of a well-defined body of content knowledge. The difficulty for the language testing profession is making a case for the need for its expertise. Although members of the profession are best placed to recognize this, it may appear self-serving to the wider population.

Conclusion

The hearings raise some difficult questions for language testing practitioners and researchers. In many of the practical contexts it is used, language assessment remains peripheral to the main task at hand, for example, applications for study or training places, recruitment to jobs in the aviation or healthcare industries. Promoting an attitude of 'leave it to the experts' would be patronizing and untenable, but it is apparent from this study that little awareness exists in the general population about concepts and processes likely to be routine in the language testing community. There is a need for a fuller understanding among policy makers, test users and the general public of the scope of language tests and the claims that can be made based on their results. Knowing where and how to start a process of education to reduce this gap requires studies like this one to be undertaken in other contexts where non-practitioners are dealing with language tests and test results, for example, immigration officials processing visa applications, or university admissions staff dealing with prospective students (see O'Loughlin, 2011). Ethnographic research could lead to a needs analysis

informing the development of training resources for such groups. Likewise, general explanatory materials which take as their starting point the perspective of non-practitioners could promote LAL in the broader population, perhaps via the web. It is clear that professional organizations, both local and international, have a role to play in developing such materials and making them accessible.

At the same time, in a general sense the language testing profession needs a stronger, more easily accessible presence for non-practitioners. To achieve this, LAL will inevitably need to extend into 'language assessment communication', comparable to the role of science communication (a sub-field in its own right), put forward as a remedy for widespread levels of scientific illiteracy (see Weigold, 2001). Language assessment communication would serve the goals of enhancing LAL and highlighting the existence of a field of expertise. To start with, this may be done through greater engagement with the media by its professional bodies (letter writing, press releases) as well as through individual members' activism (opinion pieces, blogs and web resources; e.g. see Fulcher, 2012). Once the profession is made more visible, the likelihood of its opinions being sought is higher, resulting in a more fruitful and nuanced public conversation on this important topic.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Acknowledgement

We are grateful to our anonymous reviewers for their helpful comments on earlier versions of this work.

References

Australian Medical Council. (2012, August 28). Assessment pathways. Retrieved September 10, 2012, from www.amc.org.au/index.php/ass/apo.

Bolton, S. (2010). Auditing Cambridge ESOL's Main Suite and BEC examinations. *Cambridge ESOL Research Notes, 39*, 31–33.

Bracey, G. W. (2000). *Thinking about tests and testing: A short primer in 'assessment literacy'*. Washington, DC: American Youth Policy Forum.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 126–136). Cambridge: Cambridge University Press.

Bybee, R. W. (1997). *Achieving scientific literacy: From purposes to practices*. Portsmouth, NH: Heinemann.

Department of Health and Ageing. (2011). Submission to the Inquiry into registration procedures and support for overseas trained doctors. Retrieved September 10, 2012, from http://wopared.aph.gov.au/house/committee/haa/overseasdoctors/subs/sub84.pdf.

Elder, C., Pill, J., Woodward-Kron, R., McNamara, T., Manias, E., McColl, G., & Webb, G. (2012). Health professionals' views of communication: Implications for assessing performance on a health-specific English language test. *TESOL Quarterly, 46*(2), 409–419.

Fereday, J. & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods, 5*(1), 80–92.

Fulcher, G. (2012). Language testing resources [website]. Retrieved September 10, 2012, from http://languagetesting.info.

Harper, C. A., & de Jong, E. J. (2009). English language teacher expertise: The elephant in the room. *Language and Education, 23*(2), 137–151.

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

Kaiser, G., & Willander, T. (2005). Development of mathematical literacy: Results of an empirical study. *Teaching Mathematics and its Applications, 24*(2–3), 48–60.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Lo Bianco, J. (2001). Policy literacy. *Language and Education, 15*(2-3), 212–227.

Medical Board of Australia. (2010). English language skills registration standard. Retrieved September 10, 2012, from www.medicalboard.gov.au/Registration-Standards.aspx.

Newton, P. E. (2005). The public understanding of measurement inaccuracy. *British Educational Research Journal, 31*(4), 419–442.

O'Loughlin, K. (2011). The interpretation and use of proficiency test scores in university selection: How valid and ethical are they? *Language Assessment Quarterly, 8*(2), 146–160.

Standing Committee on Health and Ageing. (2012). *Lost in the labyrinth: Report on the inquiry into registration processes and support for overseas trained doctors*. Canberra: Commonwealth of Australia. Retrieved from www.aph.gov.au/Parliamentary_Business/Committees/House_of_Representatives_Committees?url=haa/overseasdoctors/index.htm.

Stoynoff, S., & Chapelle, C. A. (2005). *ESOL tests and testing: A resource for teachers and program administrators*. Alexandria, VA: TESOL.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21–36.

Weigold, M. F. (2001). Communicating science: A review of the literature. *Science Communication, 23*(2), 164–193.

Author 1 (Corresponding author): John Pill

Position: Doctoral student

Affiliation: The University of Melbourne

Email address: tpill@unimelb.edu.au

Fax number: +61 3 8344 8032

Mail address: Language Testing Research Centre

Room 405, Babel Building

Parkville, Victoria, 3010

Australia

Author 2: Luke Harding

Position: Lecturer

Affiliation: Lancaster University

Email address: l.harding@lancaster.ac.uk

Fax number: +44 1524 843085

Mail address: Department of Linguistics and English Language

County South

Lancaster University

Lancaster LA1 4YT

United Kingdom