

Language test as boundary object: Perspectives from test users in the healthcare domain

Language Testing
2016, Vol. 33(2) 271–288
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0265532215607401
ltj.sagepub.com


Susy Macqueen

University of Melbourne, Australia

John Pill

American University of Beirut, Lebanon

Ute Knoch

University of Melbourne, Australia

Abstract

Objects that sit between intersecting social worlds, such as Language for Specific Purposes (LSP) tests, are *boundary objects* – dynamic, historically derived mechanisms which maintain coherence between worlds (Star & Griesemer, 1989). They emerge initially from sociopolitical mandates, such as the need to ensure a safe and efficient workforce or to control immigration, and they develop into standards (i.e. stabilized classifying mechanisms). In this article, we explore the concept of LSP test as boundary object through a qualitative case study of the Occupational English Test (OET), a test which assesses the English proficiency of healthcare professionals who wish to practise in English-speaking healthcare contexts. Stakeholders with different types of vested interest in the test were interviewed (practising doctors and nurses who have taken the test, management staff, professional board representatives) to capture multiple perspectives of both the test-taking experience and the relevance of the test to the workplace. The themes arising from the accumulated stakeholder perceptions depict a 'boundary object' that encompasses a work-readiness level of language proficiency on the one hand and aspects of communication skills for patient-centred care on the other. We argue that the boundary object metaphor is useful in that it represents a negotiation over the adequacy and effects of a test standard for all vested social worlds. Moreover, the test should benefit the worlds it interconnects, not just in terms of the impact on the learning opportunities it offers candidates, but also the impact such learning carries into key social sites, such as healthcare workplaces.

Keywords

Authenticity, boundary object, LSP testing, standards, test impact, washback

Corresponding author:

Susy Macqueen, Language Testing Research Centre, School of Languages & Linguistics, University of Melbourne, Parkville, Melbourne 3010, Australia.

Email: susym@unimelb.edu.au

High-stakes language tests are frequently positioned metaphorically as gates into socio-economically desirable domains. Powerful stakeholders are construed as gatekeepers who control the flow by setting a standard for entry in the form of a test score. In this article, we expand this popular metaphor to a more holistic notion of tests as boundary objects which are brought about by socioeconomic needs. We first explore the theoretical insights offered by the concept of boundary objects and we then apply it to a case study of the relevance and impact of a high-stakes specific-purpose language test, the Occupational English Test (OET).

Boundary objects

Boundary objects are dynamic, historically derived mechanisms, which develop in order to maintain coherence between intersecting social worlds (Star & Griesemer, 1989). They are constructed from a pooling of information at the intersections of the participating social worlds (Star & Griesemer, 1989). Boundary objects often become *standards*. Broadly speaking, standards are stabilized classifying mechanisms which cohere aspects of social structure and are pervasive in modern life; the ‘raw materials out of which social order is constructed’ (Busch, 2011, p. 42). Standards come to be recognized as the way things are (done) in societies, such as what to do at a red traffic light, the length of a phone number and how to gain citizenship of a country. It is a feature of the standard-saturated modern world that the genesis and reasoning behind classificatory systems become forgotten as standards are naturalized in social use (Bowker & Star, 1999; Busch, 2011).

Tests of Language for Specific Purposes (LSP) can be considered *boundary objects*, developing as they do from a social mandate, such as a need to ensure a national health-care workforce can communicate to a level that is deemed adequate for patient safety and workplace efficiency. This need brings about a classificatory infrastructure; LSP tests categorize and differentiate people in terms of how ready they are to participate linguistically in specific discourse communities. They do this by using standardized language sampling methods, which, ideally, elicit language forms and structures that are relevant to the target social sphere and are judged using methods that are also congruent with ways in which communication is evaluated in the target domain (Douglas, 2000; Jacoby & McNamara, 1999). How meaningful or relevant these samples and the attendant classificatory infrastructure (e.g. criteria, scoring, cut scores) are to the people within the boundaries of the target sphere, as well as to those hoping to enter it, is a key question for research on the impact of tests.

Test impact and authenticity

Fulcher and Davidson (2007) argue for an effect-driven approach to test development whereby impact on stakeholders not only informs design decisions but is the basis for ongoing monitoring of test validity. They define effect-driven testing as ‘a blending of procedure with fundamental virtue’, and advise that we ‘think about the intended beneficial impact as the test is built, and be willing to knock it down when things change’ (p. 177). Standards theorist, Lawrence Busch, similarly advises that the balance of benefits

and costs be considered at the design stage of a standard. Ultimately, standards must not affect social conditions in ways which restrict the democratic rights of those judged by them (Busch, 2011; Shohamy, 2004). In the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) sections are devoted to avoiding negative consequences of testing and to determining the likelihood of intended benefits, both of which can be informed by stakeholder perspectives (including test-taker experiences) as well as the congruence with target domain. To this end, impact studies ‘enable us to re-evaluate and make explicit not just the standards we promote but the very view of language we take for granted’ (Davies, 2008, p. 440). Thus, the role of test impact research is to challenge the taken-for-grantedness of the standard by interrogating its rationale and examining its effects.

The meanings of LSP tests, like those of other standards, arise in interactions between social worlds, and it is by studying the adequacy of their between-ness that we can evaluate whether an appropriate informational balance has been struck (Star & Lampland, 2009). Validation, as Moss (1998) puts it, ‘entails ongoing evaluation of the dialectical relationship between the products and practices of testing, writ large, and the social reality that is recursively represented and transformed’ (p. 11). What might be termed ‘adequate authenticity’ is at the centre of such ongoing evaluations. By this, we mean (following Messick and others) a degree of authenticity that is most likely to promote beneficial test preparation practices that are relevant to the future domain of practice (Basturkmen & Elder, 2004; Douglas, 2000; Messick, 1996). An ‘adequate authenticity’ also reflects the nature of standards more generally, which are ‘always already incomplete and inadequate (compared to some ideal)’ (Star & Lampland, 2009, p. 14). However, authenticity alone is not sufficient in test design. Needs analysis processes prioritize domain tasks and skills which should be represented on a test (Douglas, 2010). Thus, effect-driven LSP test design involves considering the relative importance of test tasks as well as their degrees of authenticity, with test effects and test content requiring periodic checks to ensure that the test and its social worlds are operating coherently.

Such investigations involve determining the relationship between the test construct and the criterion domain, both of which are generated in the test development process (McNamara, 2006). As McNamara has phrased it, ‘the criterion domain is itself an abstraction, a construct in need of validation itself’ (p. 36). Many performance tests, for instance, are underpinned by models of communicative competence (see Purpura, 2008). Target domain communication may be shaped by frameworks for communication developed from professional philosophies which are upheld through education and professional policy, such as patient-centred care which is espoused in many English-speaking healthcare workplaces (e.g. Makoul, 2001). Messick (1996) articulated the ideal of authentic assessments as those which ‘pose engaging and worthwhile tasks (usually involving multiple processes) in realistic settings or close simulations so that the tasks and processes, as well as available time and resources, parallel those in the real world’ (p. 243). Even with careful analysis and sampling of the practices of the real world, the development process of a standardized test can bleach tasks and processes of authentic characteristics to the extent that the test tasks that result may be inadequate simulations (Lewkowicz, 2000). It is helpful for the sake of investigating test effects to consider the ‘target language domain’ as an abstraction that informs test design in contrast to

real-world practices, which are too contextually complex and locally detailed to be applied in a standardized test. Impact studies can raise awareness of the degree of commensurability between theoretical models which contribute to the test instrument and those which underlie domain practices as well as the congruence between the generalized target language domain and actual communication practices in relevant local contexts.

Stakeholder voices

Stakeholder perceptions can be insightful in terms of both the degree and the importance of authenticity (Lewkowicz, 2000). LSP test stakeholders, whether test-takers or institutional score users, are cast as 'domain experts' by virtue of their workplace experience and/or training. Accordingly, they are seen as uniquely placed to articulate what matters for effective communication in the relevant language use situation (Douglas, 2000; Douglas & Myers, 2000; Jacoby & McNamara, 1999; see also other papers in this issue), to give feedback on the appropriateness of a test's content and format (Brown, 1993; Lumley & Brown, 1996), to decide what level of test performance is sufficient for effective functioning in this domain (Lumley, Lynch, & McNamara, 1994; O'Neill, Buckendahl, Plake, & Taylor, 2007; and other papers in this issue) and, in general, to determine whether a test is achieving its intended effects.

Three relevant studies of test-taker perceptions involve comparisons between the OET, an LSP test, and the IELTS,¹ a general/academic English proficiency test. These two tests have been used interchangeably as healthcare registration standards in Australia and New Zealand for some years. In an exploratory study by Elder (2007), 53 non-native-English-speaking health professionals took both the OET and IELTS in close succession. A comparison of band score/grade frequencies on each test revealed that while overall 'pass' rates on each test were comparable (i.e. similar numbers of candidates received an overall Band 7 or above on the IELTS and a B grade or above on the OET sub-tests), the candidates viewed these two tests rather differently, as indicated in their post-test survey responses. The OET reading, writing and speaking components were perceived to be somewhat easier than their IELTS counterparts, whereas the reverse was the case for OET listening, which posed challenges on account of the Australian-accented input and the need to take notes while listening (see below for a fuller description). The perceptions of relative difficulty were not, however, reflected in obvious score differences on the two tests, to the extent that these scores were comparable. While many candidates accepted the IELTS as a valid measure of their *general* English proficiency, there was almost unanimous agreement that the OET was a better measure of their *professional* competence and hence more relevant to their needs. This suggests that health specificity in the boundary mechanism allows for greater coherence between worlds. However, a study by Read and Wette (2009) showed that practical issues also are weighed up in the evaluation of test options. In their study, which elicited comparative feedback on the relevance and utility of the IELTS and OET from 23 overseas-trained health professionals enrolled for a six-month English course for health professionals in New Zealand, there was a slight preference for the OET on the grounds of its health-related content and the fact that candidates could, at

that time, retake single components of the test where they had failed to reach the required standard, rather than resitting all four sub-tests as required for the IELTS. The IELTS, on the other hand, was seen as more affordable and easier to prepare for given greater availability of practice materials. However, the authors point out that the candidates' perceptions varied somewhat over time such that 'by the third interview many health professionals seemed to have realized that both the IELTS and OET are tests of English language proficiency, not of clinical knowledge and therefore the medical content of the OET became less important as a factor determining their choice of a test' (Read & Wette, 2009, p. 39).

In a study by Oh (2010), the views of 18 international medical graduates (IMGs) were canvassed after they had taken and passed the OET and had experience in the workforce, as is the case with the current study. It is interesting to note that all but the most able participants (who had passed the OET after only a one-week period of preparation) held the view that preparing for the exam helped them develop effective communication skills for their profession. Opinions of test tasks and their relevance to the medical domain were generally favourable, although the listening test was criticized for its failure to represent non-native accents encountered in the workplace. Some also held the view that the note-taking demands on the listener were too heavy and distracted from the processing of new information. Among the limitations of the OET noted by respondents was its failure to sample communication with colleagues and its lack of general English content relevant to daily conversation. Taken together, these studies show that test-takers are affected by multiple aspects of a standard (specificity, timing, cost, availability of practice materials) and that perceived effects may vary depending on the degree to which the candidate engaged with test preparation. In the present study, we sought to build on these studies of test-taker perceptions by eliciting multiple stakeholder perspectives on the test as a boundary mechanism.

The Occupational English Test (OET)

The Occupational English Test (OET) is 'an international English language test that assesses the language and communication skills of healthcare professionals who seek to register and practise in an English-speaking environment'.² In this paper, we draw from a larger study (Macqueen, Pill, Elder, & Knoch, 2013) to exemplify the effects and meanings of the OET as a boundary object. We focus on two of the four OET sub-tests, the Speaking and the Listening sub-tests, because these offer complementary perspectives of a productive skill and a receptive skill and because these skills are common to many communicative language performance tests.

Method

We explore, through the qualitative analysis of interview data, the nature of the test as it is perceived by participants with diverse relationships to the test: those involved in setting the standard (professional bodies), those who occupy the target domain (supervisors in the workplace) and 'boundary crossers' as classified by the test (successful test-takers who have achieved the standard and since entered the target domain). The largest OET

candidate groups are doctors and nurses; it is these professions that formed the focus of the study. Taking the view that the test, as boundary object, performs a bridging function as people transition to the Australian healthcare context, our overall aim was to explore how well the test fits between the stakeholder worlds it serves to connect. Specifically, we posed the following questions.

1. How do stakeholder groups with diverse relationships to the test perceive the fit between the test content and language use in the workplace?
2. Is the language of the workplace adequately represented in the test?
3. What are the effects of the test as boundary object?

The interviews were semi-structured to enable an in-depth, contextualized exploration of relevant issues as they arose. They were approximately 30 to 45 minutes long. Interviewees were asked to comment on the relevance of the OET content to the workplace and on the language demands of the workplace for users of English as an additional language. Participants who had not taken the test (i.e. board members and most supervisors) were provided with a description of each sub-test and sample test papers so they could comment on specific tasks. Past candidates were also asked to fill out a background questionnaire, which included information about their OET scores, test preparation methods, linguistic and cultural background, professional occupation and experience, and general opinion of their test experience.

The questionnaire and semi-structured interview data were filtered through a grounded analytical process to elucidate stakeholder perceptions of test effectiveness, relevance and impact. The interview data were transcribed and analysed thematically by two researchers using NVivo, computer software designed to support qualitative coding. The analysis was an iterative, collaborative process in which descriptive themes were identified and interpreted in relation to the research aims (Miles, Huberman, & Saldaña, 2014).

Participants

As mentioned above, three stakeholder groups were interviewed for this study. These are described below.

Senior representatives of two professional bodies. One senior representative of each national board, the Medical Board of Australia and the Nursing and Midwifery Board of Australia, was interviewed. Board representatives were approached first so that they could advise on possible recruits for the sample of workforce participants (below) and give a general overview of the role of language testing in the registration process. On their advice, participants were specifically sought from urban and rural areas as well as in the field of aged care.

Supervisors of EAL (English as an Additional Language) health professionals. Five senior nurses and six senior doctors were interviewed, all of whom had several years' experience in their profession. These participants were educators and/or in senior institutional roles. Two nurses were co-ordinators at aged-care facilities (one urban and one rural) and

one of these had considerable experience in remote hospital contexts. One doctor was also a past OET candidate and so was able to offer the perspectives of both past candidate and supervisor.

Successful OET candidates who are currently employed in their health professions. The past candidates interviewed were employed in a variety of contexts and had a broad range of experience. The sample comprised 10 nurses (seven working in hospitals and four in aged care, including one in both) and eight doctors (two working in medical centres and six in hospitals, at least three of which were regional). The participants all had work experience in Australian healthcare contexts, enabling them to comment on workplace communication practices; the mean time the past candidates had worked in Australia was just under 12 months. The mean time the past candidates had worked in their professions in any context was 4 years 8 months. All past candidates had achieved a grade B or higher on all four sub-tests of the OET, with most achieving grade B. Thirteen of the 19 past candidates had also taken the IELTS and were therefore able to compare their experiences of the two tests. One of the past candidates was recruited for the study as a supervisor. The participants spoke a variety of first languages, indicating the diversity of overseas-trained health professionals who practise in Australia: Arabic (3), Bangla (1), Burmese (1), Chinese (4), Farsi (1), Filipino (1), Gujarati (1), Hindi (1), Korean (1), Punjabi (2), Russian (1), Sinhala (1).

Findings

As we are broadly concerned with the test as a boundary object, our analysis focuses on the potential for worthwhile learning to occur in the transition to the workplace. The findings are presented using themes which are derived from a theoretical convergence of work by Star and others on boundary objects and boundary crossing (Akkerman & Bakker, 2011; Star & Griesemer, 1989; Star & Lampland, 2009) as well as thinking which has informed the practice of LSP testing around task authenticity, specificity of language in relation to the target language use domain, and the interaction between language and content (Douglas, 2000, 2001; McNamara, 1996). The first section deals with the delineation of the target language use domain (i.e. where the boundary lies). The second considers individual professional trajectories in relation to the boundary (i.e. boundary crossers). The last two sections consider the quality and scope of specific parts of the test as boundary object: how sufficiently the tasks and texts are perceived to cover the critical aspects of the workplace language domain. We have presented aspects of the larger study (Macqueen et al., 2013) that elucidate a view of the test as boundary object. Voices from each of the participant groups are represented across themes, as relevant.

Extracts are attributed using a code indicating whether the interviewee is a supervisor (S) or past candidate (PC), the interviewee's profession (Nurse or Doctor) and an identification number. Board representatives' contributions are identified within the text.

Drawing the boundary: Communication skills versus language proficiency

A persistent theme relating to perceptions of the test as a between-worlds boundary object was the distinction between the healthcare construct of 'communication skills'

from the applied linguistics construct of 'language proficiency' (see also Manias & McNamara, this issue). The relevance of this distinction was evident in responses from all groups of participants. Both board representatives viewed the English language registration standards (OET and IELTS) purely in terms of general English proficiency. The nursing board representative explained that the test result gives employers assurance that the registrant has adequate non-technical language proficiency with the understanding that any new employee would need to learn context-specific language. The medical board representative outlined the limits of the language test purpose:

we don't think it's testing clinical communication skills...there are many native English speakers who don't have good communication skills in a clinical context...if the testing is congruent with practice, that's terrific, but we shouldn't be relying on that.

Similarly, a senior doctor who was also a past candidate observed that aspects of spoken communication such as 'whether you are caring, whether you are sympathetic, whether you understand what the patient's need is' should not be tested in a language test because they are 'well above the level of basic communication skills' (S-Doctor 6). Another past candidate described workplace communication as requiring 'something on top of the language itself... it is a lot about interpersonal skills in hospital so you shouldn't just do the task, you are not just task orientated, you should also relate to, build a rapport with the patient' (PC-Nurse 3). Senior doctors and nurses noted the importance of checking a patient's comprehension as a part of the health professional-patient interaction, while past candidates observed that body language was an important component of workplace communication that appeared to be absent from the OET. Participants' comments indicate that a distinction between speaking and listening skills is not particularly relevant to them; 'communication skills' appears to combine the 'macro' skills that are considered separate for test purposes.

Despite the fact that the test does not focus explicitly on assessing clinical communication skills, some past candidates assumed this was part of the test construct. As one informant observed, the specific-purpose characteristics of the test are not easily separated from 'pure' language assessment: 'it is also a test on understanding, listening skills... whether you are able to provide reassurance, you have a reassuring technique, whether you have advising skills, it is a little bit more comprehensive communication skills' (PC-Nurse 1). PC-Nurse 9 explained that the first time she took the OET she focused only on language. When she failed, she reconsidered her stance and decided for her subsequent attempt that focusing on communication ('to listen to the patient, what she really wants') was what was required. Indeed, several informants indicated how such attention to appropriate communication in context helped them succeed on the OET.

As suggested in the nurses' observations above, participants' understandings of what 'listening' entails offer insight into how clinical communication skills might be delineated from a work-ready quality of language proficiency. For the senior staff in the study, there are two overlapping listening skills: (1) listening as a clinical strategy and (2) listening comprehension as part of clinical process. S-Doctor 4 explained the importance of listening as a clinical strategy: 'often listening is the biggest task that you have to do when you are talking to a patient - it is not all about information transfer from you

to them, particularly ... if a patient is upset about something or you have got to tell them something terrible.' This kind of psychologically strategic listening, however, may be seen as a professional skill, inappropriately tested as a language skill; the same doctor concluded that this type of listening was less a language skill than a clinical one: 'It is more content in a way because that is the appropriate clinical skill at this point, just to actively listen.' All senior staff described aspects of the second type of listening, as a part of clinical process, that is, for comprehending and recording information. Links were made between the listening test task involving note-making while listening to a consultation and this kind of 'everyday' listening, as exemplified by S-Doctor 6, a medical educator and past candidate: 'you need to see whether they actually pick up all the important essential information from that patient.' Related to this was the observation made by several participants that the test does not represent the range of accents that might be encountered in the Australian healthcare workplace.

Professional journeys

Striking an informational balance between worlds is a critical feature of boundary objects (see discussion above) which relates to the interaction between content knowledge and language use that characterizes LSP assessments (Douglas, 2000). One perspective that was revealed in the data was that of the individual professional journey (or boundary crossing) of the test-taker and how relevant they perceived the test to be in relation to their professional development and goals.

The questionnaire data showed that almost all past candidates had chosen to take the OET over IELTS (which is not a specific-purpose test) because it was related to their profession. Similarly, in the interviews, informants generally characterized the OET as complementary to the professional exams they must also take (e.g. the Australian Medical Council's clinical examination for PC-Doctor 6). Two participants indicated that they perceived the OET to be easier because it was more relevant to their professions. As one wrote in the questionnaire, 'IELTS can be very tricky esp. when talking about rare kinds of animals, or fish, etc.' Participants suggested that dealing with an unfamiliar medical topic on the OET was preferable to dealing with an unfamiliar topic from any field (in the IELTS). In the view of PC-Doctor 6, the IELTS imposes a double burden of content and language knowledge:

For any English exam there are two major informations you need to know: the English itself, and some general knowledge. If I sit for exam with OET – so I already [have] the medical knowledge, the general knowledge. So the rest of the exam will be English so I can handle it later on. But in IELTS I should handle both of them.

For some candidate stakeholders, whose professional experiences may not have been based in an ethos of person-centred care, preparing to take the OET helped set their expectations of professional life and the workplace: 'it was a learning experience not just about language but also a little bit about the culture ... of nursing in English speaking contexts' (PC-Nurse 10). PC-Nurse 1 stated that the OET is 'more functional ... more practical ... we are more familiar'; the test is related to the working environment: 'I chose

OET because it is more related to what I do and what I want to do here in Australia and that is the purpose of it' (PC-Nurse 6).

Despite this perception of greater professional relevance, individual professional trajectories were not necessarily congruent with the test content. For example, one past candidate nurse (PC-Nurse 3) commented that some role-play scenarios in the speaking test would most likely not be encountered by a junior nurse with little prior experience. Similarly, PC-Nurse 6 stated that the test content is 'beyond what I do in terms of my communication [in current employment]... it is ahead of what is basic'. On the other hand, another nurse pointed out that the scenarios are limited to 'medical ward and surgical ward' and so do not match his particular experience as a psychiatric nurse (PC-Nurse 4). There was a general acknowledgement that test-takers will perform better if the topic they are given in the role-play is something they are familiar with; therefore, professional knowledge is seen to facilitate performance even if that content knowledge is not being assessed directly in the test.

Issues of specificity of topic also arose regarding the listening test. PC-Doctor 7 pointed out that the test topics are not all directly related to the medical field and include other areas of health practice; she would prefer it if they were always closely linked with medical practice. In contrast, a nurse participant (PC-Nurse 3) noted that the topics are rarely from a nursing context; she felt the topics are quite similar to some of the topics used in the IELTS and did not relate particularly to her experience of workplace communication. Another doctor felt the consultations were too often based in general practice. These diverse opinions highlight the fact that the specificity of LSP tests is a matter of both the *degree* of technicality/presumed knowledge as well as the *type* of content knowledge represented. However, while the latter may have been imprecise in terms of some individual professional trajectories, in general, the specificity of the test materials was viewed very positively when compared with more general test materials (in the IELTS) which straddle the same professional boundary.

Note-taking

Related to the degree and quality of specificity is the extent to which the test captures features of the real world (i.e. whether it achieves adequate authenticity). A prominent example of this arose when participants reported how the combination of note-taking and listening in one listening test task compared with typical workplace configurations. This part of the listening test requires candidates to take notes while listening to a consultation. An example is a doctor consulting with a patient about his back injury; candidates supply notes under headings such as 'history of recent incident,' 'details of pain,' 'treatment,' 'further details of problem,' 'doctor's initial comments,' and 'results of the examination.' All candidates, regardless of profession, do the same task.³

While this was perceived as an authentic combination of skills, its timing and structure drew some criticism in terms of authenticity. Several senior professionals reported that notes were generally not taken during the consultation process, where clinical strategies such as attention to visual cues and rapport-building took precedence. For example:

You should maintain rapport with the patient. Look at them when they are talking so you can pick up on cues including if you are examining them you look at their face if you are feeling

their abdomen to see if they are in pain because they won't tell you. You have to take all that in and if you are busy writing notes you can't do that. (S-Doctor 4)

Comparing the workplace and the OET listening test note-taking task, which uses an audio-recording played once only, a past candidate noted how real-life consultation listening was easier because information can be checked and repeated back (PC-Doctor 3). Describing work in an Emergency Department, PC-Doctor 5 explained: 'you ask a question and they answer ... and literally when you start writing the patient pauses and waits for you to finish your writing, then when you raise you head they continue'.

Although note-taking might not ideally occur while a patient is talking during a consultation, several senior professionals reported that note-taking *while* listening to a conversation did occur in ward rounds with senior team members. Other challenges of the real world are indicated in the description by S-Doctor 4, a medical educator in a major urban hospital, who describes the 'aural nightmare' in which such notes might be taken:

on that ward round they [international medical graduates] won't be given enough time to do the tasks that they are supposed to be doing. So they will be balancing the folder, writing out the notes about what is happening while simultaneously being expected to write a radiology request or a pathology request, look at the medication chart, look at the fluid balance chart, take in everything that is being said, write their own notes for what jobs they might have to do later. Already you see this is an impossible task, meanwhile the interaction between the patient and the registrar will be taking place in a crowded and busy ward room with other patients having loud conversations with their relatives or just yelling, machines will be beeping and maybe another ward round is taking place in the same room. It is an aural nightmare really, if you can't filter then you are going to have a lot of trouble.

In this description, we glimpse the complexity of real-world 'listening,' complete with layers of sounds, visual cues, multiple critical tasks, and hierarchical relationships. The test texts/tasks must enable sufficient evidence for valid inferences to be made about an individual's potential performances in interaction with such a complex and dynamic environment. Further, senior doctors observed that real-world note-taking is not indiscriminate: 'you will be writing notes about what was talked about and what was decided' (S-Doctor 5). This is not the same type of engagement required in the test, in which the headings are provided and no prioritization or selection of information is required. In relation to this, S-Nurse 4 observed that one way to demonstrate that effective listening had taken place would be for candidates to provide appropriate headings for their notes:

It seems like [test designers] have almost given people the answers by putting the headings [given in the answer booklet] ... whereas I would imagine that if I am a Registered Nurse with some background in practice that I should be able to come up with my own headings from the dialogue.

Despite this discrepancy between the real-world activity and test task, note-taking to record what has been done was considered important in much nursing work and the test was readily linked to workplace interactions by participants. In this regard the test task produced beneficial effects for boundary crossers. Examples were PC-Nurse 6, who recognized the test 'helping me jot down notes, make abbreviations like really quickly

during like telephone call from the doctor,' and PC-Nurse 10, who had learned to focus only on important words as a strategy for understanding. Preparation for the test task, therefore, has the potential to increase relevant skills for the workplace in some ways, but falls short of other workplace behaviours such as holding consultation information in one's memory for subsequent prioritization and note-making. This demonstrates the potential role of the test as boundary object in promoting learning as well as the learning opportunities omitted.

Role-play

The speaking sub-test has a role-play format (see Elder, this issue, for a description). Perhaps owing to the strong tradition of simulation assessment in Australia, doctors and nurses at supervisory level generally indicated that patient–clinician role-plays were highly relevant to workplace communication and viewed as a familiar assessment method (see also Woodward-Kron & Elder, this issue). Both medical and nursing educators reported that communication skills are inherent in professional assessments that involve simulated patients. One medical educator reported that professionals who have trained overseas may only have been assessed through written exams. Therefore, having this assessment genre represented in the language test is congruent with the healthcare context and its learning culture.

Past candidate informants recognized a general closeness between the OET role-play situations they practised when preparing to take the test and their workplace experience; one commented that 'sometimes I found that my patients really talked like from OET preparation' (PC-Nurse 7). They noted clinical aspects of the role-plays and how these are familiar to them from their everyday practice of healthcare. They also pointed out how OET role-play scenarios raised their awareness of cultural and professional expectations, e.g. about privacy of information (PC-Nurse 4). Especially among the nurses interviewed, it appears that preparation for the OET speaking test helped with confidence: 'I need to greet the patients and I have to give explanation and I give advice and reassure patient. So that ... preparation of OET helped me, everyday conversation with patients, and also yeah I could practise lots of different topics when I prepared OET' (PC-Nurse 7).

The boundary was further blurred in terms of authenticity of roles. One medical educator (S-Doctor 4) observed that role-play assessments in general offer 'a clean example of an interaction, which is unrealistic.' In real life, he noted, patients are usually more challenging in some regard; for example they are talkative, hostile, or confused. However, another doctor pointed to the complexity of involving more realistic examples: 'you [test developer] would have to pick which kind of difficult communicating you were going to test them on. Are you going to test them on a person who is reticent and quiet and you are trying to draw them out or the verbose person or the angry one?' (S-Doctor 1). Inviting in this quality of interaction implies even greater pretence, or acting ability, in the simulation. Furthermore, the co-operation of interlocutors was seen as facilitative in the test situation: PC-Doctor 1 described how, 'if [the interlocutors] ask questions and they're involved in the conversations and they sort of participate in the test actively, it really helps motivate us to do our role as well, appropriately as well, and it's healthy because it means you flow nicely' (see also Woodward-Kron & Elder, this issue, about the consensual orientation of the OET role-plays).

An observed omission in the speaking test was interaction between health professionals. As a senior doctor observed, there is a process of synthesis and translation between registers that occurs in inter-professional interactions:

I think it would be interesting to hear them [test candidates] take the same data from the doctor–patient [interaction] and have a doctor–doctor. Either listening to them make a referral or go to their senior saying, ‘I am worrying about this man/lady because of x’. Because you have got to synthesise very quickly and put it all in a very different language. It is [using] straight medicalese to your colleagues. (S-Doctor 1)

PC-Nurse 7 also noted the need she had for more complex, professional language when handing over a patient to her nurse-in-charge, a task she found challenging. Several participants considered patient handover a critical workplace interaction not represented in the speaking test. One senior nurse in an aged-care context gave this example of a communication failure between EAL-speaking staff:

I will find that we will hand something over and then the next day we will come back and it has gone through two nurses and we get something completely different ... Just say [nurse name] might hand over a certain resident is on antibiotics for a UTI [urinary tract infection] and we will get back that they are on antibiotics for a chest infection. (S-Nurse 1)

As PC-Nurse 3 pointed out, ‘lots of information can be missed from one person to the next person to the next’ in real-life handover situations.

Although inter-/intra-professional interaction was noticeably absent for participants, one effect of preparing for the professional–patient role-plays was awareness of the need to learn suitable lay terms: ‘I have to convert the professional terms into the layman terms and I was surprised at how many words are there ... my vocabulary got richer when I was preparing for speaking [test]’ (PC-Doctor 7). Similarly, S-Doctor 5 observed that EAL-speaking professionals may not know sufficient non-medical jargon for communicating effectively with patients.

The overall view of the speaking test was that it promotes the learning of one necessary register in the health professional’s transition to workplace (patient-talk), but it doesn’t adequately capture another (inter-/intra-professional talk). In our final section we will draw on these findings to illuminate how the test as boundary object offers constrained learning opportunities for individuals who encounter it.

Discussion

In this paper, we have operationalized the notion of tests as boundary objects through investigating stakeholder worlds and inquiring about the degree to which the test is operating coherently between them. Documenting the perspectives of various groups with different relationships to the test in this way invites difference and power into the discussion of test use so that tensions may emerge (Holland & Reeves, 1994). In answer to our first research question about how stakeholder groups with diverse relationships to the test perceive the test content in relation to language use in the workplace, one such tension emerged. This is represented in Figure 1, which shows the test as Boundary Object (BO), which is, to some extent, composed of two language use models: a work-readiness level of

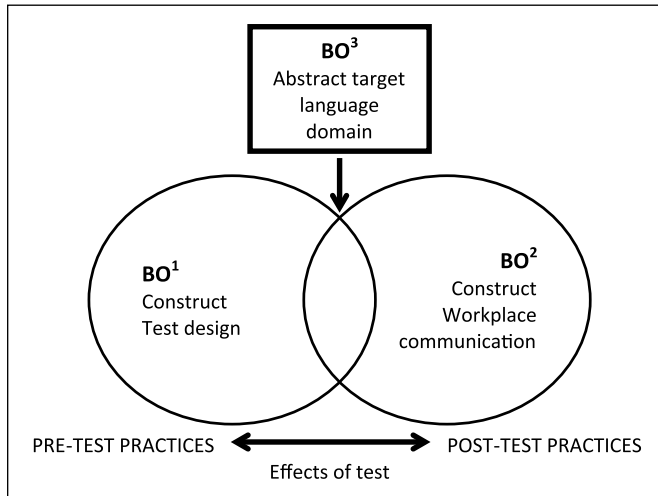


Figure 1. Boundary object (BO) manifestations and test effects.

language (macro-skill) proficiency for healthcare purposes on the one hand (BO¹), and aspects of communication skills for patient-centred care on the other (BO²). Although for some stakeholders the two constructs are kept relatively separate (e.g. the professional board view), for some past candidates the two clearly overlap, with patient-centred communication strategies emerging as role-play test strategies that are accorded some importance (BO³). This blending may be an effect of a degree of incommensurability between the applied linguistics and health professional notions of communication. That is, the separability of speaking and listening skills—quite standard in applied linguistics (e.g. Buck, 2001)—is impossible in communication frameworks associated with patient or person-centred care (e.g. Levinson, Lesser, & Epstein, 2010; McCormack & McCance, 2006).

These two ‘constructs’ have had relatively separate geneses and disciplinary trajectories (Skelton, 2013), with language tests such as the OET traceable to models of communicative competence and the concept of language proficiency (described in McNamara, 1990, 1996). The coherence of these two constructs is critical in fulfilling the social mandate to regulate equitably the entry to and quality of the healthcare workforce. Such coherence implies adequate authenticity whereby test method does not impede meaningful, relevant sampling, and sampling leaves nothing important out (Messick, 1996). The answer to our second research question about the adequate representation of workplace language is summed up in the general evaluation of the OET given by PC-Doctor 8: ‘It’s a test, but it’s relevant.’ The basic congruence with professional needs echoes the finding by Elder (2007) that health professionals viewed the OET as being far more relevant to their professional needs than the IELTS. In the examples provided in this paper, however, there are observed gaps and differences between the real world and the test materials.

Sufficiency of overlap between the constructs (adequate authenticity) can be exemplified in the listening note-taking task. Although note-taking is a skill that is perceived as relevant to the target domain, there are inauthentic features such as the practice of writing while listening (also discussed in Elder, 2007), the absence of the need to be selective

when note-taking, the absence of visual cues, the unnaturally quiet test environment, and the inability to direct the interaction to check information as necessary. Despite this test method interference, participants saw the task configuration as sufficiently representative of real-world practices. Furthermore, it is arguable that test tasks which bear closer resemblance to the tasks of the target domain are more readily critiqued by domain insiders. For example, the activity of note-taking, which was clearly linked to daily healthcare practice by participants, elicited extensive opinion. It seems unlikely that such fine-grained domain-related criticisms would surface if the discussion were about a task with a more general text topic and item types that were less related to real-world behaviour. Indeed, it may be seen as testimony to the relevance of the boundary object that it can be considered in some depth in insider terms.

A key feature of real-world communication which participants noted to be absent from the OET speaking test was intra-/inter-professional communication. This inadequacy relates again to the LSP compromise between content knowledge and language skills. Patient-talk can be mobilized in a language test because of its relationship to the language and experience of the general population (including speaking test interlocutors and raters). Inter-professional and (to a possibly greater extent) intra-professional talk calls for further delving into insider technical knowledge for which 'between-world' interlocutors and raters are ill-equipped. This brings into focus the practical nature of standards, remembering that we are speaking of boundary objects (standards) in a very general sense as mechanisms for making elements of the physical and social worlds work together. Once established, as 'naturalized' mechanisms, standards become increasingly inert and less able to change (Busch, 2011).

A further discrepancy observed by past candidates in this study was that the accents represented in the test are not representative of the broad range of accents they would typically encounter in Australian healthcare workplaces (also reported in Oh, 2010). This case illustrates the need for a standard to evolve so that, through judicious sampling, it better reflects the range of accents typically encountered in the workplace. Doing so would increase the likelihood that candidates were able to adapt to the different varieties they may later experience in the workplace (Harding, 2014). This issue relates to the fact that high-stakes standardized tests are themselves standardizing mechanisms: tests categorize people into levels or standards but they also standardize language use by sanctioning 'appropriate' varieties, for instance, as judged by raters trained to recognize deviation from norms deemed appropriate (Shohamy, 2007; Young, 2012).

Despite these perceived insufficiencies, past candidates viewed preparation for an LSP test as far preferable to investing in a general proficiency test because of the beneficial effects they perceived in terms of developing their profession-related communication skills. Indeed, they were critical of test topics for not being specific enough. To return to the notion of 'effect-driven testing' (Fulcher & Davidson, 2007) and our final research question about the effects of the OET, there were several test-related practices which informants found beneficial once they entered the workplace, such as the development of note-taking skills, the perceived similarity between test role-plays and real-world interaction with patients, and the modelling of patient consultations and of other aspects of healthcare practice in Australia (e.g. maintaining patient privacy). The overall test experience appeared to provide some level of enculturation into the Australian healthcare context. Thus, LSP test effects might be best evaluated in terms of post-test practices, as depicted in Figure 1.

Figure 1 depicts how test effects have been operationalized in this study. Using the notion of test as a boundary object that sits between worlds, we have observed two, overlapping constructs: BO¹, the test construct (and its underlying theory of language and test method overlay) and BO², the workplace communication construct (and its underlying theory of practice/communication as well as actual practices in real-world contexts). The abstract target language domain is then represented by BO³, the perceived overlap between the test construct and the workplace construct. Note here we distinguish between BO³, the abstract target domain sampled by the test to which test *scores* apply, and the ‘real world’ or actual, un-sample-able communication, to which test *effects* apply. As with all standards, which are always incomplete and inadequate (Star & Lampland, 2009), complete overlap is impossible, even undesirable, as the local specificity of real-world sampling would be unfair. However, just as a test validation argument strives for completeness and coherence, the degree of overlap is crucial in the extrapolation inference which entails an ‘ambitious leap’ from a statistical sampling procedure to the real world (Kane, 2013, p. 28).

The between-ness of LSP tests is their key characteristic, well-signified in the popular ‘gatekeeping’ metaphor. The boundary object concept allows us to approach validation as a dialectic pursuit (Akkerman & Bakker, 2011; Moss, 1998), a negotiation over the adequacy and effects of the test for the social worlds it connects. Adopting this concept to characterize test impact frees us from seeing the test purely as propelling influence backwards (implicit in the notion of washback) and instead sees it as exerting influence potentially into all stakeholder worlds, across the whole candidate trajectory, from test preparation into the workplace (described as ‘washforward’ by van Lier in Bailey, 1996; see also Tsagari & Pavlou, 2013). In Figure 1, therefore, we show test effects as bi-directional, with the quality of pre-test and post-test practices both warranting investigation. Test developers, providers and enforcers should aim for test effects which are constructive for the worlds the test interconnects, not just in terms of the pre-test learning opportunities it offers candidates, but also the potential impact such learning has on their future workplaces.

Acknowledgements

We would like to acknowledge Cathie Elder’s input in the literature review and her insightful assistance with the shape of this paper. Thanks also to Tim McNamara and Barbara Hoekje for very helpful feedback.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The research study which this paper draws on was funded by the Occupational English Test Centre, Melbourne.

Notes

1. International English Language Testing System.

2. www.occupationalenglishtest.org
3. A sample listening test can be found at www.occupationalenglishtest.org/display.aspx?tabid=2425

References

- Akkerman, S. F., & Bakker, A. (2011). Boundary crossing and boundary objects. *Review of Educational Research*, 81(2), 132–169.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, 13(3), 257–279.
- Basturkmen, H., & Elder, C. (2004). The practice of LSP. In A. Davies & C. Elder (Eds.), *Handbook of applied linguistics* (pp. 672–694). Malden: Blackwell.
- Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: The MIT Press.
- Brown, A. (1993). The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing*, 10(3), 277–301.
- Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.
- Busch, L. (2011). *Standards: Recipes for reality*. Cambridge, MA: MIT Press.
- Davies, A. (2008). Ethics, professionalism, rights and codes. In N. H. Hornberger & E. Shohamy (Eds.), *Encyclopedia of Language and Education*, Vol. 7: *Language Testing and Assessment* (pp. 2555–2569). Berlin: Springer.
- Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.
- Douglas, D. (2001). Three problems in testing language for specific purposes: Authenticity, specificity and inseparability. In C. Elder & A. Davies (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (Vol. 11, pp. 45–60). Cambridge: Cambridge University Press.
- Douglas, D. (2010). *Understanding language testing*. Abingdon: Routledge.
- Douglas, D., & Myers, R. (2000). Assessing the communication skills of veterinary students: Whose criteria? In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 60–81). Cambridge: Cambridge University Press.
- Elder, C. (2007). OET–IELTS benchmarking study. Report to the OET Centre: Language Testing Research Centre, University of Melbourne.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Abingdon: Routledge.
- Harding, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly*, 11(2), 186–197.
- Holland, D., & Reeves, J. R. (1994). Activity theory and the view from somewhere: Team perspectives on the intellectual work of programming. *Mind, Culture, and Activity*, 1(1–2), 8–24.
- Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes*, 18(3), 213–241.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.

- Levinson, W., Lesser, C. S., & Epstein, R. M. (2010). Developing physician communication skills for patient-centered care. *Health Affairs*, 29(7), 1310–1318.
- Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing*, 17(1), 43–64.
- Lumley, T., & Brown, A. (1996). Specific-purpose language performance tests: Task and interaction. *Australian Review of Applied Linguistics*, 13, 105–136.
- Lumley, T., Lynch, B., & McNamara, T. (1994). A new approach to standard-setting in language assessment. *Melbourne Papers in Language Testing*, 3(2), 19–40.
- Macqueen, S., Pill, J., Elder, C., & Knoch, U. (2013). Investigating the test impact of the OET: A qualitative study of stakeholder perceptions of test relevance and efficacy. Melbourne: University of Melbourne, Language Testing Research Centre.
- Makoul, G. (2001). Essential elements of communication in medical encounters: The Kalamazoo consensus statement. *Academic Medicine*, 76(4), 390–393.
- McCormack, B., & McCance, T. V. (2006). Development of a framework for person-centred nursing. *Journal of Advanced Nursing*, 56(5), 472–479.
- McNamara, T. (1990). Assessing the second language proficiency of health professionals. PhD thesis, University of Melbourne, Melbourne.
- McNamara, T. (1996). *Measuring second language performance*. London: Longman.
- McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, 3(1), 31–51.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241–256.
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Thousand Oaks, CA: SAGE Publications.
- Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17(2), 6–12.
- O'Neill, T. R., Buckendahl, C. W., Plake, B. S., & Taylor, L. (2007). Recommending a nursing-specific passing standard for the IELTS examination. *Language Assessment Quarterly*, 4(4), 295–317.
- Oh, H. (2010). Overseas health professionals' perspectives on the OET. Unpublished master's thesis, University of Melbourne, Melbourne.
- Purpura, J. (2008). Assessing communicative language ability: Models and their components. In N. H. Hornberger & E. Shohamy (Eds.), *Encyclopedia of language and education*, Vol. 7: *Language testing and assessment* (pp. 53–68). Berlin: Springer.
- Read, J., & Wette, R. (2009). Achieving English proficiency for professional registration: The experience of overseas-qualified health professionals in the New Zealand context. *IELTS Research Reports*, 10.
- Shohamy, E. (2004). Assessment in multicultural societies: Applying democratic principles and practices to language testing. In B. Norton & K. Toohey (Eds.), *Critical pedagogies and language learning* (pp. 72–92). Cambridge: Cambridge University Press.
- Shohamy, E. (2007). Language tests as language policy tools. *Assessment in Education*, 14(1), 117–130.
- Skelton, J. (2013). English for medical purposes. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–7). Oxford: Blackwell. doi: 10.1002/9781405198431.wbeal0379
- Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, translations and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science*, 19(3), 387–420.
- Star, S. L., & Lampland, M. (2009). Reckoning with Standards. In M. Lampland & S. L. Star (Eds.), *Standards and their stories: How quantifying, classifying and formalizing practices shape everyday life* (pp. 3–24). Ithaca, NY: Cornell University Press.
- Tsagari, D., & Pavlou, A. (2013). The nature and impact of textbook-based EFL vocabulary tests on teaching and learning. *Research Papers in Language Teaching and Learning*, 4(1), 59.
- Young, R. F. (2012). Social dimensions of language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 178–193). Abingdon: Routledge.