

Drawing on indigenous criteria for more authentic assessment in a specific-purpose language test: Health professionals interacting with patients

Language Testing
2016, Vol. 33(2) 175–193
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0265532215607400
ltj.sagepub.com


John Pill

American University of Beirut, Lebanon

Abstract

The *indigenous assessment practices* (Jacoby & McNamara, 1999) in selected health professions were investigated to inform a review of the scope of assessment in the speaking sub-test of a specific-purpose English language test for health professionals, the Occupational English Test (OET). The assessment criteria in current use on the test represent a generalized view of language and are concerned with OVERALL COMMUNICATIVE EFFECTIVENESS, FLUENCY, INTELLIGIBILITY, APPROPRIATENESS OF LANGUAGE, and RESOURCES OF GRAMMAR AND EXPRESSION. The research study focused on healthcare consultations between trainee health professionals and patients. Educators and supervisors observed these interactions and subsequently provided feedback on trainees' performances. The assumption was that, in their comments, educators would give information pertinent to trainees' acculturation to the expectations and behaviours of the profession, that is, to "what matters" to practitioners. Thematic analysis was undertaken to establish the aspects of performance that matter to health professionals in these contexts. Data for each profession were coded independently. Clear similarities across the professions became apparent as themes emerged. An exploratory conceptual model of what health professionals value in the consultation was developed, comprising three focal areas: foundation, performance and goals of the consultation. Findings from the analysis provided an empirical basis for the generation and definition of two additional, professionally relevant criteria for use in the OET speaking sub-test – CLINICIAN ENGAGEMENT and MANAGEMENT OF INTERACTION – and of a checklist of performance indicators to be used to train assessors in applying the new criteria. This process of developing, through close analysis of domain experts' commentary, test criteria that are potentially more authentic to the target language use situation is novel and may be replicated effectively in other specific-purpose language testing contexts.

Keywords

Health professional–patient interaction, healthcare communication, indigenous assessment criteria, language proficiency, LSP testing, Occupational English Test

Corresponding author:

John Pill, American University of Beirut, PO Box 11-0236/English, Riad El-Solh, Beirut, 1107 2020, Lebanon.
Email: tp04@aub.edu.lb

This paper reports on a study that investigated the *indigenous assessment practices* (Jacoby & McNamara, 1999) of professionals in a workplace context. The purpose of the study was to develop assessment criteria for use on a test of spoken language that were more closely aligned to the criteria used by practitioners in the target language use situation. Jacoby and McNamara (1999) recognize that “it is not a transparently simple matter for applied linguists to derive formal, more general performance assessment criteria from the highly specific, complex, and content-based comments raised in an indigenous assessment setting” (p. 235). The setting in this research is the interaction between health professionals and their patients in consultations and the assessment criteria are for use in the speaking sub-test of the Occupational English Test.

The Occupational English Test

The Occupational English Test (OET) is a specific-purpose language test for health professionals who have English as an additional language and professional qualifications from jurisdictions outside Australia and who are seeking registration to practise in Australia. See Elder (this issue) for a description of the test’s purpose and use and of the format of the speaking sub-test, which is the focus of the present study. The aim of the speaking sub-test is to allow a judgement to be made regarding whether test takers have the English language and communication skills to participate in clinical interactions with patients in the target language use situation.

The criteria used to assess performance on the current OET speaking sub-test are oriented to language in a generalized sense: OVERALL COMMUNICATIVE EFFECTIVENESS, FLUENCY, INTELLIGIBILITY, APPROPRIATENESS OF LANGUAGE, and RESOURCES OF GRAMMAR AND EXPRESSION. They were based on criteria used in the US Foreign Service Institute (FSI) speaking test but framed more “communicatively” (McNamara, 1996, p. 106). For example, the criterion named INTELLIGIBILITY replaces the original FSI criterion ACCENT. McNamara (1996) describes tests with criteria of this type as *weak* performance tests, in which “the focus is on *language performance* ... [and] the purpose of the assessment is to elicit a language sample so that second language proficiency ... may be assessed” (p. 44, emphasis in original). He contrasts this with *strong* performance tests, in which “performance will primarily be judged on real-world criteria, that is, the fulfilment of the task set” (p. 43). For health professionals seeking registration in Australia, assessment of language is required by law to be carried out separately from assessment of professional competence and skills (McNamara, 1996, p. 40). The OET must therefore be defined as a weak performance test.

Almost since the introduction of the OET, concern has been expressed about its effectiveness: on the one hand, test takers who fail the test may feel unjustly penalized for deficiencies in language that they consider have little, if any, effect on their ability to undertake their professional role while, on the other, supervisors and colleagues of some of those who pass express dissatisfaction with the language and communication skills of incoming health professionals (Lumley, 1995; McNamara, 1996; Wette, 2011). This dissatisfaction may be because the passing standard for the test is set too low, but it may also indicate that the test is not measuring sufficiently those aspects of performance that matter to health professionals in the workplace and is therefore unable to identify test takers who lack them. The

construct of the OET speaking sub-test represented in its assessment criteria perhaps does not accurately reflect the demands of the workplace in the opinion of experienced practitioners. This issue is anticipated in Jacoby and McNamara's (1999) claim that "using the traditional four linguistic skills to delineate special-purpose performance is inadequate to capture real-world communicative cultures and activities" (p. 234).

*Indigenous criteria*¹

The notion of *indigenous assessment practices* was described by Jacoby (1998) in her doctoral study of a group of academic physicists giving feedback to each other at regular sessions in which they rehearsed conference presentations. Although the group included speakers of English as an additional language, with particular individuals being described as having observable limitations in their language ability, the feedback rarely considered their language errors differently from those made by first-language speakers; the attention of the group was principally directed towards other issues apparently perceived as more pertinent in the context. Using methodologies of conversation analysis and ethnography, Jacoby shows how novice members of the group are socialized into the practices and expectations of a particular professional activity. The group members "call upon their own indigenous members' methods of practical reasoning and on a rich inventory of tacitly known assessment criteria" (Jacoby, 1998, p. 311). Jacoby and McNamara (1999) subsequently reflect on and develop this issue, noting that specific-purpose performance "is by definition task-related, context-related, specific and local" (p. 234). They consider how best to capture for assessment what matters in different contexts and suggest discourse analysis as a way forward (see Woodward-Kron & Elder, this issue).

Several researchers have attempted the task of capturing for assessment what matters in various contexts, some using the term *indigenous criteria* explicitly and others implying it. Brown (1993) developed test tasks and criteria for a test of Japanese for tour guides working in Australia who had Japanese as an additional language. She involved representatives of the tourism industry in the test development process as domain experts, and the resulting assessment scheme included real-world, non-linguistic issues in a "task fulfilment" criterion. Douglas and Myers (2000) compared the assessment of communication skills of a group of veterinary students by applied linguists and by veterinary professionals, finding that the applied linguists focused on language and communication aspects of the construct that they assumed they were measuring, whereas the domain experts appeared more concerned with the students' content knowledge and the professionalism of their relationship with the clients. Elder (1993) studied the assessment in English-medium high schools of trainee teachers of mathematics and science who had English as an additional language. In a later article (Elder, 2001), she reflects on this and other research into determining teachers' readiness for the workplace and notes how subject experts apply different criteria from applied linguists and have a different orientation to the assessment process.

Erdósy (2005) considered the indigenous assessment practices employed by a Canadian university professor on students' written assignments from an undergraduate course in modern Chinese history. He found that language errors of students with English as an additional language were discounted as long as meaning remained clear. However, in more linguistically demanding tasks, such as essays, poor language control

was taken into consideration, especially when combined with weak content (p. 184). Language proficiency was therefore one of the professor's indigenous assessment criteria but not the dominant one; a certain minimum threshold was assumed. Erdősy (2009) stresses the highly contextualized nature of these assessment criteria and appears pessimistic about deriving from them criteria that could be more generally relevant. However, the limited scope of Erdősy's study (involving one professor and a small group of students) cannot preclude the existence of criteria relevant to, say, undergraduate writing across several disciplines in Arts and Humanities. In another study that seeks to develop context-relevant assessment criteria, Fulcher, Davidson, and Kemp (2011) create a "performance decision tree" to structure the assessor's grading of a test taker's performance. To establish the set of binary decisions and additional questions that constitute the tree, they draw on a variety of previous applied linguistics studies of the particular domain as a means to determine essential characteristics of interaction between travel agents and customers. This process makes "what matters" in the context transparent, but their assessment scheme would clearly need to be amended to be relevant in other contexts. The ability to generalize from specific, contextualized instances of performance and draw inferences regarding likely performance in similar contexts is most important in testing (Chalhoub-Deville, 2003; Douglas 2001a), but practical studies in which such a process is attempted are lacking.

It would seem vital in an exploratory study of indigenous assessment criteria to allow the richest possible representation of insiders' views to be elicited. The applicability of these views to the particular focus of the investigation – that is, in the present study, the review of criteria used in a language test – is only to be considered subsequently. Jacoby (1998) does not explicitly impose any constraint on the scope of her seminal research; however, in his book on specific-purpose language testing, Douglas (2000) appears to suggest, perhaps understandably, that investigation be focused on issues of language and communication. Research in applied linguistics involves engagement with other disciplines, by definition. On the one hand, it may be the case that the perspective of the insiders in another discipline underplays the importance of language or pays no attention to it explicitly. In a review of studies of language use in the workplace, Roberts (2005) reminds applied linguists that "language ... does not account for *all* the conditions of working life" (p. 131, emphasis in original). On the other hand, it is also likely that applied linguistics researchers bring a common perspective to their work, orienting themselves particularly to issues of language and communication. Their goals must therefore be, first, to avoid imposing a view of language on the context being investigated that is limiting and, second, to be willing to reconsider their understanding of the scope of language to align it better with the perspective of the disciplinary insiders. These points seem especially pertinent in the field of specific-purpose language testing, which is a natural locus of research into indigenous assessment criteria and the complex relationships among language, communication and performance in context. In this field, researchers and test developers must seek to extend the boundaries of what appears practicable in terms of test design and administration.

The study

The current study is the first to investigate the indigenous assessment criteria of health professionals in the context of their interaction with patients. A multidisciplinary

research team sought to capture the aspects of performance valued by practitioners in medicine, nursing and physiotherapy. Drawing on these findings, the researchers then considered how the scope of the assessment scheme in current use in the OET speaking sub-test – and therefore also the underlying test construct – could be refocused to include more of what is valued by health professionals, while the test remained a language test (i.e., not dealing directly with issues of professional competence and skills). New assessment criteria for the test were developed and trialled (see O’Hagan, Pill, & Zhang, this issue). The three professions were chosen because they are among the main sources of the OET candidature.

The first phase of the project involved collecting data from which to determine the health professionals’ indigenous assessment criteria. In a similar way to Jacoby (1998) and Erdősy (2005), the researchers sought to access situations in which feedback was routinely articulated by experts for the benefit of novices. Through the process of novices being socialized into workplace expectations and behaviours, what matters to health professionals is made explicit. Duff (2008) notes how newcomers to workplaces “learn (or are expected) to think, act, speak, and write more like their experienced co-workers and mentors” (p. 260). It is recognized that other groups will also have opinions on what matters in health professional–patient interaction, including, perhaps most importantly, patients themselves. However, the intention of the current study was to consider the views of health professionals only.

This paper describes the process of data collection and analysis for this exploratory, qualitative study in the professions of medicine and nursing.² The aim of the paper is to demonstrate that an empirical procedure can be followed to establish authentic values of a particular group of professionals in a certain context as enacted through their indigenous assessment practices and to draw on these values to develop assessment criteria for use in a language test.

Research questions

This paper therefore considers two research questions:

1. What aspects of performance matter to health professionals – specifically, doctors and nurses – in health professional–patient interaction?
2. Which aspects of performance found to matter to health professionals are amenable to assessment on a language test?

Method

Data for the study were collected from April 2010 to September 2011. This paper concerns itself with data from two main sources:

1. a series of profession-specific workshops,³ each of 60–90 minutes and audio-recorded, in which educators and clinical supervisors watched video recordings of trainees in their profession consulting with patients and/or simulated patients and provided oral commentary on the trainees’ performance;

2. (for medicine only) a set of reports each written by a visiting educator summarizing his or her feedback after observing a doctor on a general practice (GP; family medicine) training programme as the doctor consulted with patients at routine clinics during a semester-long supervised placement.

These data for medicine and nursing were collected in parallel with data for physiotherapy. The suitability of different sources of data in terms of their authenticity was considered, but decisions were not informed directly by the findings reported in Elder and McNamara (this issue). Table 1 summarizes the feedback data collected for medicine and nursing.

Participants

The participants providing feedback in the workshops were educators and clinical supervisors who had been invited by the project team. They were involved in teaching at a large university in Australia. Some were affiliated with metropolitan hospitals, others practised in primary healthcare. Participants were selected based on professional qualification, experience, and availability. It was thought that professionals involved in trainee education would be practised in articulating their views clearly. Participants' previous exposure to trainees with English as an additional language educated outside Australia was not considered in the selection process although some participants had considerable experience of providing support for such trainees. The language background of participants was also not considered; none are thought to be from a non-English-speaking background.

For the second main source of data, the written feedback to GP trainees, reports were selected from the electronic database used to record and manage trainees' progress by the training provider. They were written for each visit made by an external educator to the clinic where the trainee was placed to practise under supervision. Two visits were made during each of three training placement periods. Reports were chosen for the study from those trainees who were active in the database in the four most recent annual cohorts and who were identified as having English as an additional language (see below). They were selected to represent different stages in the training course; two or three reports were selected for some trainees although it was subsequently found impractical to consider longitudinal development. Owing to the method of selecting reports, only a minimal profile of the visiting educators could be created (e.g., gender was determined based on their given names). Some educators visited the same trainee more than once; others visited several of the trainees. The dataset also includes reports about the same trainee from more than one educator. No information was available about the GP educators' language backgrounds.

Ethics approval for the study was obtained from the university human ethics research committee and the GP training provider. All workshop participants signed consent forms and understood the voluntary nature of their involvement. Trainees in the video stimuli played in the medicine workshops and originally recorded for a separate study gave consent for the recordings to be used. Consent was not sought from participants in the published materials used in the nursing workshops.

Table 1. Summary of feedback data collected for medicine and nursing.

	Source of data*	Venue for data collection	Feedback providers	Stimulus for feedback	Feedback data collected	Word count
<i>Medicine</i>						
1	Workshop <i>MED-wk1</i>	University medical school	8 educators/clinicians	Videos A & B: trainee with SP	Oral comments; indirect	7250
2	Workshop <i>MED-wk2</i>	University medical school	5 educators/clinicians	Videos C & D: trainee with SP	Oral comments; indirect	4850
3	Written reports <i>MED-Rxx</i> [xx = trainee ID]	Routine clinics at health centres	18 visiting educators	27 GP trainees, each with 1-5 patients	46 structured reports; direct	30,000
<i>Nursing</i>						
4	Workshop <i>NUR-wk1</i>	University department	5 educators/clinicians	Videos E & F: trainee with SP	Oral comments; indirect	8800
5	Solo interview <i>NUR-wk2</i>	University department	1 educator	Videos E & F: trainee with SP	Oral comments; indirect	3300
6	Workshop <i>NUR-wk3</i>	University department	2 educators/clinicians	Videos E & F: trainee with SP	Oral comments; indirect	6800
7	Workshop <i>NUR-wk4</i>	University department	7 educators/clinicians	Videos E & F: trainee with SP	Oral comments; indirect	6100

Notes: SP = simulated patient; GP = general practice (family medicine); direct/indirect = whether feedback was provided directly to the trainee or indirectly for researchers only; word count is approximate.

*References given in italics are used in the Findings section.

Stimuli for feedback

Where possible, the stimulus for which feedback was given involved a trainee with English as an additional language, as it was thought the feedback providers were more likely to comment on the topic of language and communication in such cases. Exceptions are noted below.

In the research workshops, each video stimulus lasted between three and nine minutes. The pair of recordings used at each workshop was selected to provide a variety of task and trainee performance. The videos (A–D) used in the workshops for medicine involved doctors who had English as an additional language and had met the English Language Skills Registration Standard of the Medical Board of Australia. They were preparing for the Australian Medical Council's Clinical Examination by roleplaying

clinical scenarios with simulated patients. The two video stimuli used in all the nursing workshops (E and F) came from published training materials. They were simulations involving participants with English as their first language. No suitable material involving nurses with English as an additional language was available.

All GP trainees discussed in the written reports had English as an additional language, as far as it was possible to determine. This information was not recorded directly in the training provider's database, so trainees were selected based on information available about their country of birth and the university where their primary medical qualification was completed. In cases where there was uncertainty about a trainee's language background, the report was excluded from the dataset. It is not known whether the visiting educators were informed of a trainee's language background (or, if so, whether they would have viewed it as pertinent).

Instruments

Workshop participants were asked to take notes as they viewed each video recording, as if preparing to give feedback to the trainee observed. The sheet provided for note-taking had the general prompts "stronger/weaker aspects of performance"; the intention was to elicit as wide a range of comments as possible without particular focus on language or communication. Participants then used their notes to offer comments on each video orally in turn. The notes were collected but were not used in the analysis, as the audio recordings of the workshops were clear and provided the information in a more contextualized form.

For the written reports on GP trainees, the electronic report template had five sections for completion by the visiting educator: "Communication skills and the patient–doctor relationship"; "Applied professional knowledge and skills"; "Population health and the context of General Practice"; "Professional and ethical role"; and "Organizational and legal dimensions." A final section invited "Overall comments and/or issues needing attention." It is clear that the educators would be likely to align their comments with the section headings, affecting any analysis of what they perceive as important. Nevertheless, the headings are themselves an indication of the values espoused by practitioners in the context of GP training.

Audio recordings made at the workshops were transcribed, and names and all other references in the transcripts and reports that might identify participants were removed for the analysis process. Word counts for each of the transcripts are given in Table 1.

Analysis

Thematic analysis (Braun & Clarke, 2006; Forman & Damschroder, 2008; Galaczi, 2014) was applied to the data for this study to draw out the indigenous assessment criteria of three health professions in the context being investigated. (This paper reports findings for two of these professions.) The method focused on the content of the data, with themes initially being derived inductively from the participant commentary. Themes were then refined and consolidated through an iterative process. Three coders worked in parallel, each taking the data from one of the three health professions. They met with

members of the research team on multiple occasions during the coding period of several months to share their insights and problems and to develop the coding scheme as necessary. Data from the initial workshops were coded first, and the draft coding scheme was then applied to data collected later (subsequent workshops and the written reports for medicine) and modified to apply to them equally well. Inter-coder reliability was therefore ascertained through the methodological process itself and confirmed in exercises in which the same data were coded by all three coders and cross-checked. Data saturation was reached in the coding process, although not all themes were found in the data for each profession. Participants' own words were used whenever possible to create the "label" identifying each theme.

Findings

The themes and their inter relationships established through analysis of the data allowed a model to be developed exploring what is valued by health professionals in a consultation with a patient. The themes in the model are presented in Figure 1 and described here under three headings, relating to the performance of the consultation, its goals, and the foundation on which the performance is based. Brief examples from the data for medicine and nursing are used to illustrate themes. The reference in square brackets following each example gives its source (see Table 1 for the references used; the reference is followed by, for reports, a number indicating the placement period, then a line number locating the start of the example in the transcript or report). In workshop extracts involving more than one speaker, researcher 1 is labelled R1, participant 1 as P1, etc.

The performance of the consultation

The participants' comments on trainees regarding the performance of the consultation fell into three main thematic groups – *Clinical skills*, *Communication skills* and *Practitioner skills* – linked by a common set of *Interactional tools*. A further term – *Consultation skills* – was also used, referring to *Clinical skills* and *Communication skills* together.

The category of *Clinical skills* was of great importance to the educators. They commented on trainees' ability to apply professional knowledge and skills as the consultation unfolded (coded with the themes *Content* and *Physical examination*) and on how trainees provided structure for the consultation and particular tasks within it (*Organization*).

Content

You offered full STI [sexually transmitted infection] screening after unprotected intercourse. You explained the tests, but didn't offer pre-test counselling for the HIV [human immunodeficiency virus] test. [MED-R02-1-12]

Physical examination

as an experienced nurse there are a lot of things you pick up from that in terms of the body temperature, whether the skin is clammy [NUR-wk3-104]

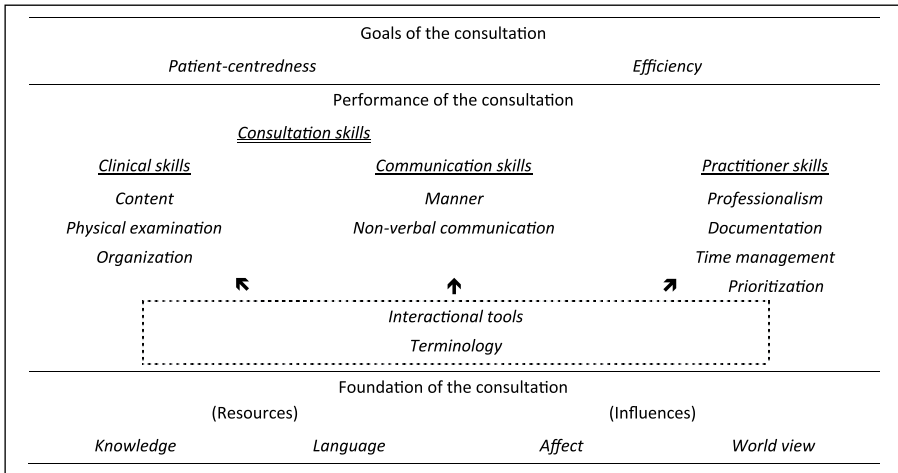


Figure 1. Model summarizing themes.

Organization – e.g., in terms of the process of information-gathering

she moved away from er the presenting complaint fairly quickly as well I thought, started to jump around a bit [MED-wk1-55]

Educators also valued the *Communication skills* of the trainees – how they related to the patient in terms of *Manner* and *Non-verbal communication*. The following extract from the data exemplifies evidence in the data for the interrelationship of themes and how superordinate categories could be defined based on the participants' comments: Participant 2 (P2) summarizes P1's comments on how the doctor *engaged* with the patient (*Manner*) and made *eye contact* (*Non-verbal communication*) as demonstrating *communication skills*.

P1 *did we say eye contact? ((very quietly)) Her eye contact was quite good.*

R1 *What absence of or –?*

P1 *No no she it was good yeah it was good. ((agreement)) She engaged she engaged really well.*

P2 *Basically her communication skills were good. [MED-wk1-128]*

Some educators' comments characterized *Communication skills* and *Clinical skills* as distinct and complementary.

I think both her communication skills were quite good and her clinical skills were quite good [MED-wk1-143]

You know like it was quite an effective communication in that [the nurse] got what she wanted out of it which was he- finding out basically her assessment of her [the patient's] obs[ervations] and things like that and and where her coughing was so I guess um yeah it was effective [NUR-wk1-318]

Interdependence between the two groups of skills was nevertheless recognized and mention was also made of *Consultation skills*, which appears to accommodate both groups. In the example below, *listening skills* were coded with the theme *Communication skills* and taking a *focused history* with the theme *Clinical skills*.

your general consultation skills have improved over the course of the general practice training. You demonstrated good listening skills, picking up on patient cues and resisted the temptation to interrupt early on. You then took a focused history of the presenting complaint. [MED-R12-2-5]

The set of *Practitioner skills* was noted mainly in the non-simulated contexts with the GP trainees, and also in the nursing workshops. Educators commented on trainees' *Professionalism* and the skills demonstrated regarding *Documentation*, *Time management*, and *Prioritization*. Their comments often related to behaviours that increased the professional's *Efficiency* (see below). The term *Practitioner skills* was not used by the educators themselves.

Professionalism

[concluding a discussion of how the patient should be addressed by the nurse] *I suppose that's what I'm always looking at, that when if I'm observing a student or staff member with a patient, I want to see that professionalism coming through* [NUR-wk4-307]

Documentation

Your note taking was perfectly adequate. The main factor that slows you down is your typing [MED-R06-3-50]

Time management – note also link to the theme Organization in “consultation structure”

There was a lot to cover with this patient and you were efficient and methodical in your consultation structure. [MED-R16-2-34]

Prioritization

You handled the seeing of different patients well, while at the same time dealing with an emergency patient with a lacerated face [MED-R26-1-18]

The behaviours in the three main thematic groups in the performance of the consultation are supported by a common set of *Interactional tools*. These were a variety of techniques and resources – realized predominantly through language – that accomplished the “business” of the consultation. In the following examples, *paraphrasing* and *checking understanding* are *Interactional tools*. (Educators did not use the term themselves.)

[the doctor] was – had a nice rapport with the patient and I thought the paraphrasing she did very well [MED-wk1-134]

you need ... to make sure that you clarify I think one of the things that um when you discharge patients home that um that whoever discharges the patient ... they give instructions but sometimes they don't clarify with the patient um exactly what does the patient understand [NUR-wk3-396]

A particular sub-set of *Interactional tools* concerned the theme *Terminology*. Participants commented on trainees' choice of words and their use of lay terms or jargon as enhancing or impeding interaction.

using the word 'illicit' for drugs which we tend to sort of say 'recreational' [MED-wk2-60]

then she used an acronym to him and he had no idea what that was and that upset him [NUR-wk3-239]

The goals of the consultation

When discussing which particular actions and behaviours were useful or not in the performance of the consultation, the educators also made comments to explain and support their views. In so doing, they highlighted the goals of the consultation, that is, what the performance seeks to achieve. The two principal goals were found to be *Patient-centredness* and *Efficiency*. Examples from the data relating to *Patient-centredness* follow. The examples for *Practitioner skills* above can be seen to link the issues of “what” and “why” for *Efficiency*; in addition, other aspects of performance, elements of *Interactional tools* and *Organization*, for example, were similarly linked to efficient practice (see below).

Patient-centredness

You have to pick up on the cue. You know, [the patient] commented at least once or twice about going home. So ... you had a great big open door there for an important conversation. [NUR-wk4-346]

you negotiated with the patients in a way that empowered them, e.g., when you were talking about what to do about the first woman's tummy pain, you said “we can decide together”. [MED-R10-3-7]

not tuning in at all to the patient's concerns about the possibility of cancer [MED-wk1-269]

Efficiency

[the nurse] was quite um prepared to explain who she was, what she was going to do and ah all that sort of stuff um but I think that she did drag it out a bit too long [NUR-wk3-113]

You are highly organised and manage your patients effectively and efficiently [MED-R02-1-42]

The foundation of the consultation

The third and final set of themes based on the educators' comments was concerned with two underlying resources that trainees draw on in their performance of the consultation and two influences that come to bear on their performance. The major resource for the educators was the trainees' *Knowledge*.

Your definition of pleurisy being any inspiratory chest pain is incorrect and I would like you to review your knowledge. [MED-R05-1-21]

In the analysis, it was noted how educators distinguished between *Knowledge* as a resource and *Content* as its application in the consultation – a theme in the group of *Clinical skills* above. In the first example below, the educator comments first on the trainee’s “knowledge base” (i.e., *Knowledge*) then on her ability “to apply it” (i.e., *Content*). In the second example, the educator mentions a lack of “clinical aptitude” (i.e., *Knowledge*) and an inability to make judgements in context (i.e., the real-time application of *Knowledge*, coded as *Content*).

You have a very good knowledge base esp[ecially] for mental health issues and osteoporotic risk fractures and you were able to apply this [MED-R24-1-19]

that sort of lack of um clinical aptitude really which she just didn't seem to be able to make those contextual judgements [NUR-wk3-112]

The second resource that the educators commented on was *Language*. Although it was much less a focus for them than *Knowledge* and related aspects of trainees’ performance, there was evidence that this theme mattered to the educators in the study. The terms in square brackets link the topic of the example to the existing assessment criteria used in the OET speaking sub-test. This is relevant to discussion below of the valued aspects of performance that inform the assessment criteria.

Some of the patients, mainly elderly, had difficulty with your accent [INTELLIGIBILITY; MED-R04-1-3]

the nurse seemed to speak English quite fluently [FLUENCY; NUR-wk3-114]

I think his questions were at least phrased clearly [RESOURCES OF GRAMMAR; MED-wk2-261]

You asked the patient “you still drink?”, which I thought was a bit abrupt [APPROPRIATENESS OF LANGUAGE; MED-R14-1-16]

In the same way that *Knowledge* has been presented as a resource for the health professional to draw on (as *Content*) in their performance, *Language* provides fundamental support in the performance of the consultation, in particular, for the theme of *Interactional tools*, that is, for the techniques through which many aspects of both *Communication skills* and *Clinical skills* are realized. In some cases, an educator’s feedback gives a trainee a specific realization to use – that is, particular words or a phrase to say – to perform most appropriately in a particular context (see the first example below). In other cases, the educator highlights the issue through description, for example, in the use of “medical jargon” and “signposting” in the second and third examples, respectively. The educators in the study consider such issues to be aspects of clinical communication, whereas the researcher, through the lens of applied linguistics, sees them as language related. This point is discussed below.

I still think she could have done more to embrace him in you know addressing his needs by using ‘I’ statements: ‘I can understand how you must feel’ and you know ‘we’ve probably been a bit neglectful we’re quite busy but, you know, I understand that now, what what can I do to help you now’ [NUR-wk1-164]

In this case, you explained several times about the cause. You used medical jargon and I think this led to [the patient's] lack of understanding. I would like you to practise using simpler everyday language – match it to the patient's background. [MED-R13-3-9]

she's not signposting enough ... she sort of said ou- almost out of the blue 'Have you had an STD [sexually transmitted disease] before?' [MED-wk2-169]

Finally, the first of the two influences on trainees in the foundation of the consultation was termed *Affect* and concerned issues relating to personality and emotional state, which were seen by the educators as having an impact on the performance of the consultation.

You need to be a bit more confident of yourself [MED-R03-2-12]

at one stage he was clearly getting nervous [MED-wk2-247]

The second influence on performance was the theme *World view*. This involved a trainee's view about the world in a broad sense and how people are and behave, perhaps drawn from their social and cultural background. Educators considered this theme with great circumspection. To prompt the extract below, participants in a medicine workshop had watched a video stimulus in which the trainee's history-taking process failed to pick up the simulated patient's cue that her boyfriend was away and link this to the patient's acquisition of a sexually transmitted infection. Two educators discussed how to attribute the reason for the trainee's poor performance, that is, her failure both in clinical and communication terms. They considered the trainee's possible naivety along with cultural beliefs about monogamy and taboo.

P3 *would she would she even be naïve enough to – or just culturally not aware that –*

P4 *well she might accept the patient's –*

P3 *that this person's just monogamous and that people are just monogamous? (xxx)*

R1 *or she might think it's offensive ((laugh))*

P3 *offensive yeah [MED-wk2-205]*

Themes were generally consistent between the two professions presented here and in the physiotherapy data. Particular inter-professional differences were noted occasionally: for example, the theme *Organization*, found in the medicine data, was not found in the nursing data. This might reflect the different professional roles and perhaps also the particular contexts in which data were gathered and the stimuli used.

Discussion

In response to the first research question about what matters to health professionals, the thematic analysis of the commentary established a wide range of themes that mattered to educators as they gave feedback on the performance of novices in their professions. Clinical and communication skills are construed as distinct, but also as working together to achieve the goals of the consultation, which are patient-centredness and efficiency.

Value is clearly placed on health professionals' underlying knowledge resource; their language resource is also valued. The impact of other influences on performance – for example, personality and cultural background – are acknowledged but possibly viewed as beyond the professional remit of educators.

The second research question asks whether those aspects of performance found to matter are amenable to inclusion in assessment criteria for a language test. In a direct performance test, the assessment must be concerned with what is observable. The foundation themes (resources and influences) support the performance of the consultation but are not themselves directly observable; they comprise the underlying competence of the test taker. Similarly, the goals of the consultation are important in guiding the choices a test taker makes in its performance, but they cannot be measured directly, only through the performance itself. Therefore, the valued aspects of the *performance* of the consultation are what must be considered for inclusion in assessment criteria, taking into account their use to achieve its goals.

Also constraining inclusion for assessment is the issue of which aspects of performance can be considered to relate to language. In the Australian context, as already noted, the assessment of health professionals' language proficiency is required by law to be separate from that of their professional competence. Themes directly related to professional issues must therefore be excluded (i.e., the groups of *Clinical skills* and *Practitioner skills*). OET speaking assessors have backgrounds in language teaching and testing and receive training to assess the roleplay performances, but they are not qualified healthcare professionals who are competent to make judgements on, for example, a test taker's skills at physical examination or diagnostic reasoning.

The presence in the model of the theme *Language* indicates that this resource matters to the educators. Examples in the Findings section above show how educators commented on aspects of language performance that are covered by the existing analytic assessment criteria for the OET speaking sub-test – FLUENCY, INTELLIGIBILITY, RESOURCES OF GRAMMAR AND EXPRESSION, and APPROPRIATENESS OF LANGUAGE. The sub-set of *Terminology in Interactional tools* links in particular to APPROPRIATENESS OF LANGUAGE and word choice for the context of the consultation. This evidence therefore supports the retention of these criteria in the assessment scheme. To create further criteria based on the analysis, aspects of performance coded with themes under *Communication skills* and *Interactional tools* can be considered from a language perspective, that is, in addition to a clinical perspective. As the data show, behaviours coded as *Interactional tools* facilitate the consultation in terms of communication as well as clinically. The OET assessors have the capacity to assess the quality and appropriateness of test takers' use of *Interactional tools*, such as paraphrasing or signposting, *in language terms*, while the quality and appropriateness of their *clinical application* – potentially observable by a healthcare educator – are left unassessed in the context of a language test. In this way, the view of language in the assessment criteria is expanded to include more of what matters to the healthcare educators. While test takers' paraphrasing, signposting and use of other *Interactional tools* might always have been considered by experienced OET assessors under the holistic OVERALL COMMUNICATIVE EFFECTIVENESS criterion (or even another existing criterion), the transparency afforded by developing new criteria that define more

clearly these valued aspects of performance is a positive step. These criteria can replace the existing holistic criterion.

The goals of the consultation were found to be *Patient-centredness* and *Efficiency*. The data show that both are achieved through the effective performance of the clinical task in the consultation and that language is the predominant means of realizing this performance. The goals may therefore inform the creation of new assessment criteria for a language test. From concrete examples of performance – those coded with the themes in the study found amenable to inclusion in a language test – 24 prototypical indicators of performance were derived (as suggested by Douglas, 2001a). These were put into four groups: indicators of professional manner and of patient awareness, and indicators for information-gathering and for information-giving. The resulting checklist of indicators, which was subsequently used as a training tool for OET assessors, was further abstracted to generate two assessment criteria (see O'Hagan, Pill, & Zhang, this issue). The criterion CLINICIAN ENGAGEMENT considers issues of professional manner and patient awareness and perhaps represents aspects of performance that achieve the goal of *Patient-centredness* as recognized by this study; the criterion MANAGEMENT OF INTERACTION involves mainly issues of information-gathering and information-giving, perhaps representing more the means to achieve the goal of *Efficiency*.

Non-verbal communication was a theme highly valued by the educators in the study. At current administrations of the OET at test centres around the world, the speaking sub-test is recorded for subsequent assessment and the recordings made are audio-only. This practical constraint therefore prevents any direct assessment of issues of non-verbal communication beyond paralinguistic features of speech accessible via an audio recording. The study did not consider further how non-verbal communication might be assessed – for example, with a further criterion of its own or through building it into the proposed new criteria, focusing on how it promotes CLINICIAN ENGAGEMENT or MANAGEMENT OF INTERACTION – although, in future forms of the OET or in another test free of the limitations described, assessment of this aspect of performance is desirable in order to represent more completely the values of the practitioners.

Conclusion

The research introduced in this paper is a novel attempt to derive contextually relevant criteria for use in the assessment scheme of a specific-purpose language test, using an empirical method drawing on the values and views expressed by practitioners about the target language use situation. It sets an example of abstracting from highly specific data to create more general criteria and therefore may be considered an initial, practical step in bridging the divide noted in the literature between context-bound and more generalizable criteria. It also provides a methodological framework for further studies in different specific-purpose language testing contexts.

The revision of the scope of the assessment criteria has clear implications for the construct underlying the test. The two criteria proposed in this study add a contextually relevant, interactional perspective to the existing criteria, which focus more on characteristics of the individual test taker. Test takers will be assessed on their ability to use the language resources available to them to respond to the concerns of their patients and to

manage aspects of the consultation efficiently. The argument for including these new aspects for assessment is that they are valued by practitioners and are performed through language. In specific-purpose language tests, the inseparability of the task and the language with which to do it has long been recognized (Douglas, 2001b).

This study is limited in the sources from which data were collected and it can be foreseen that data collected in other healthcare settings or prompted by different stimuli might have produced a different set of themes. Nevertheless, the findings presented here were recognized by members of the reference group for the overall project, who were experienced practitioners in the professions studied. In the data collection process, educators might have stressed aspects of language and communication because, as noted in Elder and McNamara (this issue), they knew the disciplinary background of the researchers. However, there is no apparent difference between the themes represented in the GP trainee report dataset and the other sources of data, and the GP reports were not prepared with the research study in mind. Finally, while the study considered data for three health professions (with two of these being the focus of this paper), the OET currently serves 12 health professions, and the validity of the proposed criteria for each profession must be ascertained before they are introduced at routine test administrations.

Acknowledgements

The author wishes to thank his companion data collectors and coders, Sally O'Hagan and Diana van Die, for their insights and collaboration during the coding phase of the project. He recognizes the generous support of the members of the ARC project team, in particular his three doctoral supervisors: Cathie Elder, Tim McNamara and Robyn Woodward-Kron. He acknowledges the vital contribution of the educators and trainees who participated in the study.

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 author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Australian Research Council [Linkage grant number LP0991153].

Notes

1. The literature review in this paper considers the findings of studies undertaken on indigenous criteria while Elder and McNamara (this issue) focus on methodological issues raised in these and similar studies.
2. The physiotherapy data are discussed separately (see Elder & McNamara, this issue). The findings for medicine have been presented in some detail in the author's doctoral thesis (Pill, 2013) and, in a preliminary form, in Elder et al. (2012). Profession-specific findings have been published for physiotherapy (Woodward-Kron et al., 2012) and nursing (O'Hagan et al., 2014).
3. One nursing "workshop" involved an interview with a single participant. The procedures followed were nevertheless the same.

References

- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Brown, A. (1993). LSP testing: The role of linguistic and real-world criteria. *Melbourne Papers in Language Testing*, 2(2), 35–54.
- Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, 20(4), 369–383.
- Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, UK: Cambridge University Press.
- Douglas, D. (2001a). Language for Specific Purposes assessment criteria: Where do they come from? *Language Testing*, 18(2), 171–185.
- Douglas, D. (2001b). Three problems in testing language for specific purposes: Authenticity, specificity and inseparability. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. F. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 45–52). Cambridge, UK: Cambridge University Press.
- 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, UK: Cambridge University Press.
- Duff, P. A. (2008). Language socialization, higher education, and work. In P. A. Duff & N. H. Hornberger (Eds.), *Encyclopedia of language and education*, Vol. 8: *Language socialization* (2nd ed., pp. 257–270). New York: Springer.
- Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing*, 10(3), 235–254.
- Elder, C. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing*, 18(2), 149–170.
- 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.
- Erdősy, M. U. (2005). Responding to native and non-native writers of English: A history professor's indigenous criteria for grading and feedback in an undergraduate Sinology course. Unpublished doctoral dissertation. University of Toronto.
- Erdősy, M. U. (2009). Chasing Proteus: Identifying indigenous assessment criteria in academic settings. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment* (pp. 111–133). Frankfurt am Main, Germany: Peter Lang.
- Forman, J., & Damschroder, L. (2008). Qualitative content analysis. In L. Jacoby & L. A. Siminoff (Eds.), *Empirical methods for bioethics: A primer* (pp. 39–62). Oxford, UK: Elsevier.
- Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, 28(1), 5–29.
- Galaczi, E. D. (2014). Content analysis. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1325–1339). Hoboken, NJ: Wiley.
- Jacoby, S. W. (1998). Science as performance: Socializing scientific discourse through the conference talk rehearsal. Unpublished doctoral dissertation. University of California, Los Angeles.
- Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes*, 18(3), 213–241.
- Lumley, T. (1995). The judgements of language-trained raters and doctors in a test of English for health professionals. *Melbourne Papers in Language Testing*, 4(1), 74–98.
- McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

- O'Hagan, S., Manias, E., Elder, C., Pill, J., Woodward-Kron, R., McNamara, T., Webb, G., & McColl, G. (2014). What counts as effective communication in nursing? Evidence from nurse educators' and clinicians' feedback on nurse interactions with simulated patients. *Journal of Advanced Nursing*, 70(6), 1344–1354.
- Pill, T. J. H. (2013). *What doctors value in consultations and the implications for specific-purpose language testing* (Unpublished doctoral dissertation). University of Melbourne.
- Roberts, C. (2005). English in the workplace. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 117–135). Mahwah, NJ: Lawrence Erlbaum.
- Wette, R. (2011). English proficiency tests and communication skills training for overseas-qualified health professionals in Australia and New Zealand. *Language Assessment Quarterly*, 8(2), 200–210.
- Woodward-Kron, R., van Die, D., Webb, G., Pill, J., Elder, C., McNamara, T., Manias, E., & McColl, G. (2012). Perspectives from physiotherapy supervisors on student–patient communication. *International Journal of Medical Education*, 3, 166–174.

Transcription conventions

- (xxx) Parentheses indicate text that was unclear; each “x” represents a syllable.
- ((agreement)) Double parentheses enclose summaries of a general response in a workshop setting.
- she [the nurse] Square brackets enclose information added to aid understanding of the extract or to replace information that would identify a participant.
- ... Ellipsis indicates omitted text.
- he said ‘illicit’ Single quotation marks show speech reported by the speaker.
- but she didn’t Non-italic text indicates a syllable/word stressed in speech.