

How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals

Language Testing
2016, Vol. 33(2) 217–234
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0265532215607402
ltj.sagepub.com



John Pill

American University of Beirut, Lebanon

Tim McNamara

University of Melbourne, Australia

Abstract

This paper considers how to establish the minimum required level of professionally relevant oral communication ability in the medium of English for health practitioners with English as an additional language (EAL) to gain admission to practice in jurisdictions where English is the dominant language. A theoretical concern is the construct of clinical communicative competence and its separability (or not) from other aspects of professional competence, while a methodological question examines the technical difficulty of determining a defensible minimum standard. The paper reports on a standard-setting study to set a minimum standard of professionally relevant oral competence for three health professions – medicine, nursing, and physiotherapy – as measured by the speaking sub-test of the Occupational English Test, a profession-specific test of clinically related communicative competence. While clinical educators determined the standard, it is to be implemented by raters trained as teachers of EAL; therefore, the commensurability of the views of each group is a central issue. This also relates to where the limits of authenticity lie in the context of testing language for specific purposes: to represent the views of domain experts, a sufficient alignment of their views with scores given by the raters of test performances is vital. The paper considers the construct of clinical communicative competence and describes the standard-setting study, which used the analytical judgement method. The method proved successful in capturing sufficiently consistent judgements to define defensible standards. Findings also indicate that raters can act as proxies for occupational experts, although it remains unclear whether the views of performances held by these two groups are directly comparable. The new minimum standards represented by the cut scores were found to be somewhat harsher than those in current use, particularly in medicine.

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

Keywords

Analytical judgement method, healthcare communication, language proficiency, LSP testing, Occupational English Test, standard setting

This paper deals with the difficult question of establishing the minimum required level of professionally relevant oral communication ability in the medium of English for health practitioners with English as an additional language (EAL) to gain admission to professional practice in medicine and allied health fields in jurisdictions where English is the dominant language. A lot hangs on this question. On the one hand, there is the right of readmission of migrant professionals to the clinical workplace, a workplace within which they may have been operating successfully for years, in the medium of other languages; in some cases, the issue is a right of permanent admission to a workplace in which they are already working, under provisional registration, in which English is the main medium of communication. On the other hand, there is the safety and health of patients, which may be compromised by professionals who are not fully capable of understanding patients' needs and health status, of negotiating adherence to treatment regimes, or of communicating successfully with other health professional colleagues, all through the medium of spoken English. The complexity of resolving the issue is a result of two factors: the construct of clinical communicative competence and its separability (or not) from other aspects of professional competence, a theoretical question; and the technical difficulty of determining a defensible minimum standard, a methodological question.

In this paper, we report on a standard-setting study which attempts to address these two questions. The study implies a position on the construct of clinical communicative competence as much as it is a methodological study of standard setting. It involves the establishment of a minimum standard of professionally relevant oral competence in English as an additional language for three health professions – medicine, nursing, and physiotherapy – as measured by the speaking sub-test of the Occupational English Test (OET), a profession-specific test of clinically related communicative competence. While clinical educators and professionals were involved in determining the standard, its implementation is in the hands of raters trained as teachers of EAL, so the issue of the commensurability of the views of each group is central to the study. This also relates to where the limits of authenticity might lie in the context of testing language for specific purposes: if a test is to represent the views of domain experts authentically, an adequate alignment of their views with the scores given by the raters of test performances is vital.

The paper begins with a discussion of the construct of clinical communicative competence, and then reports on a standard-setting study involving clinical educators in establishing minimum standards of oral communicative competence for immigrant health professionals. Two kinds of data were available from the research study; quantitative evidence of judgements are presented in the current paper, while qualitative comments from the health professionals involved are considered separately (Manias & McNamara, this issue).

Defining a construct of clinical communicative competence

Establishing a relevant construct of clinical communicative competence as the basis of a language test can be complicated by policies affecting professional licensure. For example, in Australia, national regulatory bodies apply a *language skills* standard to health professionals who have English as an additional language and qualifications from outside their jurisdiction. This acknowledges the importance of health professionals having the language and general communication skills necessary to practise effectively in English in the Australian context. Achieving certain grades or bands on a recognized language test – principally the specific-purpose Occupational English Test (OET) or the general-purpose International English Language Testing System (IELTS) – is the usual way that applicants for professional registration meet the language skills standard.

Standards of *professional knowledge and skills* also are set, and these are measured separately from language skills, as required under Australian law. As an example, doctors who follow the “standard pathway” of the Australian Medical Council (AMC) to obtain registration in Australia are required to pass a multiple-choice question test to demonstrate their clinical knowledge and then a clinical examination comprising a series of interactions with patients and simulated patients to demonstrate application of clinical skills and the associated clinical communication skills (Australian Medical Council, 2015).

There are thus two questions to consider about the use of language tests in the context of health professional registration. The first is whether the language tests used are in fact appropriate measures of the particular language skills test takers require in the target language use situation for which they function as a gate-keeping mechanism. With regard to the subject of this paper, the speaking sub-test of the Occupational English Test, this question has been discussed elsewhere (Lumley, 1995; O’Hagan, Pill, & Zhang, this issue; Woodward-Kron & Elder, this issue), and the assumption made in this standard-setting exercise is that the test, particularly with the proposed expansion of the set of assessment criteria (see Pill, this issue), is a sufficiently valid assessment tool for its specific purpose. The paper on the qualitative data from our standard-setting study also addresses this first question (Manias & McNamara, this issue). The subsequent question is how to determine what level of performance on the specific-purpose language test, scored by raters with training as teachers of EAL, and as expressed through the test score, is sufficient in the eyes of health professionals to allow a practitioner to participate safely and effectively in the Australian workplace, at least as far as his or her language and general communication skills are concerned. In this standard-setting study, the views of health professionals were sought as to the overall adequacy or otherwise of a series of performances on the oral component of a clinically relevant communication skills test, and these views, expressed as judgement scores, were then compared with the scores given to the same performances by language-trained raters, in order to evaluate how commensurable they were.

These questions together raise a broader theoretical issue in specific-purpose language testing about whether language-trained raters scoring language test performances can be an effective proxy representing the views of other groups, for example, health professionals, who inevitably bring a different perspective to the situation, and who may

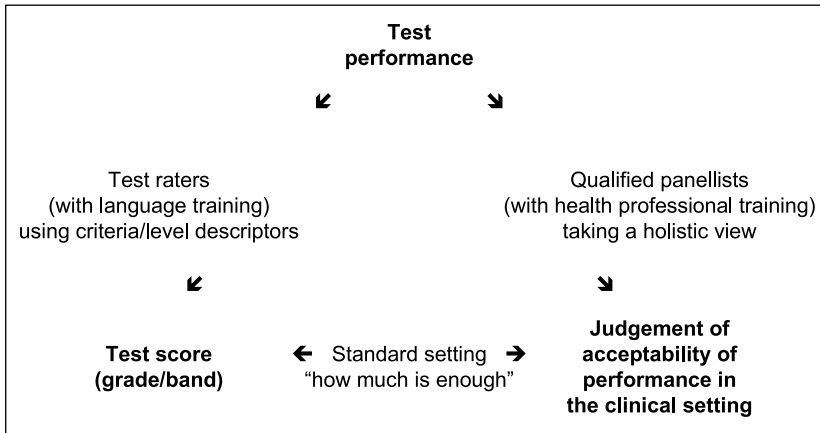


Figure 1. Interpreting test scores through standard setting.

not be able to separate out language and communication skills from other knowledge and skills that form part of a test performance. Their general opinion of a test performance may be based on implicit criteria that are very different from those used explicitly by the language-trained raters. The overarching question is consequently whose views – or what weightings of different views – are more useful for the context in which the test results are interpreted and take effect. Furthermore, it also may be the case that practitioners in different health professions do not share the same view on performance standards. While a relatively homogeneous perspective was found on the various aspects of performance that matter for health practitioners in three professions (Pill, this issue), it does not necessarily follow that the standards of readiness to practise will align in a similar way. In the particular case of the OET, current practice applies the same cut scores to all test takers, regardless of profession. It is therefore important to look for variation in the accepted standards among the health professions under investigation here.

Figure 1 shows the relationships between test performances, raters' scores and panellists' assessments, and the basis for the interpretation in each case. Standard setting can be seen as a means of "triangulating" the data and their interpretations.

Our study draws on the opinions of health professionals and educators in three health professions: medicine, nursing, and physiotherapy. Because the specific-purpose language test simulates professional interaction with a patient (see Elder, this issue, for description of the speaking test format and task), it was anticipated that the health professional participants (panellists) in the standard-setting exercise would be able to recognize the quality of the test takers' performances and match them easily to standards of workplace readiness. As Kenyon and Römhild (2014) note, standard setting originates in certification testing, for example, for teaching licensure. In such a context, and in the methodology of standard setting generally, there is an assumption that the views of the panellists setting the standard correspond with the criteria used to assess the performance; that is, they are drawn from the same underlying construct. Also, this assumption is made to some degree in the current study but, as already noted, in this case we can use

the standard-setting methodology to explore how different the perspectives elicited in our study are, that is, the perspectives of language-trained raters on the one hand and of domain experts – health professionals – on the other. We can investigate whether the two groups have a sufficiently similar understanding of at least one part of the underlying construct (i.e., the part relating to language and communication skills) so that the raters' scores using the test criteria correlate with the domain experts' assessments of the performances overall, thereby assuring an authenticity of assessment in this regard.

Standard setting

Kenyon and Römheld (2014) describe standard setting as “a process by which qualified panellists, following carefully developed and documented procedures that mitigate against arbitrariness, assign interpretative meaning to performances on tests” (p. 944). For instance, panellists provide judgements on the quality of a test performance regarding whether it indicates the test taker's readiness to participate in the target language use situation. Kenyon and Römheld (2014) further note that “Although the process of standard setting relies on psychometric and statistical tools, the process itself is a socially moderated one” (p. 945).

There are various methods used for standard setting. Some, such as the modified Angoff methods (Brandon, 2004) and the bookmark method (Mitzel, Lewis, Patz, & Green, 2001), involve independent panellists reviewing selected-response test items or test criteria and descriptors to determine the likely performance of an imagined “minimally competent” test taker. In contrast, holistic methods (Cizek & Bunch, 2007) present panellists with products of assessment (a “body of work”) by a test taker and require them to make a holistic judgement about the overall competence demonstrated. These methods are therefore suitable for use with spoken or written performances. They are easy to explain to the panellists, as what is required is a decision that is intuitive for them, that is, similar to those made in other assessment situations based on an actual performance with which they are likely to be familiar. The methods often focus on performances that fall around the borderline between, say, pass and fail, because panellists' views of performances in this area are used to establish the cut score. The panellists' decisions are independent of scores assigned previously to the performances using the established rating scheme for the particular context; different standard-setting methods then involve a variety of mathematical and statistical approaches to determine the final cut score.

The analytical judgement method

The analytical (or analytic) judgement method, the one used in this study, was described by Plake and Hambleton (2001). They initially suggested that panellists sort performances into 12 categories: four categories named (for example) “novice,” “apprentice,” “proficient,” and “advanced,” each having three sub-categories, “low,” “mid,” and “high.” The highest-level category is therefore “high advanced,” and the lowest is “low novice.” The focus for the purposes of setting cut scores is on the borderline categories, for example, the adjacent pair of “high apprentice” and “low proficient.” All the

performances that panellists place into these two categories are used to establish a single cut score between “apprentice” and “proficient” by calculating the mean of the scores they were given using the established rating scheme. Cizek and Bunch (2007) note that, although various methods of calculating the cut score are proposed, “a simple averaging approach appeared to work as well as the others (p. 122).” It can be seen that this method may be used to create cut scores between more than one level of performance, which is also pertinent to the context of the OET speaking sub-test under investigation, where results are reported using a series of grades, currently from A (highest) to E (lowest).

Plake and Hambleton (2001) also suggest an adapted version of the analytical judgement method where the pairs of borderline categories are merged into one and the two highest and two lowest categories are also merged, creating a set of seven rather than 12 categories. This reduced set of possible categories is thought easier for panellists to manage and, in practice, was found not to affect the placement of cut scores greatly (Cizek & Bunch, 2007, p. 122). The seven-category model used in the current study is presented in the Methods section below.

Standard setting for language proficiency in health professional contexts

Published examples of standard setting carried out in health professional contexts are those initiated by the United States National Council of State Boards of Nursing (NCSBN) to set passing standards for nurses trained outside North America on the revised (paper-based) Test of English as a Foreign Language (TOEFL; O’Neill, Tannenbaum, & Tiffen, 2005), TOEFL internet-based test (TOEFL iBT; Wendt, Woo, & Kenny, 2009), International English Language Testing System (IELTS; O’Neill, Buckendahl, Plake, & Taylor, 2007), the Pearson Test of English Academic (PTE Academic; Woo, Dickison, & de Jong, 2010) and the Michigan English Language Assessment Battery (MELAB; Qian, Woo, & Banerjee, 2014). The panellists in these studies included nurses with English as an additional language, nurse educators, nurse regulators, and consumers. Berry, O’Sullivan and Rugea (2013) recommended new standards for IELTS for international medical graduate doctors in a report to the General Medical Council (GMC) in the UK based on a standard-setting exercise involving separate groups of doctors, nurses, allied health professionals, patients, and members of the public, and responsible officers and medical directors; however, the recommendations in the report were taken up only to some extent in revisions to GMC language standards made subsequently. It also should be noted that none of these English tests seeks directly to reflect the healthcare workplace in the content, task types, or assessment criteria used. There appears to be no similar standard-setting research for other professions.

Domain experts scoring the Occupational English Test

Lumley’s (1995) study compared the rating of OET speaking sub-test performances by language-trained raters and by medical practitioners. This was not a standard-setting exercise, however, and the medical practitioners were specifically asked to use the same scale as the raters. The purpose of the study was to investigate whether raters were assessing test takers too leniently, following complaints that migrating doctors

were passing the test but subsequently not meeting the standards expected of them. The study found that, overall, the doctor participants rated test performances slightly more leniently than the OET raters; in terms of mean scores, there was no significant difference between the two groups. The doctors gave a single score only based on the Overall Communicative Effectiveness criterion of the assessment scheme, although they also were aware of the scope of the other linguistic analytic criteria used on the test. Their scores, given using a single six-point scale, were compared only with the scores given by OET raters on the same criterion, although the OET raters had scored the performances using the full assessment scheme of six criteria. The findings therefore show that doctors can apply a criterion of communicative effectiveness in the same way that OET raters apply it. However, this does not directly address whether the doctors are satisfied with the performance of the test takers in terms of their demonstrating readiness for the workplace in a more general sense, that is, aside from a language-focused perspective, which might be less pertinent to health professionals (see Elder & McNamara, this issue). The question of what performances the doctors deemed satisfactory and unsatisfactory was not considered directly, although the interpretation of the scale they used explicitly stated that scores of four and higher indicated a “pass.” (The criteria, level descriptors and cut scores used at present on the OET are somewhat different from those described in Lumley’s study.) While Lumley’s study provides some information about doctors’ views of medical test takers’ performances, there is no evidence that similar findings would emerge if parallel studies were carried out for nursing and physiotherapy.

The cut scores in current use on the OET speaking test are well established, although their provenance is unclear. Five grades are used to report test results: A–E (highest to lowest). The cut scores for the speaking and writing sub-tests have remained the same at each test administration for a number of years and are expressed as fair scores in logits. Fair scores are the output of routine Rasch analysis using Facets software (Linacre, 2014), drawing on sets of raw scores given by two raters independently for each test taker. Raters choose from a set of level descriptors – scored 1 (lowest) to 6 – for each criterion. The current five criteria for the speaking sub-test are Overall Communicative Effectiveness, Intelligibility, Fluency, Appropriateness of Language, and Resources of Grammar and Expression. In the proposed revision to the assessment scheme, the criterion Overall Communicative Effectiveness is replaced by two new criteria, Clinician Engagement and Management of Interaction, which were scored 1 to 4 (see O’Hagan, Pill, & Zhang, this issue). The probabilistic many-facet Rasch measurement (MFRM) model includes rater performance as a facet in the data analysis to take account of the relative severity of each rater in the group marking the set of test-taker performances. The fair score derived for each test taker is expressed as a number between 1.00 and 6.00 for the existing set of criteria and between 1.00 and 5.33 for the proposed set of criteria. This reduction in score range is due to the use of shorter scale lengths for the two new criteria. Note how this difference in the score ranges means that fair scores derived from one set of criteria cannot be compared directly with those from the other. Also, it means that the existing cut scores cannot be applied to fair scores based on the proposed set of criteria. Establishing new, empirically based cut scores for the proposed set of criteria is a practical goal of this study.

Table 1. Summary of numbers of panellists, workshops and samples graded.

	Panellists	Workshops	Samples graded
Medicine	13	3	26
Nursing	18	2	25
Physiotherapy	8	1	24

Research questions

The standard-setting exercise addresses these questions:

- (1) Can health professionals (for three health professions) consistently assess the performance of OET test takers on speaking test tasks in terms of the language and communication skills demonstrated?
- (2) To what extent are their holistic ratings of test-taker performance similar to those made by language-trained raters using OET speaking assessment criteria specifically expanded with the intention of including more of what health professionals value?
- (3) In what ways do the cut scores established in the standard-setting exercise using data from the health professionals differ from those in current use on the OET speaking sub-test?
- (4) Is there any variation in the cut scores established for the three different professions?

Methods¹

The procedural rigour in a standard-setting exercise itself constitutes an element of the validation of the exercise (Kenyon & Römhild, 2014). To minimize arbitrariness and randomness of results, a defensible and principled approach is required in the methods used.

Participants

Panellists were educators and practitioners in each of the three professions being investigated who had experience working with trainees, including those with English as an additional language. Several had been involved in data collection for an earlier phase of the same project (see Pill, this issue); their further involvement was deliberately sought. Panellists consented to their data being used and were paid for their involvement in the standard-setting workshops. Table 1 summarizes the number of panellists for each profession, the number of workshops held, and the number of different samples graded. There is no clear standard for how many panellists and samples are required to achieve satisfactory outcomes in this procedure (Plake & Hambleton, 2001, pp. 309–310); the current study provides further information to consider in this regard (See Discussion below).

Procedure

In each workshop, panellists were introduced briefly to the context of the OET and the format of the speaking sub-test. The grading procedure was explained and the scoresheet introduced (see Instrument section below). Panellists were asked to take a holistic view of the test performances but to exclude from their decisions the test taker's clinical knowledge and skills apparent in the recording, as language-trained raters would not be able to evaluate these domains. Selected samples of test-taker performances were presented to the panellists as audio recordings, reflecting how OET raters routinely assess performances. The workshop facilitators acknowledged that potentially useful visual information relating to non-verbal communication was not available from the audio recordings. Samples of performance used in the workshop – and in the post-workshop assessment pack – were chosen carefully to represent a range of levels of performance based on the fair scores resulting from analysis of the OET raters' application of the proposed six criteria (see O'Hagan, Pill, & Zhang, this issue). This set of criteria was chosen for this purpose because, congruent with the premise of the whole project, it was anticipated to reflect to the greatest extent possible the scope of panellists' views. Panellists were not told the OET raters' scores, however.

In the workshop, panellists revealed the grade they had given for each sample presented and discussed the reasons for their decisions. As Plake and Hambleton (2001) explain, "the goal of the discussion is to allow panellists who rated a [performance] differently to provide their rationale, with the intent of sharing insights that might have been overlooked or missed by the other panellists" (p. 290). Discussion also allowed the facilitators to focus panellists' attention on the task set if their comments indicated that they appeared to be deviating from it, for example, showing influence relating to a test taker's apparent lack of professional knowledge. Panellists were invited to amend their initial grade following the discussion if they so wished, but their final grade was not shared with other panellists – the analytical judgement method does not seek a group consensus. Between four and six samples were discussed in this way at each workshop. (For analysis of discussion from standard-setting workshops, see Manias and McNamara, this issue.) To conclude the workshop, the facilitators introduced a post-workshop assessment pack with a further 20 samples for grading. These were presented to panellists on audio CDs containing different sequences of recordings to avoid any order effect. An instruction sheet was provided. Written feedback was collected from the panellists at the end of the workshop and on their completion of the post-workshop assessment task to gain understanding of the panellists' perceptions of the validity of the process.²

Instrument

Data collected for analysis were each panellist's grade for each sample of performance. Following an adapted analytical judgement method, panellists could choose one of seven levels of performance. Figure 2 presents an example for nursing of the scoresheet format used. Four anchor categories were chosen: "unsatisfactory", "not yet competent", "competent" and "strong". The use of "not yet competent" and "competent" was intended to signal clearly a "fail/pass" threshold. There was no explicit discussion to define a "minimally

Mark **only one** of the seven boxes on the right with a cross (X).

You will hear a nurse who has a qualification from outside Australia. Consider this performance for the purpose of the nurse's participation in entry-level / supervised clinical practice involving interaction with patients, co-workers and supervisors, and as if you were his/her supervisor.

I rate the communication skills in this performance as ...

STRONG

7

⇄ between STRONG and COMPETENT

6

COMPETENT

5

⇄ between COMPETENT and NOT YET COMPETENT

4

NOT YET COMPETENT

3

⇄ between NOT YET COMPETENT and UNSATISFACTORY

2

UNSATISFACTORY

1

Figure 2. Scoresheet for the nursing standard-setting exercise.

competent performance," as it was anticipated that panellists would have different views of this and these were to be captured in the grading process itself. Panellists were encouraged to use the "between" categories as freely as the others. These categories are important because they provide the data required for cut scores to be established. Each scoresheet presented the same information about the context in which the grading decisions were to be made (see Figure 2); this context was also explained in the set-up phase of each workshop. The contextualization of the decision – in this case, relating to a health professional trained outside Australia and his or her demonstration of readiness to participate in supervised clinical practice – is essential to the cut scores being set appropriately for the intended use and interpretation of the test results. The instruction that the panellist should make the decision as if he or she were the test taker's supervisor was intended to create a sense of personal involvement and responsibility for the health professional in question.

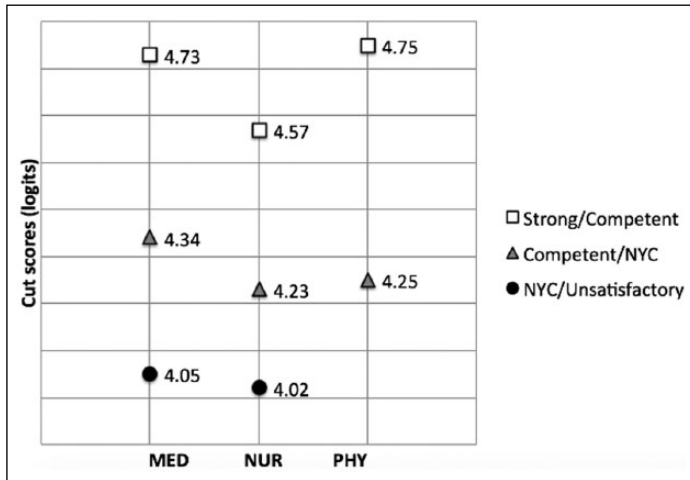


Figure 3. Comparing cut scores across the three professions. NYC = Not yet competent; MED = medicine, NUR = nursing, PHY = physiotherapy.

The process of analysis of the data collected is explained concurrently with the study’s findings in the following section.

Analysis and results

New cut scores

The data were collected into spreadsheets for each profession. The analytical judgement method was applied as described above. Thus, the fair score was determined by a prior Facets analysis of the OET raters’ scores, using the expanded set of six criteria; this score was added up for every performance in each “between” category, weighted by the number of times it occurred in the category. Therefore, if two panellists put the same performance in a “between” category it was counted twice in the sum, and so on. The mean fair score for all the performances in the “between” categories is taken as the cut score between the categories above and below. Cut scores were calculated in this way between each of the four anchor categories and reported as fair scores on a scale from 1.00 to 5.33 for each profession. See the Appendix for an example of the calculations for the data for medicine. The cut scores established are shown in Figure 3: “Strong (A)/Competent (B),” “Competent (B)/Not yet competent (C),” “Not yet competent (C)/Unsatisfactory (D).” In current practice, the lowest grade used for OET results is grade E; however, this is rarely awarded in routine test administrations and, for the purposes of this study, grades D and E were conflated.

The cut scores differed slightly for each profession. The cut scores for nursing were lowest, which raises the possibility that different cut scores should be used for each profession in operational administrations. Note that it was not possible to determine the lowest cut-score level (i.e., between “Not yet competent” and “Unsatisfactory”) for physiotherapy due to the paucity of test takers placed into these levels; more data would be needed to resolve this issue (see Discussion below).

Table 2. Summary of MFRM analysis findings for three professions.

	Medicine	Nursing	Physiotherapy
Number of panellists	13	18	8
Panellist leniency/severity: Measure outside range from -1.0 to 1.0	1 "lenient"	1 "severe"	–
Panellist inconsistency: Infit MnSq >1.5	1	3	1
Number of performances	26	25	24
Reliability of measures	0.92	0.94	0.89
Misfitting performances: Infit ZStd >2.0	1 (no. 12)	–	1 (no. 21)

Consistency of panellists' grading

In order to investigate the consistency and coherence of grading by the health professional panellists using the seven-point scale, a MFRM analysis was carried out using Facets software for the data collected for each profession. Selected results are shown in Table 2.

The analyses indicate that most panellists were rating consistently, while exhibiting somewhat different standards of harshness and leniency in terms of allocation to the given categories, an entirely expected result, and one in line with previous studies of rating behaviour (e.g., Eckes, 2011). One medicine performance was problematic in terms of its categorization for the medical practitioner panellists (i.e., "misfitting" in the analysis); the Appendix shows how this performance (no. 12) was placed by panellists into five of the seven available categories. Similarly, one physiotherapy performance was problematic for the physiotherapy panellists.³ The reliability of the test-taker measures derived from the ratings was high.

Comparing health professional and language rater judgements

The health professionals' assessment of the performances using the standard-setting instrument provided judgement scores that, when subjected to the same MFRM analyses, yielded an overall measure (fair score) for each test-taker performance. These health professional fair scores could then be compared – in terms of how they ranked performances – with fair scores for the same performances produced using the scores of the regular OET raters applying the expanded set of criteria. Although the instruments used by the two sets of judges, namely, health professionals and OET raters, were different, this became a test of the extent to which the construct – clinical communicative competence – appeared to be shared between the two sets of rankings. Performances were first ranked according to the OET raters' fair scores and coded with the grade category (A, B, C or D) they fell into when the new cut scores were applied. They were then re-sorted according to the health professionals' fair scores. Performances were counted as clearly differently ranked if, when re-sorted, they were next to performances that were not in the same or an adjacent grade category. For example, a performance was clearly differently ranked if it was categorized as grade A by the OET raters but was then placed, according

Table 3. Classification using old and new criteria.

Classification		Medicine		Nursing		Physiotherapy	
Grade	Category	Existing	New	Existing	New	Existing	New
<i>Existing</i>	<i>New</i>						
A	Strong	3 (10%)	8 (27%)	2 (7%)	10 (33%)	3 (10%)	8 (27%)
B	Competent	17 (57%)	8 (27%)	13 (43%)	3 (10%)	20 (67%)	15 (50%)
C	Not yet competent	7 (23%)	10 (33%)	10 (33%)	7 (23%)	5 (17%)	
D	Unsatisfactory	3 (10%)	4 (13%)	5 (17%)	10 (33%)	2 (7%)	7* (23%)
	Total	30 (100%)	30 (100%)	30 (100%)	30 (100%)	30 (100%)	30 (100%)

*It was not possible to determine the “Not yet competent”/“Unsatisfactory” cut score for physiotherapy.

to the health professionals, with performances categorized as grade C. There was generally good agreement in the ranking of performances. In medicine, only 4 of the 26 performances (15%) were clearly ranked differently in the two sets of judgements. For nursing, 6 of 25 performances (24%) were in this category, and for physiotherapy, 3 of 24 performances (13%). This finding indicates the extent to which the two groups of raters are likely to be drawing on similar aspects of the performances to make their judgements (see Discussion below).

Comparing old and new standards

It was possible to compare the classification of a sample of 30 test performances from each profession using the existing cut scores, and based on the currently used criteria, with the classification of the same sample using the new cut scores, and based on the proposed criteria. This comparison is made based on the assumption that the four categories used by the panellists in the standard-setting exercise (“Strong,” “Competent,” etc.) can be set against the four main grades used to report OET results (A, B, C, or D).⁴ The results are reported in Table 3. This shows that using the new cut scores, and the new criteria, is likely to have an impact on rates of passing overall and of classification into the different categories if extrapolated from this relatively small sample to the wider test-taker population. In the data for medicine, more test takers were given grade A (i.e., moved up from B to A), but fewer test takers overall passed (i.e., there were fewer Bs and more Cs). For nursing, the impact on overall pass rates was slightly less, despite the fact that the proposed cut scores for nursing are lower than those currently used. Many more nurses were classified as grade A than as grade B. For physiotherapy, there was no impact on overall pass rates but, again, more test takers were classified as grade A than as grade B.

Discussion and conclusion

The study has shown the usefulness and practicality of the analytical judgement method in garnering the views of health professionals as to the minimum acceptable standards of spoken communication skill required of non-native-English-speaking health professionals in three professions – medicine, nursing and physiotherapy – wishing to enter supervised clinical practice. It enabled a range of levels of performance to be defined, by using scores given by language specialists associated with performances judged to be defining of a particular level of competence by the health professionals. The judgements were sufficiently consistent to permit the definition of agreed and defensible standards using this method. The inclusion of new cut scores using the proposed criteria will potentially lead to greater confidence in the attributes held by candidates, since these cut scores are based on the communicative qualities which health professionals consider to be important and on a view shared among them of what minimum standards are required.

We note, however, the need for an adequate number of standard-setting panellists to contribute to the process and for these panellists to be as fully representative of their profession as possible. In their chapter on standard setting, Tannenbaum and Katz (2013) summarize others' findings to propose that 10 to 15 panellists should provide acceptable results; the somewhat problematic findings in this study for physiotherapy would appear to support the need for this suggested range. Furthermore, a sufficient number of stimulus performances that lie around the cut-score boundaries (as perceived by panellists) must be available. Problems may arise if only a few performances are placed in the borderline categories, as the analytical judgement method does not use the non-borderline samples in the calculation of the cut scores. In our study, we encountered this issue in the physiotherapy data, as there were insufficient observations to define adequately the lowest level of performance. Another issue with stimulus materials is the quality of the stimulus tasks, and the quality of the interlocutor performance. If these are perceived as inadequate or distracting in some way by panellists, there may potentially be some impact on their standard-setting judgements. In their feedback on the procedures followed in this study, panellists raised some concerns relating to these areas, although the general consensus was positive towards the exercise (see Manias & McNamara, this issue).

When the ranking of test performances by health professionals was compared with the ranking given to the same performances by language-trained raters, there was considerable agreement, with most candidates ranked in a similar order. This was reassuring in that it suggested either that the health professionals and the language raters were oriented to similar qualities in the performances or, if they were not, that the impact of differences in orientation was relatively slight. It also suggested that the cut scores determined from the views of the health professionals are meaningful in terms of the judgements of language raters.

The performances used in this study had been scored by language-trained raters twice, once using the current set of criteria, based in applied linguistic theories of communicative competence and relatively insensitive to the orientation of health professionals, and a second time using the proposed reformed criteria, which incorporate more of what health professionals specifically value in clinical communication. As a result, it was

found that, when existing pass/fail cut scores using the current criteria were replaced by pass/fail cut scores using the proposed criteria, and based now to a greater extent on the judgements of health professionals, decisions about who should pass and fail were different in many cases. This suggests that adopting the new criteria and allowing health professionals to set cut scores will lead to more valid conclusions about admission to supervised clinical practice. In particular, the new criteria and the new standards lead to greater recognition of strongly competent communication, and a slightly tougher minimum standard for admission to clinical practice, particularly in medicine. Because of the confounding of the use of the proposed criteria and the use of the new standards, it is not clear which is the more influential, or whether their impact is synergistic. What is needed to clarify this is to see the effect of the new standards on scores using the old criteria. As in practice the recommended changes to current procedures involve both using new criteria and implementing new standards, the relative contribution of each then becomes more of an academic question than a practical one, and we were interested in understanding the combined impact of the changes.

It was noted that different cut scores were recommended for each of the professions involved. While those for physiotherapy and medicine were similar, the cut scores for the equivalent grades for nurses were uniformly lower. Reflection of this finding in policy would not mean changing the existing system of reporting levels of performance on the OET, but would have the cut scores associated with each level (i.e., final grade) defined differently for each profession as necessary. For example, a B grade would continue to represent the minimum requirement for most regulatory bodies, but the exact meaning of that grade, in terms of the range of fair scores it represents, would vary according to a test taker's profession.

New cut scores have been proposed for the three health professions included in the current study. Following the model adopted for this study, additional standard-setting workshops should be conducted with the remaining health professions and the impact of resultant changes to cut scores on pass rates should be considered.

Finally, while the evidence of this study on the whole is encouraging about language specialist raters having the capacity, when trained to be oriented to features of communication shown to be important to clinical supervisors, to act as proxies for the health professionals as judges of performance, the complementarity of perspectives remains an issue. In the context of specific-purpose language testing, and particularly in a case where the language test is legally required not to assess more than language, the language and communication skills and the professional skills, while "inextricably intertwined" (Jacoby & McNamara, 1999, p. 234), may involve a different focus of attention in each case. Consequently, the views of language-trained raters on a language test may not match well those of domain experts whose sense of what matters in this context is much broader and differently focused; potentially, at least, these views may still be incommensurable. It is unclear, for example, to what extent health professional panellists were able to discount professional knowledge and skills – such as professional conduct and adequate diagnostic reasoning – in their judgements, as they were instructed to do. The paper by Manias and McNamara (this issue), addresses this question, but further research, for example using verbal protocols to study the raters' decision-making processes, is called for. However, the issue of commensurability is not a problem for this test

alone; it is fundamental to the whole field of specific-purpose language teaching and testing, and thus goes far beyond this study.

Acknowledgements

The authors thank the participants in the standard-setting workshops for their time and opinions. Thanks are also due to members of the project team who facilitated the workshops.

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. Approval for the study was given by the relevant Human Ethics Advisory Group at the University of Melbourne.
2. A further reason for selecting the analytical judgement method for this study was its relative practicality in terms of panellist involvement.
3. Issues that may affect panellists' judgement are discussed in Manias and McNamara (this issue).
4. Note also that the sample test performances used in the study were drawn from regular OET administrations, and the grades using existing cut scores and criteria are those resulting from the routine analysis undertaken at the time. The grades were therefore based on the rating of *two* roleplay performances by each test taker, while only one of those performances was re-rated, using the proposed criteria, in this study. This comparison is consequently not a totally direct one.

References

- Australian Medical Council. (2015, August 1). AMC examinations (Standard Pathway). Retrieved from www.amc.org.au/assessment/pathways/standard/exams
- Berry, V., O'Sullivan, B., & Rugea, S. (2013, February). *Identifying the appropriate IELTS score levels for IMG applicants to the GMC register. Report submitted to the General Medical Council*. (150 pp.). Centre for Language Assessment Research (CLARe), The University of Roehampton, London.
- Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17(1), 59–88.
- Cizek, G. J., & Bunch, M. B. (2007). The body of work and other holistic methods. In G. J. Cizek & M. B. Bunch (Eds.), *Standard setting: A guide to establishing and evaluating performance standards on tests* (pp. 117–153). London: Sage.
- Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main: Peter Lang.
- Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes*, 18(3), 213–241.

- Kenyon, D. M., & Römheld, A. (2014). Standard setting in language testing. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 2, pp. 944–961). Hoboken, NJ: Wiley-Blackwell.
- Linacre, J. M. (2014). *Facets* [computer software]. Beaverton, OR: Winsteps.com.
- 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.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum.
- 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.
- O’Neill, T. R., Tannenbaum, R. J., & Tiffen, J. (2005). Recommending a minimum English proficiency standard for entry-level nursing. *Journal of Nursing Measurement*, 13(2), 129–146.
- Plake, B. S., & Hambleton, R. K. (2001). The analytic judgement method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283–312). Mahwah, NJ: Lawrence Erlbaum.
- Qian, H., Woo, A., & Banerjee, J. (2014, October). *Setting an English language proficiency passing standard for entry-level nursing practice using the Michigan English Language Assessment Battery*. NCLEX Technical Brief. National Council of State Boards of Nursing. Chicago, IL. Retrieved from www.ncsbn.org/14_NCLEX_technicalbrief_SettinganEnglishLanguageProficiencyPassing.pdf
- Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology*, Vol. 3: *Testing and assessment in school psychology and education* (pp. 455–477). Washington, DC: American Psychological Association.
- Wendt, A., Woo, A., & Kenny, L. (2009). Setting a passing standard for English proficiency on the Internet-based Test of English as a Foreign Language. *JONA’s Healthcare Law, Ethics, and Regulation*, 11(3), 85–90.
- Woo, A., Dickison, P., & de Jong, J. (2010, June). *Setting an English language proficiency passing standard for entry-level nursing practice using the Pearson Test of English Academic*. NCLEX Technical Brief. National Council of State Boards of Nursing. Chicago, IL. Retrieved from www.ncsbn.org/NCLEX_technicalbrief_PTE_2010.pdf

Appendix. Calculation of cut scores using analytical judgement method data for medicine.

Rater> Perf.V	Workshop 1						Workshop 2						Workshop 3						Count					
	rater 1	rater 2	rater 3	rater 4	rater 5	rater 6	rater 7	rater 8	rater 9	rater 10	rater 11	rater 12	rater 13	2	4	6	fair score	product 2	product 4	product 6				
	score	score	score	score	score	score	score	score	score	score	score	score	score	score	score	score	score	score	score	score				
1	5	5	5	5	5	6	6	5	6	5	6	6	6	0	0	5	4.89	0.00	0.00	24.45				
2	2	4	4	4	4	4	5	4	5	2	3	1	3	4	4	3.82	11.46	15.28	0.00					
3	4	5	4	3	1	4	3	5	5	4	5	3	3	0	4	4.74	0.00	18.96	4.74					
4	3	3	3	4										0	1	0	4.06	0.00	4.06	0.00				
5	4	4	3	4	3	4	4	3	4					0	7	0	4.44	0.00	31.08	0.00				
6														0	0	0	4.56	0.00	0.00	0.00				
7	6	5	5	5	6	5	6	6	6	3	3	3	5	0	0	7	4.84	0.00	33.88	0.00				
8	5	5	4	3	5	3	3	5	5	6	6	5	5	0	1	2	4.21	0.00	4.21	8.42				
9	4	3	3	2	4	3	4	5	4	5	3	5	5	1	4	0	4.22	4.22	16.88	0.00				
10	3	3	1	3	3	1	2	1	3	3	2	1	2	2	0	3.69	7.38	0.00	0.00					
11	5	4	5	4	5	3	6	4	6	5	5	6	6	0	2	4	4.77	0.00	9.54	19.08				
12	4	4	1	3	5	3	6	5	3	5	5	6	6	0	2	2	4.60	0.00	9.20	9.20				
13	4	5	2	5	4	4	4	4	3	6	4	5	3	1	6	1	4.44	4.44	26.64	4.44				
14	4	4	4	4	4	2	3	5	4	1	4	4	4	1	8	0	4.19	4.19	33.52	0.00				
15	4	4	3	5	5	2	2	4	4	2	3	3	3	3	4	0	4.19	12.57	16.76	0.00				
16	6	6	6	6	6	5	5	7	7	5	7	7	7	0	0	6	4.97	0.00	0.00	29.82				
17	5	4	4	6	5	4	3	5	5	6	4	5	5	0	4	2	4.46	0.00	17.84	8.92				
18	5	5	3	4	4	2	4	5	6	5	3	5	5	1	3	1	4.36	4.36	13.08	4.36				
19	4	3	4	2	4	3	3	1	3	4	2	1	1	2	4	0	3.57	7.14	14.28	0.00				
20	5	5	3	6	4	4	4	6	4	5	4	4	4	0	6	3	4.19	0.00	25.14	12.57				
21	5	4	7	6	4	4	6	5	6	7	5	6	6	0	3	4	5.22	0.00	15.66	20.88				
22	4	4	2	5	4	4	3	3	5	6	5	4	4	1	5	1	4.28	4.28	21.40	4.28				
23	4	3	5	4	3	3	4	2	4	6	1	3	3	1	4	1	4.18	4.18	16.72	4.18				
24	2	4	4	4	3	3	3	4	6	5	4	3	3	1	5	1	4.47	4.47	22.35	4.47				
25	2	5	4	4	3	3	3	4	3	1	3	1	1	1	3	0	4.20	4.20	12.60	0.00				
26	5	4	4	6	5	4	5	4	4	5	3	5	5	0	4	1	4.78	0.00	19.12	4.78				
	18 84 42																		72.89	364.32	198.47			
	Divided by																		18	84	42			
	Mean =																		4.05	4.34	4.73			
	cut score																		C/D	B/C	A/B			