The judgements of language-trained raters and doctors in a test of English for health professionals[1]

Tom Lumley

Abstract

Occupational experts are commonly used as informants during the process of development of spoken language assessment procedures in occupational settings. The rating process, however, tends to be conducted solely by language-trained specialists, normally teachers. Research to date has produced conflicting findings concerning the relative harshness and other characteristics of language-trained raters versus 'naïve' native speaker or occupational expert raters.

This question is considered in the context of a recent standard-setting project carried out for the Occupational English Test (McNamara 1990, Lumley, Lynch & McNamara 1994), an occupation-specific test of English for overseas-trained health professionals administered on behalf of the Australian Government. 20 audio recordings of role plays from recent administrations of the test were each rated by both 10 trained ESL raters and 10 medical practitioners.

The ratings produced by the two groups of judges were analysed to compare the extent of agreement they showed concerning candidates' language proficiency, as well as differences in their interpretations of the rating scale used. Broad similarities in judgements found between the two groups indicate that the practice of relying on ESL-trained raters can be justified.

[1] This research was conducted with the assistance of funds from the National Languages and Literacy Institute of Australia (NLLIA). An earlier version of this paper was presented at the 19th Congress of the Applied Linguistics Association of Australia, The University of Melbourne, July 1994. Thanks are due to Annie Brown and other colleagues in the NLLIA-LTRC for comments on earlier drafts of the paper.

Melbourne Papers in Language Testing 1995 Volume 4.1 pp. 74-98. The Language Testing Research Centre, The University of Melbourne.

1. Introduction

As part of the process of development of spoken language assessment procedures in occupational settings, occupational experts are sometimes used as informants (Brown 1993). The rating process, however, more commonly relies exclusively upon the judgements of language trained specialists (e.g., McNamara 1990).

Research to date has produced conflicting findings concerning the relative harshness and other characteristics of language trained raters versus 'naïve' native speaker or occupational expert raters (Galloway 1980, Barnwell 1989, Brown 1993). Barnwell (1989), for instance, found that a group of native speakers of Spanish, who had received no particular language training, were consistently harsher in their ratings of American students' performances in oral interviews than was an ACTFL-trained rater. Powers and Stansfield (1985) reported reasonable, though not high, levels of agreement between ESL teachers trained as raters and both nurses and consumers (patients), in a study of the Test of Spoken English (TSE). They found median correlations between scores produced by two kinds of judges (nurses and consumers) and TSE scores (produced by pairs of ESL-trained raters) of 0.66 to 0.68.

In the context of a test of Japanese for Tour Guides, an advanced level test of occupational language proficiency, Brown (1993) found that there were variations in rater behaviour between raters depending on whether or not they had experience in the tourist industry. These variations showed not in the level of harshness displayed by the two groups of raters, but in their sensitivity to particular assessment criteria. The raters with a teaching background but no industry experience also showed themselves reluctant to use the full range of score points on the rating scale used to assess candidates. The point of Brown's study was to establish whether or not the two groups could provide fair ratings for candidates, using the scale provided. Rasch analysis showed similar levels of consistency (or fit), and of overall harshness, amongst the two groups, but more variability in severity levels amongst the non-teachers. In addition, the two groups interpreted different criteria in different ways: the teachers rated more harshly on linguistic categories (grammar and expression, vocabulary and fluency), while non-teachers were harsher on pronunciation, as well as on one of the criteria, 'task fulfilment', on the task, 'Dealing with an upset or worried client', which requires particular skills relevant to the task of being a successful tour guide.

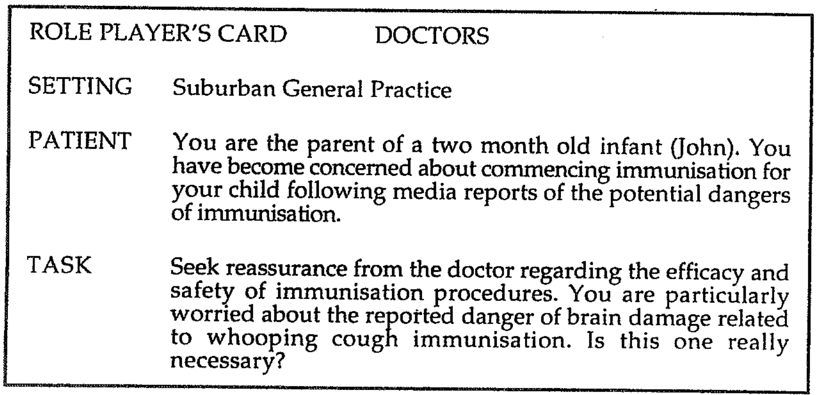
2. The Occupational English Test

The present study reports on the extent of agreement found in ratings given by two groups of judges, 1) ESL-trained raters and 2) doctors, during a standard-setting exercise carried out for a test of English proficiency for health professionals.

The Occupational English Test (OET) (McNamara 1990) is a four-skills test (speaking, writing, listening and reading) currently used by eleven health professions in Australia as part of their accreditation procedures for overseas-trained professionals wishing to practise in Australia.

The test includes common tasks for all professions for the listening and reading components, while the speaking and writing sub-tests have materials developed for each profession. The speaking sub-test, which is the focus of this paper, takes the form of an interview (unassessed), followed by two clinically-based role-plays.

The interaction takes place between an interlocutor, in the role of patient/client or the relative of a patient/client, and the candidate, who adopts his/her professional role (Figure 1).



ROLE PLAYER'S CARD DOCTORS

SETTING Suburban General Practice

PATIENT You are the parent of a two month old infant (John). You have become concerned about commencing immunisation for your child following media reports of the potential dangers of immunisation.

TASK Seek reassurance from the doctor regarding the efficacy and safety of immunisation procedures. You are particularly worried about the reported danger of brain damage related to whooping cough immunisation. Is this one really necessary?

Figure 1. Demonstration stimulus materials

Assessment is conducted live, only if the interlocutor is also an accredited assessor; otherwise it is conducted later from an audio tape of the interaction. All candidates are double rated. A six-point rating scale of the semantic differential type is used, for the six categories shown in Figure 2. The assessor carries out three assessments, one for each of the role plays, followed by a third, final assessment, based on impressions of the whole performance; this last is the assessment used for reporting candidate performance.

OVERALL COMMUNICATIVE EFFECTIVENESS

	6		5		4		3		2		1	
Near-native flexibility and range	___	:	___	:	___	:	___	:	___	:	___	Limited

INTELLIGIBILITY

Intelligible ___ : ___ : ___ : ___ : ___ : ___ Unintelligible

FLUENCY

Even ___ : ___ : ___ : ___ : ___ : ___ Uneven

COMPREHENSION

Complete ___ : ___ : ___ : ___ : ___ : ___ Incomplete

APPROPRIATENESS OF LANGUAGE

Appropriate ___ : ___ : ___ : ___ : ___ : ___ Inappropriate

RESOURCES OF GRAMMAR AND EXPRESSION

Rich, flexible ___ : ___ : ___ : ___ : ___ : ___ Limited

Figure 2. Occupational English Test—Rating categories and scale used

The pass score was originally set at a minimum of 4 for the category Overall Communicative Effect, plus an average score of 4 for the remaining 5 assessment categories.

The largest group of candidates is medical practitioners, but there are also significant numbers of nurses, dentists, vets and others.

It is not considered part of the role of the OET to assess the adequacy of a candidate's communication skills for professional practice in an unsupervised setting, but rather to make a judgement about whether or not the candidate should be able to participate successfully in the next stage of accreditation. This is normally a supervised, clinically-based bridging programme in a teaching hospital, during which time the candidate's English proficiency may be expected to improve with exposure to the communicative demands of the professional situation. Before being registered for practice in Australia, candidates are generally required to pass further tests: for doctors, these comprise a test of clinical competence and a test of medical knowledge.

There has recently been criticism from bodies representing health professionals that the pass standard for the OET was too low, so that candidates were passing the test with inadequate proficiency in English to cope with the demands of their profession.

This view received anecdotal support from other quarters, too, including some of the teachers involved in preparing candidates for the test. Their concern appeared to be motivated by problems candidates may face if they pass the test with levels of proficiency too low for them to gain entry to or be successful in the clinically-based bridging courses, or too low for them to gain employment in their field. It would appear more productive from the point of view of test candidates to spend more time acquiring a sounder grasp of English than in struggling against odds which are already stacked fairly heavily against them, with the added burden of communication difficulties.

The issue of setting standards in language tests is largely a political process (as Lumley, Lynch and McNamara 1994 discuss in greater detail). In the case of the OET a tension exists between the views of advocates of the immigrant professionals (who generally press for a more lenient standard), and those of the representatives of professional registration boards (who typically advocate more stringent criteria). In recent years, the views of the advocates of the immigrants have held greater sway, with the result that the OET has not been a difficult test, often having pass rates of 70–80% or more. The major decisions regarding candidates are thus moved to one of the examination procedures conducted by the councils representing the health professions once candidates have passed the OET.

It was recognised by the test developers, nevertheless, that the recent criticisms of low pass standards may not be unfounded. For example, it was considered possible that the raters' view of the criterion level required to pass the test had slipped since the introduction of the test in its present format in the late 1980s. It was therefore decided to conduct a study to determine whether a revised pass level is necessary for the test. Because most of the criticism focussed on candidates' oral interaction, the speaking sub-test was selected as the initial area of investigation. For practical reasons it was also necessary to restrict the scope of the study to candidates of a single profession, and doctors were chosen, as the dominant group in numerical terms.

3. Purpose and methodology

The purpose of the standard-setting study was to establish a new criterion level for performance on the speaking sub-test. Following Powers and Stansfield (1985) it was decided to employ the judgements of 1) representatives of the medical profession and 2) trained ESL raters who regularly rate test performance. The ratings given by the doctors would then be compared with the ratings given during the study by the ESL raters. Clearly, in such a context, the issue of difference or similarity between these two groups in their perceptions of candidates' language proficiency is important. It is necessary to consider whether the judgements were in fact comparable, or whether the two professional groups perceived candidate proficiency in this context in quite different terms.

The following questions will be considered in this study:

Question 1: To what extent did the two groups agree on classification of candidates as pass/fail?

Question 2: Were ESL raters as a group more lenient than doctors?

Question 3: What evidence is there for differences between judgements made by the individual raters?

10 ESL raters, trained and experienced in assessment of the OET, and 10 doctors, were initially selected to take part in the study. The doctors were for the most part chosen on the basis of their having had extensive experience working with overseas-trained doctors working in Australia in the clinically-based bridging programmes mentioned earlier, giving them familiarity with the issues faced by these doctors in professional settings. Two doctors were included as representatives of the Australian Medical Council (AMC), the professional body responsible for accreditation of medical practitioners in Australia. One of these had similar experience to the other eight, while the other occupied a senior position on the Examining Body of the AMC.

20 audio tapes of test candidates were selected from recent test administrations, from a range of the national and language groups most commonly represented in the test population. They covered a range of score points, above clear fail, but most were clustered in the range of an average score across rating categories of between 4 and 5 (on a scale of 1 to 6), the range in which it was anticipated the new pass level would fall. All participants rated these 20 tapes. Due to pressure of work, one doctor was unable to complete the task, and data were only collected from 9 of the doctors.

A significant difference between Powers and Stansfield's (1985) study and the present one lies in the samples of language on which judges were asked to make decisions. Powers and Stansfield asked judges (both nurses and consumers) to make judgements about candidates' English proficiency for three different general situations in which nurses might be engaged (hospital nursing, public health nursing and teaching). In making these judgements they relied on samples of oral language elicited by tasks on the TSE, a test of general proficiency with content not specifically related to any health profession, and a test in which the ESL-trained raters are not asked to consider any specific occupational context when making operational judgements. In the OET, by contrast, judges (both the doctors who participated in this study and the ESL-trained raters who conducted the rating for this study, under operational conditions) are asked to make judgements about candidates' proficiency to function generally in the communicative contexts of a particular medical setting (a clinically-based bridging course in a hospital). In making these judgements they rely on candidates' performance in a test with content designed specifically for health professionals, with a task simulating a situation medical practitioners might expect to encounter routinely, i.e. a medical consultation.

4. Briefing process

The doctors were all given, either individually or in small groups, a short briefing session (30–45 minutes), the main purpose of which was to clarify the judging task. During this session, most of them expressed an opinion about the issue of English proficiency of overseas-trained doctors. Views of individual participants varied: generally most, but by no means all, overseas-trained doctors were perceived as having adequate proficiency in English. One or two participants thought the English language proficiency of overseas-trained doctors was a very serious problem; others felt that issues related to communication by these doctors in professional settings are not necessarily best conceptualised as a language problem, but may include a wide range of other factors, many of them cultural, and that there are potent reasons of equity which should not demand standards of communication from immigrant doctors that are not assessed in native English speakers.

The practical point of interest in this study was a simple judgement of whether or not the candidate was considered to have adequate proficiency in English to participate successfully in a supervised clinical bridging programme. It was felt impossible in this context to expect useful judgements from the doctors on the full range of linguistic assessment categories, without extended discussion of how each individual category should be interpreted. In effect this would have required a lengthy training session, which would have been incompatible with the pressure of their work as well as running the risk of unduly influencing their judgements. They were therefore provided with a list of the categories, with a brief gloss for each one and asked to make only a single holistic judgement, using the category, 'Overall Communicative Effect', on only the first role-play from each candidate. A full set of the instructions provided to the doctors is given in Appendix 1.

ESL raters	Doctors
language experts	occupational experts
received training as raters for the test	no training as raters
reliability established after training	reliability not established
used 5 explicit linguistic categories of assessment plus 'Overall Communicative Effect'	one judgement only, no particular linguistic categories: 'Overall Communicative Effect' only

Figure 3. Features of the two groups of judges

The ESL raters received no particular briefing, since they had all been trained previously as raters for the OET, and had all taken part in regular rating for the test recently. In order to be able to make meaningful comparisons between the judgements produced by the two groups of judges, only the judgements on the category 'overall communicative effectiveness' for the first role play were analysed. As reported by McNamara (1990) this holistic category represents the best summary of the ratings provided for all categories of assessment, although it is likely to be influenced by additional features of the candidate's performance to the linguistic ones specified.

The differences between the two groups of judges are summarised in Figure 3. As can be seen, the two groups are polarised in a number of ways, and it would therefore not be surprising if we were to observe substantial differences between them.

candidate ID no. (N=20)	ESL raters (N=10)	Drs (N=9)	Total (N=19)	
414	**10**	**9**	**19**	most proficient
416	**10**	**9**	**19**	
291	**10**	**9**	**19**	
46	**10**	**9**	**19**	
126	**10**	**9**	**19**	
311	9	**9**	18	
199	**10**	7	17	
293	**10**	8	18	
110	9	**9**	18	
174	8	7	15	
141	9	6	15	
91	7	**9**	16	
53	5	7	12	
129	6	6	12	
104	4	3	7	
249	3	5	8	
114	2	2	4	
176	3	1*	4**	
179	**0**	2	2	
66	**0**	**0**	**0**	least proficient

Complete agreement within each group is marked in bold type
* [of 7]
**[of 17: poorly audible; 2 ratings missing]

Table 1. No. of raters classifying each candidate as 'pass' (raw score = 4.0 or more), Overall Communicative Effect only

5. Results

Question 1: To what extent did the 2 groups agree on classification of candidates as pass/fail? This question is of course central in the context of standard setting, as well as in considering the comparability of the two groups of judges.

Table 1 shows that, as indeed one might expect with single, holistic ratings on a subjectively marked test, there was considerable variation in levels of agreement within and between the two groups of raters over pass/fail categorisations. Only the 5 most able candidates were universally judged by both groups as passing; at the other end of the scale, there was complete agreement only over the least proficient candidate, no. 66. A further 3 candidates were passed by all the doctors, but failed by at least one ESL rater, while two other candidates were passed by all ESL raters, but failed by one or more doctors. The ESL raters also agreed that candidate no. 179 should fail. This leaves 8 candidates over whom there was larger disagreement.

Question 2: Were ESL raters as a group more lenient than doctors, as had been predicted would be the case?

Using the scores allocated by both groups of raters on the single category of overall communicative effect, then the answer is, counter to expectations, no, as has been reported in Lumley, Lynch and McNamara (1994). Table 2 shows that with a raw score pass level of 4.0, whereas the average score given by the doctors as a group would allow on average 13 of the sample to pass, the average score given by the ESL raters would pass only 11, so if anything, the ESL raters appear harsher than the doctors. Examination of mean scores produced by the two groups, on the other hand, shows there was no difference between them.

This table also sheds more light on the issue of consistency between the two groups: it is worth noting that all candidates passed as a group (i.e. with a mean score of 4.0 or above) by the ESL raters were also passed by the doctors, while no candidate failed by the doctors as a group (i.e. with a mean score below 4.0) was passed by the ESL raters.

candidate ID no.	ESL raters	Drs' judgements	Combined ratings, ESL & Drs	
414	6.0	5.6	5.8	most proficient
416	5.4	5.4	5.4	
291	5.0	5.3	5.2	
46	4.6	5.0	4.8	
126	4.6	4.7	4.6	
311	4.2	4.6	4.4	
199	4.6	4.0	4.3	
293	4.4	4.2	4.3	
110	4.1	4.3	4.2	
174	4.0	4.0	4.0	
141	3.9	4.0	4.0	
91	3.7	4.1	3.9	
53	3.6	4.0	3.8	
129	3.6	3.6	3.6	
104	3.3	3.2	3.3	
249	3.1	3.3	3.2	
114	2.9	3.0	3.0	
176	3.1	2.1	2.6	
179	2.6	2.3	2.5	
66	2.4	1.9	2.2	least proficient
	3.94	3.93	3.93	mean
	0.93	1.06		s.d.

Table 2. Mean scores produced by ESL raters and Doctors: 'Overall Communicative Effect' only

Question 3: What evidence is there for differences between the individual raters?

Table 3 shows the number of candidates passed by each judge.

ESL Raters (N = 10)

Judge ID no.	No. Passed	No. Failed	
211	18	2	lenient
251	17	3	
226	14	6	
248	14	6	
255	14	6	
278	14	6	
279	13	7	
202	12	8	
229	11	9	
246	8	12	harsh

Doctors (N = 9)

Judge ID no.	No. Passed	No. Failed	
101	17	3	lenient
108	16	3	
102	16	3	
106	15	5	
103	14	6	
105	14	6	
107	13	7	
104	13	7	
109	8	12	harsh

Table 3. No. of candidates passed by each judge,
Overall Communicative Effect only

We can see that there are in fact very substantial differences here in the degree of severity shown by individual raters, with the ESL raters each passing between 8 and 18 of the candidates, and the doctors each passing between 8 and 17. ESL rater no. 246 and Doctor no. 109 are both considerably harsher than the rest of either group. There also appears to be slightly less variation amongst the doctors concerning the number of passing candidates than among the ESL raters.

A product-moment correlation table (Table 4) was also produced for pairs of judges, in order to show trends in consistency of agreement between each judge and every other.

rater	202 esl	211 esl	226 esl	229 esl	246 esl	248 esl	251 esl	255 esl	279 esl	278 esl	101 dr	102 dr	103 dr	104 dr	105 dr	106 dr	107 dr	108 dr
211	.776																	
226	.921	.921																
229	.934	.664	.874															
246	.807	.331	.639	.859														
248	.936	.908	.956	.862	.65													
251	.788	.979	.932	.707	.397	.914												
255	.936	.911	.974	.882	.651	.975	.924											
279	.945	.824	.937	.891	.755	.949	.873	.955										
278	.919	.902	.974	.866	.64	.949	.929	.965	.938									
101	.814	.975	.93	.711	.429	.932	.981	.935	.871	.939								
102	.794	.956	.905	.691	.387	.919	.968	.906	.852	.913	.978							
103	.911	.915	.992	.861	.642	.957	.932	.967	.94	.968	.938	.913						
104	.93	.813	.933	.856	.739	.926	.843	.915	.933	.943	.86	.843	.935					
105	.912	.87	.95	.837	.675	.957	.907	.937	.954	.945	.919	.907	.964	.963				
106	.894	.93	.967	.834	.597	.957	.949	.967	.935	.966	.956	.941	.967	.93	.944			
107	.918	.82	.935	.88	.687	.92	.83	.936	.909	.929	.846	.85	.918	.868	.866	.911		
108	.786	.959	.917	.696	.356	.893	.953	.903	.816	.911	.943	.963	.909	.847	.867	.94	.845	
109	.815	.323	.632	.857	.976	.651	.388	.636	.737	.641	.425	.407	.636	.755	.682	.598	.679	.385

Table 4. Correlations between judgements on 'Overall Communicative Effect', ESL raters and Doctors

Correlations between pairs of raters of less than 0.8 are shaded. For the most part there is a correlation between each pair of judges of greater than 0.8, which while not especially high, nevertheless constitutes a reasonable level of agreement[2]. It can be seen that one ESL rater, no. 246 and one doctor, no. 109, are responsible for most of the disagreement that is apparent. Thus these two judges are out of line with the others, in both groups, not simply in terms of their harshness (as we saw in Table 3), but also their consistency (i.e., the extent to which they agree with the others about the rank order of ability of the candidates). This is further illustrated by the mean correlations for each rater with all the other raters, which varied between 0.83 and 0.91 for all raters except 246 and 109, who each had a mean correlation with the other raters of only 0.64. It is interesting, nevertheless, that these two judges show a very high level of agreement with each other (0.976). The levels of correlation between the pairs of judges generally are quite impressive given that they are the result of single ratings only on a single sample of language. It is worth noting that these figures are noticeably higher than the (median) correlations reported by Powers and Stansfield (1985) between judges (including both nurses and consumers) and TSE scores (produced by pairs of ESL-trained raters) of 0.66 to 0.68.

With regard to Brown's (1993) finding that language specialists may be reluctant to use the full range of score points on a scale when rating occupational language tests, Table 5 shows that neither group uses the lowest score category much at all, although there is clearly a greater reluctance on the part of the ESL raters than the doctors to use the two lowest score points. At the other end of the scale, however, the highest score, 6, is used equally by both groups. Generally, the ESL raters do seem to prefer to use the middle points on the scale, with 62% of their ratings falling into the categories of 3 or 4, compared to the doctors, who only used those categories for 52% of their scores. Possibly this is accounted for by the ESL raters reserving the use of the lowest two categories for the weakest performances that they sometimes encounter when rating test

[2] It should be noted here that for the purposes of reporting individual candidates' scores operationally, there are two mechanisms which improve the reliability of candidates' scores. Firstly, all candidates are rated twice. Secondly, multi-faceted Rasch analysis (Linacre 1989) is used, which takes into account the relative harshness or severity of the judges rating each candidate, and compensates accordingly, building this into the scores reported for candidates.

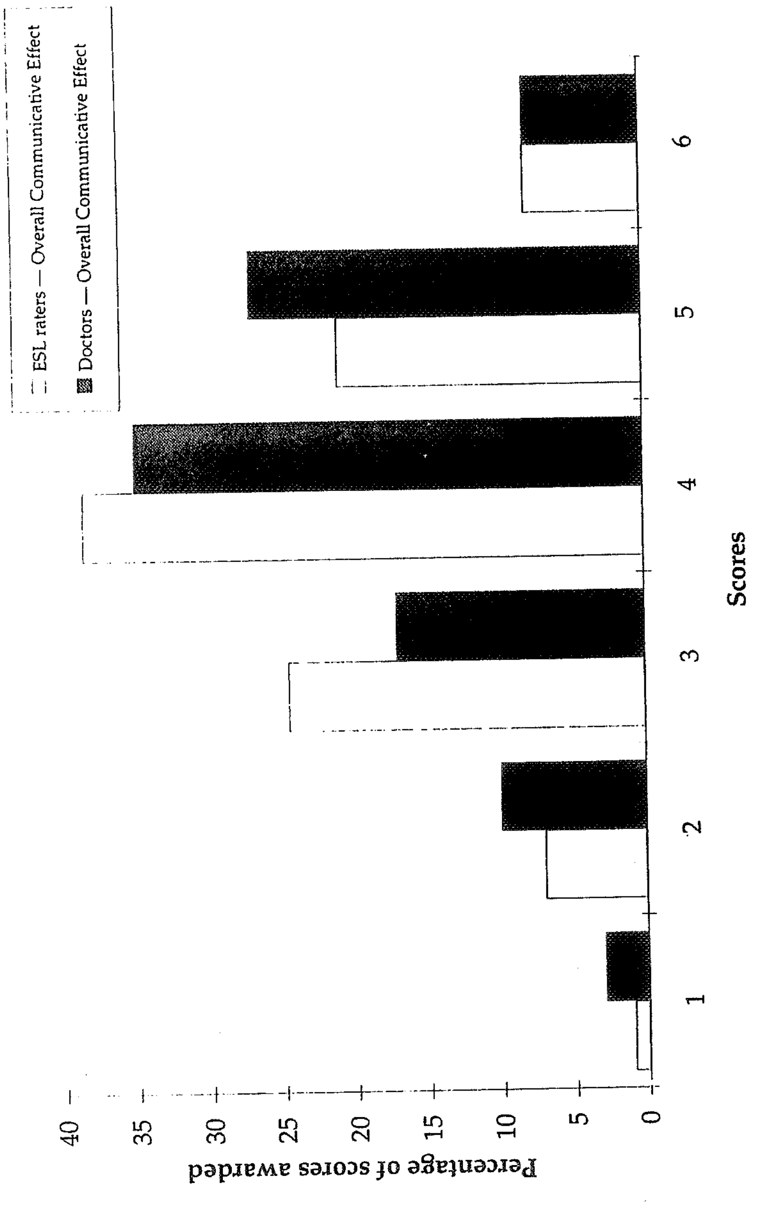


Table 5. Use of Score Categories, ESL raters and Doctors

administrations, examples of which were not selected for this study, whereas the doctors have no such experience to draw on.

Discussion

The most significant finding to emerge from this study is that at a global level there seems to be considerable agreement between the two groups (although markedly different judgements were produced by one member of each group). In other words, it seems quite reasonable for ESL raters to make judgements in this sort of occupational setting: it would appear that it does not matter too dramatically which of the two groups conducts the rating. This is a reassuring finding in the context of this test, providing clear evidence for the validity of the ratings made operationally by ESL teachers.

It is possible that the relatively high levels of agreement generally shown between pairs of judges, whether doctors or ESL raters, reflect the fact that the OET is a specific-purpose, rather than a general-purpose test, unlike the TSE. This may result in judges, albeit of different professional backgrounds, being more likely to agree on candidates' performance when relating it to the same context (clinical consultations), and when their judgements rely on a language sample relevant to that context.

Nevertheless, considerable variation has been observed between individual members of each group. This study emphasises once again the need for more than a single rating of performances on subjective tests such as this one, and points also to the need for some form of mechanism that can compensate for differences in relative harshness or leniency of raters, as shown by those involved in this study. It is clear that even despite training, differences will remain.

It is somewhat surprising that rater no. 246, with many years' experience rating the OET, should appear to be so out of step with the others. However, this kind of phenomenon, where a very experienced rater may suddenly lose his/her consistency with the group, has been observed, if not well chronicled, elsewhere in performance assessment. If training assists raters to be internally consistent; as has been found (Cushing 1993, Weigle 1994), then a need for retraining would appear to be demonstrated. Whether or not retraining would help for the inconsistent rater identified here is unclear, but is clearly a point worth investigating. Given that this rater has a long history of reliable rating on the test, and so the inconsistency was not predicted, then regular retraining would seem appropriate for all raters. Another approach would be to investigate the reasons for the substantial agreement found between this judge, no. 246, and no. 109, comparing the criteria on which they made their decisions with the criteria influencing the other judges. This clearly relates to the construct validity of the test, as represented by the ratings made by all judges in this study, whether doctors or ESL raters. Such an approach might also shed light on the reasons for the differences that were observed in Table 1 between the groups' perceptions of individual candidates. We may speculate that the doctors were more influenced by the content of what was said by candidates, while the ESL raters concentrated more on purely linguistic aspects of the interaction, but this suggestion needs careful investigation.

The wider variation observed between the individual ESL raters, compared to the doctors (see Table 3 above), concerning the number of candidates who should pass the test, may be partially attributable to the different rating processes employed by the two groups: it is conceivable that the more complex rating task conducted by the ESL raters (considering six categories of assessment) leads them to produce more diverse ratings than the doctors. It is quite possible that different linguistic features are dominating the judgements of different raters for each candidate. Again, it would be interesting to determine what these features are and how they influence ratings.

This study resulted in the OET speaking sub-test becoming slightly harder to pass (Lumley 1994); this decision was made largely in response to the demands of the political context in which the test is used. The raw scores, as we saw, showed the doctors to be no harsher than the ESL raters, in fact if anything the reverse. This point requires discussion, raising as it does questions about the validity of the test, since in another sense the principal complaint made had been that the OET raters were perceived as too lenient, a suggestion which finds no support in this study.

It may be that either the tasks presented in the role-plays or the communicative demands of the test situation do not adequately represent the kind of oral communication where test candidates may in real life show themselves to be lacking proficiency. This is not really surprising, given that the test was designed as an example of a 'weak' performance test, to use McNamara's (1990) term; that is to say, its primary purpose is the elicitation of a sample of language which can be assessed, and the occupational focus is only used to provide a context that appears generally relevant to the participants in the test. For example, it may be that there is a problem with the interlocutors, who are for the most part middle-class, well-educated, articulate native speakers of a rather standard variety of Australian English, whose ESL training has alerted them to the potential for miscommunication in spoken interaction, and which, one may fairly safely presume, they would take some trouble to avoid in a testing situation. They may not represent sufficiently well the kind of patients or clients with whom health professionals need to interact, or, at least, the range of patients with whom they work. Involved here is very likely the intractable issue of breadth of comprehension, involving perhaps the ability, or lack of it, to process idiomatic language (possibly avoided or simplified by ESL teachers), a skill which is largely untested in the OET. Another feature which may be insufficiently considered is the ability of candidates to clarify, expand or rephrase explanations and courses of action in different ways when interacting with patients of different backgrounds or with different needs. There are possibilities here for further research concerning the authenticity of the task and of the interaction between candidate and interlocutor. Alternatively, the issue may be less involved with language proficiency than with cultural expectations. There are numerous other possibilities.

So, do we need a 'stronger' performance test than this, in McNamara's terms, involving judgements about the candidate's professional competence, in some way? If so, who would be competent to make the assessments? Assessors already express concern on occasion over the extent to which they should be making judgements about the candidate's knowledge or understanding of professional terminology. They may be influenced by their own personal experiences with health professionals, which would represent an exceedingly incomplete basis for making consistent decisions. In essence, ESL-trained raters are neither permitted nor competent to pass judgement on such matters. Doctors, on the other hand, would a) require training and b) be too expensive, to make the kind of complex ratings made by the ESL raters.

In conclusion, lacking answers to all of these questions, there seems to be no convincing argument yet for using other than the ESL raters, who appear to agree reasonably well with the occupational experts, the doctors; provided, of course, that their reliability is continually monitored and shown to be acceptable.

References

Barnwell, D. (1989) Naïve native speakers and judgements of oral proficiency in Spanish. *Language Testing*, 6, 2: 152–163.

Brown, A.D. (1993) The effect of rater variables in the development of an occupation-specific language performance test. Paper presented at the *Language Testing Research Colloquium*, Cambridge, August.

Cushing, Sara. (1993) Effects of training on raters of ESL composition. Paper presented at the *American Association for Applied Linguistics* Annual Meeting, 1993.

Galloway, V. (1980) Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal*, 64:428–433.

Linacre, J.M. (1989) *Many-faceted Rasch Measurement*. Chicago: MESA Press.

Lumley, T. (1994) *Recalibration of the pass standard for the Occupational English Test*. (Report submitted to the National Office for Overseas Skills Recognition, Canberra, May.)

Lumley, T., Lynch, B., & McNamara, T.F. (1994) A new approach to standard-setting in language assessment. Paper presented at the *American Association for Applied Linguistics* Annual Meeting, Baltimore, March.

McNamara, T.F. (1990) *Assessing the Second Language Proficiency of Health Professionals*. Unpublished Ph.D. thesis, The University of Melbourne.

Powers, D.E. & Stansfield, C.W. (1985) Testing the oral proficiency of foreign nursing graduates. *The ESP Journal*, 4: 21–35.

Weigle, Sara C. (1994) Using FACETS to model rater training effects. Paper presented at the *Language Testing Research Colloquium,* Washington, DC, March.

Appendix 1

Occupational English Test (OET): Speaking

Standard-setting Project, 1994

Information sheet for supervisors of clinical bridging programmes

The aim is to elicit opinions, from supervisors who have experience of training overseas-trained medical practitioners, of the minimum working knowledge of English required in a supervised clinical setting. You will be asked to listen to a series of 20 audio recordings from recent administrations of the OET, and on the basis of these make a judgement concerning the adequacy of each candidate's English for participation in supervised clinical practice.

Assessment in the speaking sub-test of the OET is carried out on the basis of performance on two tasks, each lasting approximately 4 to 8 minutes. These tasks take the form of role plays: simulated consultations between the candidate (adopting his/her professional role) and a native speaker of English (in the role of patient or client or the relative of a patient/client). For the purposes of the current study, you will listen only to the first of these role plays for each candidate.

Before and during the test the candidate is constantly reassured:

1. that the purpose of the interaction is to elicit a sample of language on the basis of which a judgement may be made about his/her English language proficiency; and

2. that no judgements are made concerning the candidate's medical knowledge.

The medical content of the interaction, and the quality of the advice given, are therefore irrelevant to decisions made during this test of language. You should therefore completely set aside any judgement of the candidate's clinical knowledge or experience.

The following scale is used in rating candidates:

OVERALL COMMUNICATIVE EFFECTIVENESS

	PASS					FAIL	
Near-native flexibility and range	6	5	4	3	2	1	
	___ :	___ :	___ :	___ :	___ :	___	Limited

The points on the scale should be interpreted as follows:

6: There is no doubt about the candidate's ability to communicate effectively in English.

5: The candidate would clearly be able to cope successfully with the linguistic demands of a supervised clinical bridging programme.

4: The candidate has the minimum competence necessary to cope with the linguistic demands of a supervised bridging programme in a clinical setting.

3: The candidate does not quite have the minimum competence necessary to cope with the linguistic demands of a supervised bridging programme in a clinical setting.

2: The candidate would clearly fail to cope with the linguistic demands of a supervised clinical bridging programme.

1: The candidate has no more than a fairly elementary level of competence in English, and should probably not even be taking this test.

The scale is thus meant to indicate a range from a very advanced to a fairly elementary competence. Candidates who pass the OET may be eligible to apply for a place in a supervised bridging programme in a teaching hospital, provided they also pass any additional screening tests of medical knowledge/clinical competence that the programme may require as part of its admission procedure. A passing level (nominally mid-way between score points 3 and 4) will therefore represent the minimum competence with which a candidate could cope with a bridging programme in a clinical setting, involving interaction with patients/clients, clinical teachers and colleagues.

In making your decision you should consider the following questions:

- Could this person cope without undue embarrassment to him/herself or to others (supervisors, clinical teachers, patients, relatives of patients, colleagues) with the communicative demands of this supervised setting?

- Do you think this person would find the communicative demands of such a setting unreasonably stressful?

- Could you manage to communicate effectively with this person in a clinical bridging programme you were supervising?

- Could your patients manage necessary communication with this person?

- Could your colleagues manage to communicate effectively with this person in a supervised clinical bridging programme?

Language features which may contribute to your decision include the following (this is not an exhaustive list):

Intelligibility

(e.g. How easy is it to understand the candidate's pronunciation? Does it require undue strain to listen to the candidate? Does it become easier to understand him/her as you get used to the accent/style of speech?)

Fluency

(e.g. How evenly does the candidate speak? Does speech flow at a rate which enables the listener at least to follow the conversation?)

Comprehension

(Does the candidate appear to understand most of what the patient expresses about his/her concerns?)

Appropriateness Of Language

(e.g. Are appropriate expressions used in explaining medical conditions or courses of action to the patient? Is any inappropriate choice a real barrier to communication?)

Resources Of Grammar And Expression

(e.g. Does the candidate have adequate vocabulary and control of grammatical expression to express necessary ideas clearly and unambiguously? Are any deficits here so serious as to form a real barrier to communication?)

At the end of the role play, enter an assessment using the six-point scale of overall communicative effectiveness as shown above. Use a cross to mark which of the six points on the scale best locates the candidate's performance in that category. Please DO NOT place a mark between two score points.