A New Approach To Standard-Setting In Language Assessment[1]

Tom Lumley, Brian K. Lynch and T.F. McNamara

1. Introduction

Background and purpose

In this paper, we consider a standard setting exercise involving the Occupational English Test (McNamara 1989, 1990a, 1990b), a specific purpose test of English as a second language for health professionals. This Australian test is used as part of the screening of immigrant and refugee health professionals prior to the resumption of their professional careers in their new country of residence.

The purpose of the paper is to consider the role of new technical advances available in Generalizability theory and multi-faceted Rasch measurement in addressing familiar problems of eliciting judgments from experts. The place and limitations of these technical advances within the political and social context of this test will be considered. The paper argues that the new techniques can assist in and improve aspects of familiar approaches to standard setting, but that the essential decisions remain questions of value, hence political, not merely technical ones.

Technical advances

The new techniques used in this paper are Generalizability theory (Brennan 1983; Shavelson & Webb 1991), implemented through the program GENOVA, and multi-faceted Rasch measurement (Linacre, 1993) implemented through the program FACETS. The use of these approaches in investigating the behaviour of raters in performance assessment settings has been the subject of recent papers (Bachman

[1]The research for this paper was made possible by a grant from the Australian Commonwealth Department of Employment, Education and Training administered through the National Languages and Literacy Institute of Australia (NLLIA). The research was carried out at the NLLIA Language Testing Research Centre in the Department of Applied Linguistics and Language Studies at the University of Melbourne.

Melbourne Papers in Language Testing 1994 Volume 3.2 pp. 19-40. The Language Testing Research Centre, The University of Melbourne.

et al. 1993; Lynch and McNamara 1994). Briefly, the approaches permit investigation of the nature and extent of variability between and within raters, and interaction effects involving raters and particular groups of candidates or particular test tasks. To date, these techniques have not been used to investigate the qualities of expert Judges in standard setting exercises in language testing contexts, but have obvious potential as the judging task is central to approaches to standard setting such as that reported in Powers and Stansfield (1985), in which nurses and patients were asked to rate the acceptability of performances by non-native speaking nurses on the Test of Spoken English.

Social and political background

As part of its annual intake of immigrants and refugees, some hundreds of overseas-trained health professionals are currently entering Australia each year as permanent residents. The majority of these are medical practitioners, but a number of other health professional groups (nurses, dentists and several others) are also represented. The process of registration for practice in Australia typically involves the following three stages after an initial verification of documentation:

1. The Occupational English Test (OET), an English language proficiency test taken by each of the professions involved; its development and administration are in the hands of the National Languages and Literacy Institute of Australia on behalf of the National Office for Overseas Skills Recognition.

2. Profession-specific pencil-and-paper tests of professional clinical knowledge, developed by the relevant professional examining body, for example the Australian Medical Council.

3. Performance-based tests of clinical competence, again conducted by the relevant professional examining bodies.

Stages 2 and 3 are demanding for those health professionals whose clinical experience is restricted to contexts rather different from Australia, a country with a technically sophisticated health care system, where diseases are linked to the lifestyle of a relatively affluent industrial and post-industrial society. This is true for a majority of candidates, who most frequently come from countries in

Eastern Europe, the Middle East, the Indian sub-continent and South-East Asia. In order to have a chance of success on these practical, clinically based tests, reflecting roughly the standard of final year medical training in Australia, candidates must have some access to experience of clinical practice in Australia. A limited number of places on hospital based bridging programs is available. Even then, it is normal for it to take up to two years between the time of arrival in Australia and the completion of the registration process. This is both expensive and potentially damaging to the health professionals' confidence and level of clinical knowledge and skill, which will deteriorate if it is not kept up by practice.

The entry of large numbers of their professional colleagues from other countries presents a challenge to the Australian health professions, many of whom are under pressure to reduce significantly the output of graduates from Australian medical schools and other health professional training institutions. For example, the current level of new registrations of overseas-trained medical practitioners represents the equivalent of the output a major medical school, many of which might be fighting pressure to substantially reduce the scale of their operations. The situation is exacerbated by the fact that access to the necessary clinically based retraining, some of it fairly substantial, is in the hands of those clinical educators most conscious of the pressures on the local training institutions. Despite this, the professions have on the whole maintained a cooperative attitude, and there are many instances of individuals who have worked long and hard in the interests of the overseas trained.

In such a situation it is tempting for the local professions to use the stages of the registration process, including the OET, to create barriers to registration. The predecessor of the OET in the form in which it existed in the 1970s certainly was so used. The test included difficult and obscure literary texts, and there was a very high failure rate. The situation was reformed in the mid-1980s, and the current form of the English test, which is based on the communicative tasks of the workplace, was introduced in 1987. The test currently has 4 subtests, one for each macroskill, with profession-specific material in two of the subtests (speaking, writing) within a common format. In the case of the speaking sub-test, the candidates is required to adopt his or her professional role and to engage in a role play with an interlocutor who will simulate

a patient or the relative of a patient. An example of the role play materials used in the speaking sub-test is given in 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 immunization for your child following media reports of the potential dangers of immunization.
TASK	Seek reassurance from the doctor regarding the efficacy and safety of immunization procedures. You are particularly worried about the reported danger of brain damage related to whooping cough immunization. Is this one really necessary?

Figure 1: Demonstration stimulus materials

At around this time pressure was removed from the OET in this regard, and attention was transferred to the stage 2 and 3 tests.

Crucial to this shift was a decision made in the mid 1980s about the standard of the passing score on the OET. It was set in relation to performance on the Australian Second Language Proficiency Rating Scale (ASLPR) (Ingram, 1984), an Australian oral proficiency interview (OPI) test related closely to the FSI and ACTFL scales. Initially, ASLPR 4, roughly equivalent to FSI level 4, was set as the benchmark, but this was strongly resisted by the overseas-trained health professionals and their advocates, including English language teachers, as it was felt that it would unnecessarily delay their entry to the all important clinically based retraining programs, where in any case their English language skills could be expected to improve. The campaign to lower the required level was successful, with the result that the OET was no longer a difficult test; pass rates of up to 80% became frequent.

In other words, the initial setting of benchmarks was a political process, which in this instance was won by the candidates. The Australian Medical Council, responsible for the clinical examinations and deeply involved in the process of registration of overseas-trained doctors, continued to make their case. As

complaints grew from clinical supervisors and examiners over what were perceived to be the poor communication skills of non-native speaker doctors, pressure grew for the benchmark of ASLPR 4 to be reinstated. It was in this context that the present study was undertaken, as a way of rationally determining what was an acceptable level of performance in the eyes of clinicians with experience of working with overseas-trained doctors in a supervisory capacity. Following Powers and Stansfield (1985), it was decided to elicit judgements from these doctors of the relative acceptability of performances on the OET speaking subtest. These would then be compared with the standards being applied by OET examiners, who were repeatedly being criticized for being too lenient.

2. Methodology

Performances on the speaking sub-test are rated by two raters working independently, using a rating scale reproduced in Figure 2 (Note: a "4" on this scale is considered equivalent to a "3" on the ASLPR). An audiotape recording is made of each interaction.

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

Data Collection

20 of these audiotapes were selected from recent test administrations (18 from 1993 and 2 from 1991) (see Table 1). There were 10 male candidates and 10 female candidates, from the most common ethnic/language groups represented in the test: Middle East, Spanish, E. European, S. Asian (Indian sub-continent), S.E. Asia and Africa.

Candidates were selected with a variety of raw scores, covering the range of scores (averaged across rating categories) from 3.0 (= clear fail) to 6.0 (maximum score), with most concentrated in the range 3.5 - 5.0, close to the likely standard (nominally 4.0).

A total of 8 different tasks were used by the 20 candidates. It was not feasible to select equal samples of each task, so task could not be examined as a facet.

10 ESL raters were selected from the pool of trained and experienced raters.

10 medical practitioners were contacted: their qualifications for inclusion in the study were either that they had experience of working in a supervising/teaching role with overseas-trained doctors in clinical bridging programmes, or that they were examiners for the test of clinical competence, or, in some cases, both. Two of the 10 were examiners selected to represent the Australian Medical Council: of these, one had experience teaching in bridging programmes, the other was the Chairman of the Examiners' Committee.

The 10 ESL raters were not given any special preparation, since they had all received formal rater training and had recently been involved in rating test administrations. They followed the operational procedure normally used in test administrations: each candidate is rated three times for each of six rating categories (Figure 2), using a 6-point scale, for the first roleplay, for the second roleplay and a final, definitive assessment. This final set of 6 scores is the one used in the test for reporting candidate performance.

The 10 doctors were given a short briefing (30-45 minutes), either individually or in small groups. For all of them, the purpose of the study was explained, including reference to the widespread perception, generally (though not universally) voiced or echoed by themselves, that the standard of English required to pass the test was too low.

Candidates:

- Audio recordings from recent test administrations (N = 20)
- male = 10, female = 10

Judges:

- Doctors (N = 9) Task 1 only rated for each candidate
- ESL raters (N = 10) Tasks 1 & 2 rated for each candidate

Rating categories used:

- Doctors 1 category only: Overall Communicative Effect
- ESL raters 6 categories including Overall Communicative Effect

Table 1: Data from Occupational English Test

Because of the constraints upon the time the doctors were able to devote to this study, it was decided that they would make a single rating only of each candidate, based upon the first of the two test tasks only. This rating was the same as the global rating category used operationally by the ESL-trained raters, i.e. 'Overall Communicative Effect'. This category has been shown (McNamara 1990) to be by far the best predictor of the final score awarded to candidates, representing in effect a summary of the other judgements made during the assessment.

The 6-point scale was presented to them in similar terms to those used in training sessions for the ESL-trained raters. As with the ESL raters in their training sessions, the doctors were instructed to base their judgements on their perceptions of how well they considered the candidate would cope with the demands of a supervised clinical

setting in a bridging programme. A recording of an interaction (additional to the 20 tapes used in the study) was played to each participant during this session, to clarify any questions about the rating task. Owing to pressure of work, one of the doctors was unable to complete the task, and the study is only able to report on data from 9 doctors.

Data Analysis

Two approaches to analysis were used, Generalizability-Theory (G-theory) and multi-faceted Rasch measurement.

Generalizability theory

The Generalizability-study design used in our research consisted of a random effects model with one facet: Judges. Our universe of admissible observations - that is, the conditions of the measurement procedure that we willing to consider acceptable - consisted of doctors that either had experience working in a supervising/teaching role with overseas-trained doctors in clinical bridging programmes, or as examiners for the test of clinical competence, or both. There were nine conditions, or nine doctors, for the judge facet. This facet was considered random in that the nine Judges were considered interchangeable with any other set of nine Judges within the theoretical framework of G-theory. All analyses were done with the GENOVA program, version 2.2, for the Macintosh computer (Crick & Brennan 1984). In planning our Decision-studies, we were interested in generalizing to different numbers of raters. We were primarily interested in absolute decisions rather than relative decisions (see Hudson & Lynch 1984; Shavelson & Webb 1991); that is, we wanted information that would help us make decisions about the standing of test candidates in relation to a standard of performance. We were also restricted to the use of 19 candidates in the GENOVA analysis, due to the absence of data for one of the candidates (and the requirements of GENOVA for balanced designs).

Multi-faceted Rasch measurement

Multi-faceted Rasch measurement (Linacre, 1989), implemented through the computer program FACETS (Linacre and Wright, 1992), relates the chances of success on a performance task to a number of

aspects of the performance setting. All of the terms in the equation are estimated as probabilities, expressed mathematically in units of equal interval called logits. In addition, FACETS offers a feature known as bias analysis which identifies specific interactions between elements of facets which deviate from the overall pattern of analysis.

In the first analysis, using the ratings provided by the doctors, two facets were taken into account, judge severity and candidate ability, using a rating scale with 6 points, on the single item, overall communicative effect.

A similar analysis was then carried out using the combined data set of the judgements of the doctors and the ESL raters on the single item, Overall Communicative Effect.

3. Results

GENOVA Results

Using the variance components from the D-Study with the same sample sizes as our original dataset (19 persons, nine Judges) we see that the greatest percentage of the total variance - 95 per cent - is attributable to persons, or universe score variance (see Table 2). This tells us that, in this assessment context, a relatively high proportion of our test score variance can be dependably associated with our test takers' ability in speaking ESL. Comparing this variance component to those for the effects of the facet Judges, we see that there is a small effect (1.5 per cent of the total variability) for the Judges on test score variability. A somewhat larger variance component is attributable to the persons by Judge interaction. This variance component includes random error, which because of the one-facet design is confounded with the interaction effect. We cannot sort out what is variance due to particular combinations of persons and Judges and what is random variability in the judgments. Still, the percentage of information uniquely attributable to persons — the information we are interested — is quite high.

Effect	Variance Component	Standard Error	Percentage of Total Variance
persons (p)	.951	.313	94.5 %
Judges (J)	.015	.007	1.5 %
pJ, e	.040	.005	4.0 %
Total	1.006		

Table 2:D-study Variance Components

Given the relative lack of effect for Judges, what is the dependability of the scores observed in this assessment context? Table 3 gives the generalizability coefficients (G-coefficients) associated with the different numbers of Judges. The G-coefficients, parallel to reliability coefficients in classical test theory, represent the degree of accuracy with which we can generalize from the test taker's observed score to their universe score. They are calculated with an error term that reflects relative decisions, or how well the observed scores differentiate the test takers on the ability being measured. Table 3 also gives the corresponding F coefficients, or dependability estimates for absolute decisions, which are the type of decisions of greatest interest in this assessment context.

# of Judges	G-Coefficient	F
1	.724	.658
2	.840	.794
3	.887	.852
4	.913	.885
5	.929	.906
6	.940	.921
7	.948	.931
8	.954	.939
9	.959	.945

Table 3: Dependability Estimates for Different Numbers of Judges

Discussion of results

The value of the G-Theory analysis in standard setting is as a preliminary stage to suggest how many Judges (in this case Doctors) are really required to take part in the exercise in order to produce acceptable levels of reliability. This has important practical implications for later data collection for use in a subsequent FACETS analysis, given the logistical difficulties of involving large numbers of doctors in the process. Once an acceptable number of Judges has been determined, as a result of the G-study, ratings of a larger number of candidates' tapes can be collected from this smaller number of Doctors to increase the sample size and the Once an acceptable number of Judges has been determined, as a result of the G-study, ratings of a larger number of candidates' tapes can be collected from this smaller number of Doctors can be used to increase the sample size and the associated reliability.

We can see that as the number of Judges increases, the dependability of our decisions, or inferences made, using the observed scores on the OET also increases. However, we can also see that using only one or two Judges results in unacceptable levels of dependability, especially in terms of absolute decisions (F). If, in our context, we were willing to accept absolute decisions based on observed scores that were 85 per cent attributable to a person's universe score (unaffected by other sources of variability), then we would be able to use only three Raters. This would be parallel to saying we were willing to accept test results associated with .85 reliability.

Our D-studies also allow us to estimate the change in dependability with various cut-scores. In our study, the doctors were informed as a part of their training that a "4" on the OET rating scale represented the standard for passing candidates (i.e., judging their language ability acceptable for entry into the medical certification "bridging program"). The information in Table 4 indicates that raising the

Cut Score	Dependability
3.5	.904
4.0	.875
4.5	.896
5.0	.935

Table 4: Dependability Estimates at Different Cut-Scores (4 Judges)

standard from 4 to 4.5 or 5.0 (using the original sample size of four Judges) would increase the dependability of our inferences from the observed scores. Of course, it also indicates that lowering the standard to 3.5 would similarly increase the dependability.

FACETS Results

On the basis of the first analysis performed using facets, an ability value was derived for each candidate, as shown in Table 5. We see from this table that 13 of the 20 candidates were awarded a passing score, i.e. 4.0 or more by the doctors, corresponding to a logit value of 0.76. Only one high scoring candidate (4916, the 2nd highest) was misfitting, and then only slightly. It is not uncommon for scoring candidates with extreme or near extreme scores to be misfitting in Rasch analyses, and the candidates' pass/fail decision is not affected.

Obsvd(raw) Average	Calib Logit	Model Error	Infit MnSq	Infit Std	Candidate
5.6	**6.65**	0.71	1.5	1	**4914**
5.4	**6.17**	0.69	2.3	2	**4916**
5.3	**5.70**	0.68	0.8	0	**291**
5.0	**4.33**	0.67	0.3	-2	**46**
4.7	**3.04**	0.64	0.9	0	**126**
4.6	**2.64**	0.63	1.2	0	**311**
4.4	**2.25**	0.62	0.9	0	**91**
4.3	**1.87**	0.62	0.4	-1	**110**
4.2	**1.49**	0.61	0.8	0	**293**
4.0	**0.76**	0.59	1.2	0	**141**
4.0	**0.76**	0.59	0.9	0	**174**
4.0	**0.76**	0.59	0.9	0	**199**
4.0	**0.76**	0.59	0.6	0	**53**
3.6	**-0.49**	0.53	1.0	0	**129**
3.3	**-1.04**	0.51	0.7	0	**249**
3.2	**-1.29**	0.50	1.2	0	**104**
3.0	**-1.78**	0.49	0.5	-1	**114**
2.7	**-1.95**	0.54	0.5	-1	**176**
2.3	**-3.21**	0.49	1.4	0	**179**
1.9	**-4.24**	0.53	1.0	0	**66**
4.0	1.16	0.59	0.9	-0.	Mean(Count: 20)
1.0	2.95	0.07	0.5	1.	S.D.

RMSE 0.60 Adj S.D. 2.89 Separation 4.86 Reliability 0.96
Fixed (all same) chi-square: 482.76 d.f: 19 significance: .00
Random (normal) chi-square: 18.96 d.f: 18 significance: .39

Table 5: Doctors: Candidate Measurement Report

The doctors' characteristics as Judges were also examined (see Table 6). It was found that they exhibited significant differences in severity (separation index = 2.45, reliability = 0.86, with a standard deviation of 1.04 logits). However, they generally demonstrated good internal consistency, with only one misfitting judge (107).

Obsvd Average	Measure Logit	Model Error	Infit MnSq	Infit Std	Dr ID
3.4	1.84	0.37	0.6	-1	109
3.7	0.86	0.38	0.9	0	105
3.9	0.44	0.38	0.6	-1	103
3.9	0.30	0.38	1.9	2	107
4.0	0.17	0.38	0.5	-2	106
4.0	0.01	0.38	1.0	0	104
4.2	-0.44	0.39	0.9	0	101
4.5	-1.50	0.43	1.2	0	108
4.6	-1.68	0.43	0.8	0	102
4.0	0.00	0.39	0.9	-0.4	Mean (Count: 9)
0.4	1.04	0.02	0.4	1.2	S.D.

RMSE 0.39 Adj S.D. 0.96 Separation 2.45 Reliability 0.86
Fixed (all same) chi-square: 59.87 d.f.: 8 significance: .00
Random (normal) chi-square: 7.94 d.f.: 7 significance: .34

Table 6: Doctors: Judge Measurement Report

It was then necessary to compare these judgements with the ratings given by the ESL-trained raters. Table 7 shows that when the doctors' judgements are combined with those of the ESL raters, no candidates appear as misfitting.

Secondly, as shown in Table 8, when the ESL raters' judgements are added, the number of candidates with an average raw score exceeding the passing level of 4.0 falls from 13 to 11. Candidates 141 and 53 now have average scores below 4.0. There are also some changes in the ranking of candidates according to ability. These are mostly minor, but two worth noting are for candidate 199, who moves from a rank order of 11.5 (on the doctors' judgements), to 7 (on the combined judgements of ESL raters and doctors), with an increase in average score from a borderline 4.0 to a more comfortable 4.3; and candidate 91, moving in rank order from 7 to 10, accompanying a reduction in average score from 4.4 to 4.1. Neither of these

candidates would, however, appear to be reclassified as failing on this second analysis.

Obsvd Average	Calib Logit	Model Error	Infit MnSq	Infit Std	Candidate
5.8	8.24	0.60	1.4	0	4914
5.4	6.40	0.47	1.6	1	4916
5.2	5.31	0.47	0.7	0	291
4.8	3.82	0.45	0.8	0	46
4.6	3.23	0.44	1.3	1	126
4.4	2.28	0.43	1.0	0	311
4.3	2.10	0.43	1.2	0	199
4.3	2.10	0.43	0.8	0	293
4.2	1.72	0.43	0.6	-1	110
4.1	1.17	0.43	1.1	0	91
4.0	0.99	0.43	0.9	0	174
3.9	0.81	0.42	0.7	0	141
3.8	0.29	0.41	0.9	0	53
3.6	-0.37	0.40	0.8	0	129
3.3	-1.26	0.38	1.0	0	104
3.2	-1.40	0.37	0.7	-1	249
2.9	-1.89	0.38	0.7	-1	176
2.9	-2.08	0.36	0.8	0	114
2.5	-3.22	0.35	1.1	0	179
2.2	-3.96	0.35	1.0	0	66
Obsvd Average	Calib Logit	Model Error	Infit MnSq	Infit Std	Candidate
4.0	1.21	0.42	1.0	-0.	Mean (Count: 20)
0.9	3.08	0.05	0.3	0.	S.D.

RMSE 0.43 Adj S.D. 3.06 Separation 7.17 Reliability 0.98
Fixed (all same) chi-square: 992.05 d.f: 19 significance: .00
Random (normal) chi-square: 18.94 d.f.: 18 significance: .40

Table 7: ESL Raters and Doctors: Candidate Measurement Report

Analysis 1: Doctors only Candidate measures, highest to lowest:			Analysis 2: ESL raters/Doctors combined Candidate measures, highest to lowest:		
Obsvd(raw) Average	Calib Logit	Candidate	Obsvd Ave	Calib Logit	Candidate
5.6	6.65	4914	5.8	8.24	4914
5.4	6.17	4916	5.4	6.40	4916
5.3	5.70	291	5.2	5.31	291
5.0	4.33	46	4.8	3.82	46
4.7	3.04	126	4.6	3.23	126
4.6	2.64	311	4.4	2.28	311
4.4	2.25	**91**	4.3	2.10	**199**
4.3	1.87	110	4.3	2.10	293
4.2	1.49	293	4.2	1.72	110
4.0	0.76	**141***	4.1	1.17	**91**
4.0	0.76	174	4.0	0.99	174
4.0	0.76	**199**	3.9	0.81	**141***
4.0	0.76	**53***	3.8	0.29	**53***
3.6	-0.49	129	3.6	-0.37	129
3.3	-1.04	249	3.3	-1.26	104
3.2	-1.29	104	3.2	-1.40	249
3.0	-1.78	114	2.9	-1.89	176
2.7	-1.95	176	2.9	-2.08	114
2.3	-3.21	179	2.5	-3.22	179
1.9	-4.24	66	2.2	-3.96	66

* = candidates passed by doctors only but failed by combined group, before taking measurement error into account.

Table 8: Rank order of Candidates on 2 analyses

The report for the Judges (Table 9) shows very similar differences in severity to the first analysis (separation index = 2.39, reliability = 0.85), with a standard deviation of 1.05 logits. The same doctor (107) was shown to be misfitting.

Examination of the bias analysis (Table 10) offered by the FACETS program revealed that this doctor was responsible for three of the eight instances of bias with a Z-score exceeding 2.0; this combined with the evidence of misfit suggests that he is a possible candidate for exclusion from the standard-setting exercise. It is perhaps worth noting that this analysis also shows that of the eight instances of bias, the doctors were responsible for 6, and the ESL raters only 2.

Obsvd Average	Measure Model Logit	Error	Infit MnSq	Std	Num	Prof
3.4	1.92	0.38	0.9	0	109	DR
3.4	1.92	0.38	0.6	-1	46	ESL
3.7	1.03	0.39	1.1	0	29	ESL
3.7	0.88	0.39	1.2	0	105	DR
3.8	0.57	0.39	0.7	0	2	ESL
3.9	0.42	0.40	0.6	-1	103	DR
3.9	0.26	0.40	2.0	2	107	DR
4.0	0.10	0.40	0.6	-1	106	DR
4.0	0.10	0.40	0.5	-1	26	ESL
4.0	0.10	0.40	0.7	-1	48	ESL
4.0	-0.04	0.40	1.3	0	104	DR
4.0	-0.04	0.40	1.2	0	79	ESL
4.1	-0.22	0.40	0.6	-1	78	ESL
4.1	-0.38	0.41	0.6	-1	55	ESL
4.2	-0.55	0.41	0.9	0	101	DR
4.2	-0.55	0.41	0.9	0	11	ESL
4.5	-1.70	0.44	1.5	1	108	DR
4.6	-1.90	0.45	1.1	0	102	DR
4.6	-1.95	0.44	0.9	0	51	ESL

Obsvd Average	Measure Model Logit	Error	Infit MnSq	Std	Num	Rater
4.0	-0.00	0.40	0.9	-0.3	Mean (Count: 19)	
0.3	1.05	0.02	0.4	1.1	S.D.	

RMSE 0.41 Adj S.D. 0.97 Separation 2.39 Reliability 0.85
Fixed (all same) chi-square: 123.38 d.f: 18 significance: .00
Random (normal) chi-square: 17.88 d.f.: 17 significance: .40

Table 9: ESL Raters and Doctors: Judge Measurement Report

Obsvd Score	Exp. Score	Bias+ Model Logit	Error	Z-Score	Candi	logit	Num	Prof
5	6.0	5.44	2.12	2.6	4914	8.24	102	Dr
4	5.4	5.16	1.94	2.7	4916	6.40	**107**	**Dr**
4	5.1	4.20	1.94	2.2	126	3.23	51	ESL
3	4.3	4.14	1.59	2.6	199	2.10	104	Dr
2	3.5	3.68	1.50	2.4	129	-0.37	**107**	**Dr**
3	4.1	3.21	1.59	2.0	91	1.17	79	ESL
6	4.5	-4.22	2.01	-2.1	91	1.17	108	Dr
4	2.4	-4.45	1.94	-2.3	179	-3.22	**107**	**Dr**

Obsvd Score	Exp. Score	Bias+ Model Logit	Error	Z-Score	Candi	logit	Num	Rate
4.0	4.0	-0.00	1.92	0.0	Mean (Count: 378)			
1.1	1.0	1.66	0.39	0.9	S.D.			

Table 10: Doctors and ESL Raters: Bias Calibration Report

Discussion of results

Two particular points of similarity and difference between the two FACETS analyses deserve comment.

1. The doctors, contrary to all expectations, turn out to be more lenient than the ESL raters. It was not possible to use professional background as a facet, since no judge appeared under more than one condition. A t-test comparing the logit values attached to the Judges' individual harshness; showed the ESL raters to be harsher than the doctors, but the difference was not significant (p = .81), as Table 11 shows.

Unpaired t-Test X ₁: Profession Y ₁: Harshness (logits)

DF:	Unpaired t Value:	Prob. (2-tail):
17	.247	.8077

Group:	Count:	Mean:	Std. Dev.:	Std. Error:
ESL	10	.058	1.02	.322
Medical	9	-.068	1.198	.399

Table 11: t-Test Results: ESL Raters vs. Doctors

Altough this difference was not significant, it was clearly noticeable: two more candidates pass when the ESL raters are excluded from the analysis. Two of the three harshest Judges are ESL raters, while two of the three most lenient are doctors (Table 9).

2. The doctors generally interpret the scale consistently with the ESL raters: there is no additional misfit amongst Judges, and no candidates are misfitting, when the two groups are combined. They do, however, exhibit slightly more bias than do the ESL raters (as we saw in Table 10).

4. Conclusion

The question remains of how to use the information from this study to address the practical issue of determining a passing standard for future administrations of the test.

It is not possible to anchor the scale produced by the doctors, who used only a single category, to another analysis involving the normal operational procedure of rating each candidate three times for each of six categories, with the final set of assessments used for reporting purposes. A scale composed of intervals defined using only the relatively crude information provided by the doctors would

cause serious misfit if then used as the anchor for the finer grained information obtained from the more complex and, as proposed by the test developers, more valid ratings given by the ESL raters.

As shown by Lumley and McNamara (1993) Judges vary in both harshness and internal consistency over time. The measurement error associated with rater harshness also adds so much unpredictability to the process that a more reliable form of anchor is needed.

A further factor is that the doctors participating in this study will not be involved in rating future terst administrations.

In the present instance, the only constant, therefore, will be the ability of candidates as determined during this study. Further investigation of possible courses of action is therefore proposed, using the following procedure as a starting point. For future test administrations, this set of ratings could be incorporated with the live test results, providing a substitute for formal anchoring. The 20 candidates from this study would occur at varying points along the scale, as can be seen in the map in Table 12, and the cut-off can be determined in relation to them: it is clear that candidates 91, 174 and 141 occur closest to the desired standard. It will be evident from such a map which candidates are clearly a fail or a pass; closer inspection of logit values and a final decision concerning the error values attached to candidates in the borderline categories will have to be made at the time, and in the context of the other factors relevant to the test. This will be discussed again in the conclusion section. All future candidates can then be allocated a pass or fail score on the basis of the logit values for their ability, in relation to the 20 candidates from this study. In effect, this procedure will narrow the band of possible scores within which the standard may fall.

Measr	+Candidate ID	-Rater ID	Scale Steps
9			(6)
	4914		
8			
7			
	4916		
6			
	291		
5			5
4			
	46		
	126		
3			
	311		
2	199 293	AS PM	
	110		
	91		
1	174	KH SB	4 standard
	141	CA	
	53	AC AWJ	
0	*	* AG BJ JH JM LG	*
	129	CP LB	
		LLF PC	
-1			
	104 249		
		MK	
-2	114 176	JD PN	3
-3			
	179		
-4	66		(1)

Table 12: ESL Raters and Doctors: All Facet Summary.

Basically, from this study we have discovered various possibilities for incorporating technical analyses such as G-theory and multi-faceted Rasch measurement into the process of standard setting. The

specific value of these technical analyses are summarized in Table 13.

G-theory

- allows us to determine the dependability of the Judges for both relative and absolute decisions
- allows us to determine how many Judges we will need for different levels of dependability
- allows us to determine the relative dependability of Judges' decisions at different cut-scores

Multi-faceted Rasch

- allows us to determine the rating severity of individual Judges
- allows us to identify individual Judges who are "misfitting"
- allows us to identify particular combinations of Judges and persons which are problematic
- allows us to express the cut-score in terms of logits (to relate the ESL judgments to the professional Judges' scale)

Table 13: Value of Technical Analyses for Standard Setting

It should also be noted that in the process of conducting a standard setting investigation, information of relevance to test validity can sometimes surface. For example, several of the doctors in our Judges sample commented that they were not certain that the test task to which the candidates responded was sufficiently representative of the clinical situation in which they would ultimately be judged. This indicates that it would be useful in future test development to attempt to secure reactions of the relevant professional group to test tasks by having them actually rate the language sample from such a task.

Nevertheless, despite these advantages, there can be no purely technical solution to the problem of standard setting in this context. Although the study reveals the doctors and ESL raters to be applying similar standards, which might suggest that the cut score be left as it is, the question is more complex than it seems. There is an inevitable margin of error around the estimate of the logit ability associated with a score of 4 from the doctors. A confidence interval of 1.6 times the standard error of this estimate means that there is a range of logit values associated with an endorsement of a pass standard. Where in this range should the pass standard be set? Do we want to err on the side of certainty, in the interests of the patient (and as it happens of the local profession), or to give the benefit of the doubt in the interests of the overseas trained candidate? There is also uncertainty around the estimate of the candidate's ability: again, should we err on the side of caution and restriction of entry, or again consider the interests of the candidate? These essentially value-laden decisions remain intractably ethical and political; no amount of technical sophistication will remove the necessity for such decisions.

5. References

Bachman, L. F.; Lynch, B. K.; and Mason, M. 1993. *Investigating variability in tasks and rater judgments in a performance test of foreign language speaking.* Paper presented at the 15th Language Testing Research Colloquium, Cambridge, August 1993.

Brennan, R. L. 1983. *Elements of generalizability theory.* Iowa City, Iowa: The American College Testing Program.

Crick, J. E. & Brennan, R. L. 1984. *GENOVA: A general purpose analysis of variance system.* Version 2.2. Iowa City, IO: The American College Testing Program.

Hudson, T.D. & Lynch, B. K. 1984. A Criterion-referenced measurement approach to ESL achievement testing. *Language Testing*, 1(2), 171-201.

Ingram, D. E. 1984. *Report on the formal trialling of the Australian Second Language Proficiency Ratings (ASLPR).* Canberra: Australian Government Publishing Service.

Linacre, J. M. 1989. *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. & Wright, B. D. 1993. *A User's Guide to FACETS: Rasch- measurement computer program, version 2.62*. Chicago, IL: MESA Press.

Lumley, T. & T.F. McNamara 1993. *Rater characteristics and rater bias: implications for training*. Paper presented at the 15th Language Testing Research Colloquium, Cambridge, August 1993.

Lynch, B. K. & McNamara, T. F. 1994. *Using g-theory and multi-faceted Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants*. Paper presented at Language Testing Research Colloquium, Center for Applied Linguistics, Washington DC, March 5-7, 1994.

McNamara, T. F. 1989. *The development of an English as a second language speaking test for health professionals. Part Two of a report to the Council on Overseas Professional Qualifications on a consultancy to develop the Occupational English Test*. Parkville, Victoria: University of Melbourne, Department of Russian and Language Studies.

McNamara, T. F. 1990a. *Assessing the second language proficiency of health professionals*. PhD thesis, University of Melbourne.

McNamara, T. F. 1990b. Item Response Theory and the validation of an ESP test for professionals. *Language Testing*, 7(1), 52-76.

Powers, D. E. & Stansfield, C. W. 1985. Testing the oral English proficiency of foreign nursing graduates. *The ESP Journal*, 4(1), 21-36.

Shavelson, R. J. & Webb, N. M. 1991. *Generalizability theory: A primer*. Newbury Park, CA: Sage.