DOCUMENT RESUME

ED 364 066 FL 021 400

AUTHOR McNamara, T. F.; Lumley, Tom
TITLE The Effect of Interlocutor and Assessment Mode Variables in Offshore Assessments of Speaking Skills in Occupational Settings.
PUB DATE Aug 93
NOTE 12p.; Paper presented at the Annual Language Testing Research Colloquium (15th, Cambridge, England, and Arnhem, The Netherlands, August 2-8, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *English for Special Purposes; Foreign Countries; Language Skills; *Language Tests; Statistical Analysis; Tape Recordings; *Testing Problems; *Verbal Tests; *Vocational English (Second Language)
IDENTIFIERS Australia

ABSTRACT

This study investigated potential problems associated with one method of testing English-as-a-Second-Language (ESL) skills in work settings in foreign countries. The method, used by the Australian government to assess speech skills of individuals in locations outside Australia, involves having native English-speakers carry out a series of oral interactions with the examinee, that are taped and later evaluated by a small team of trained raters. The study reported here analyzed data from assessments of about 70 examinees who were administered an advanced ESL test for health professionals. Assessment of the recordings included multiple ratings of examinee performance, competence of the interlocutor, rapport established between examinees and interlocutor, and audibility of the interaction. The data were subjected to statistical analysis. Results suggest that interlocutor variability and audiotape quality do affect ratings, and that this method of analysis is useful for examining variables in the testing situation. (Author/MSE)

THE EFFECT OF INTERLOCUTOR AND ASSESSMENT MODE VARIABLES IN OFFSHORE ASSESSMENTS OF SPEAKING SKILLS IN OCCUPATIONAL SETTINGS[1]

T.F. McNamara and Tom Lumley
NLLIA Language Testing Centre, University of Melbourne

Address for correspondence: NLLIA Language Testing Centre, Department of Applied Linguistics and Language Studies, The University of Melbourne, Parkville, Victoria 3052, AUSTRALIA *Telephone*: #61 3 344 4207 *Fax*: #61 3 344 5163 *e-mail*: t.mac@unimelb.edu.au; tom_lumley@muwayf.unimelb.edu.au

Abstract

The increasing demand for performance assessment of speaking skills in second languages has led to logistic complications, for example, the delivery of tests in offshore locations. One solution to the problem has been to train native speaker interlocutors to carry out a series of oral interactions with the candidate, with assessment from audio recordings of the test session postponed and conducted centrally by a small team of trained raters. This technique is currently used in two large scale occupationally related ESP tests administered internationally on behalf of the Australian Government. But these procedures raise questions about the effect of such facets of the assessment situation as interlocutor variables and the quality of the audiotape recording. Recent developments in multi-faceted Rasch measurement (Linacre, 1989) have significantly broadened the possibilities for investigation of these issues.

The research presented in this paper investigates potential problems associated with the above approach to the offshore testing of speaking skills. Data from audiotape-based assessments of approximately 70 offshore candidates from two administrations of the Occupational English Test, an advanced level ESP test for health professionals, are considered. In addition to multiple ratings of candidate performance, each recording is rated for perceptions of the competence of the interlocutor, the rapport established between the candidate and the interlocutor, and the audibility of the interaction. These aspects of the assessment situation are treated as facets in a multi-faceted Rasch analysis of the data.

The results of the analysis reveal the effects of interlocutor variability and audiotape quality on ratings. The paper concludes with an evaluation of the overall feasibility of the procedure, and implications for test administration are considered. The study is also a further demonstration of the application of multi-faceted Rasch measurement in performance assessment settings.

Introduction

The increasing demand for performance assessment of speaking skills in second languages has led to logistic complications, for example, the delivery of tests in offshore locations. One solution has been to train native speaker interlocutors (who may or may not be trained ESL teachers) to carry out a series of oral interactions with the candidate, with assessment from audio recordings of the test session postponed and conducted centrally by a small team of trained raters. This is the solution used by the Occupational English Test (OET) (McNamara, 1990) and the **access:** test (Wigglesworth and O'Loughlin, 1993), two large-scale ESL tests for intending migrants to Australia administered internationally on behalf of the Australian Government. However, these procedures raise questions about the effect of such facets of the assessment situation as interlocutor variables and the quality of the audiotape recording. This paper will examine these issues in the context of the Occupational English Test.

[1]This paper was originally presented at the Language Testing Research Colloquium, University of Cambridge, August, 1993. The research was made possible by a grant from the National Languages and Literacy Institute of Australia.

Interlocutor and assessment mode variables in the Occupational English Test (OET)

The Occupational English Test (OET) (McNamara, 1990) is an ESP test for health professionals which was developed in its present format in 1987. It is administered in Australia and overseas to members of 11 different health professions (doctors, nurses, dentists, veterinary surgeons, dietitians and physiotherapists, among others) who have obtained their professional qualifications overseas and who wish (after being accepted as migrants or refugees) to practise in Australia. This study concerns itself with the role play based speaking sub-test of the OET, which uses materials specific to the profession of each candidate. This sub-test has three phases: an introductory interview, focusing on the candidate's professional status (which is not assessed), followed by two short role plays (each lasting approximately 5 minutes) in which the interlocutor adopts the role of a patient or client, or the relative of a patient or client, while the candidate assumes his/her professional role. An example of the stimulus materials used in the role plays is given in Figure 1. This is the card used by the role player; the candidate has a matching card with the relevant information.

Figure 1: Demonstration stimulus materials

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?

Assessment is carried out by raters who have participated in a training session followed by rating of a series of audio-taped recordings of speaking test interactions to establish their reliability. The assessment is either carried out live, during the test, by a trained rater acting as interlocutor, or later, by a trained rater using an audio-tape of the interaction. In either case the interaction is recorded. Where possible (in large Australian cities) the assessment is done live, with confirmation of the rating from tape; but in overseas centres, or in remote Australian centres, no assessment is made at the time of the interaction. Instead it is made from tape at a central location by pairs of trained raters.

In offshore or remote Australian settings, the training and competence of the interlocutor become issues. The training received by the interlocutors varies, but at least it should consist of watching a video-tape of a typical speaking test interaction, to allow interlocutors to familiarize themselves with appropriate procedures for conducting the test, and familiarization with the test administration kit. The issues of interlocutor competence and tape quality noted above have been raised in anecdotal complaints made by raters for the OET, who have expressed concern about possible disadvantage to candidates.

Concerns have focused in particular on two features of the audio recordings: 1) the audibility of the interaction due to misuse of or poor equipment, and 2) the competence of the interlocutor.

Table 1: Facets investigated in this study

1) Audibility 2) Facets of interlocutor competence: a) The general competence in conducting the test b) The specific competence in adopting the role of patient or client c) The rapport established between the participants 3) Rater severity

With regard to the competence of the interlocutor, three factors were identified as possibly playing a role. The first of these was the general competence of the interlocutor: his/her ability to conduct the various procedures required by the test seriously and in an appropriate manner. The second was the competence of the interlocutor in adopting realistically the role of patient or client during the simulated consultations. The third the emotional climate established between the participants, or *rapport*.

Four facets of audio-taped recordings are thus identified as potentially problematical in the assessment situation:

1. Audibility of recording
2. General competence of interlocutor
3. Specific competence of interlocutor in adopting appropriate role in role play
4. Ability of interlocutor to establish rapport with candidate

These facets were the subject of the present study, and were considered in relation to the way each interacted with the facet 'rater'.

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. These aspects, or facets, will include the ability of the candidate and the difficulty of the task, but also the characteristics of the rater and other characteristics of the context in which the performance is elicited and rated. These facets are related to each other as increasing or reducing the likelihood of a candidate of given ability getting a given score on a particular task. This is expressed in the following way (Figure 2):

Figure 2: Multi-faceted Rasch Measurement

Probability of a given score on a rating scale = B - D - J - K - O (etc)

where B = ability of candidate
D = difficulty of task
J = severity of judge
K = 'step' difficulty for the particular score point on the rating scale
O = other aspect (facet) of the assessment situation.

All of the terms in the equation are estimated as probabilities, expressed mathematically in units called logits.

The number of facets of potential interest is large, and research in the field at the moment is marked by a phase of exploration, in which various aspects of the assessment setting are being conceptualized and modelled using multi-faceted measurement. This research is motivated by two factors: a research motivation, to try to identify aspects of the assessment context which can be shown to significantly affect scores; and a practical motivation, to build in a compensation for those facets which can be shown to exert a significant influence on the chances of success in an examination. This paper is a contribution to this ongoing task.

In this paper, the four aspects of the assessment setting above are treated as facets of the assessment setting in the analysis. A number of analyses are reported.

An additional feature of multi-faceted measurement is its capacity to investigate interactions between elements of facets, that is, interactions between particular raters and particular conditions of each facet of interest. It is possible, for example, that only certain raters may be affected by interlocutor competence, or audibility, but not others, and that no overall or general pattern emerges across raters. In this case, instead of an across the board compensation, an appropriate strategy may be to give feedback to individual raters on these interactions, in the hope that this feedback will remove the unwanted interaction effect.

Method

Table 2: Data

Data:

Occupational English Test administrations, 1992;

Audio recordings of Speaking sub-test interactions

	N
tapes	70
raters	7

Each tape rated twice; questionnaire completed for each rating

Data was gathered using material collected during 1992 offshore test administrations. 70 audio recordings of speaking test interactions were each rated twice. Seven raters were involved. Each rater completed a questionnaire (see Appendix 1) for each tape, in which they were asked to evaluate the audibility of the tape, and three aspects of the interlocutor's performance. In completing the part of the questionnaire dealing with *audibility*, raters gave scores to each tape for audibility, on a five-point scale, with 1 so inaudible as to render the interaction impossible to assess, and 5 as perfectly audible without effort on the part of the listener. In effect, this produced a 4-point scale, points 2 to 5. The frequency with which each of these categories was perceived is reported in Table 3.

Table 3: Audibility of tapes

Degree of audibility (5 = perfectly audible, 2 = least audible)	Frequency	%	Recoded
5	86	61	perfect (2)
4	29	21	imperfect (1)
3	20	14	imperfect (1)
2	6	4	imperfect (1)

It is clear from Table 3 that the incidence of imperfectly audible tapes was unacceptably high.

A difficulty with the categorization using four levels of audibility was the low numbers of tapes in some categories of audibility (e.g. ratings from a maximum of 6 tapes for the least audible category).

The data were therefore recoded into dichotomous categories. Candidates' tapes were categorized as 'perfectly audible' if both raters agreed the tape was fully audible (rating point 5), or 'imperfectly audible' if either or both raters rated the tape as having problems (all the other rating points, without distinction).

As far as the *interlocutor variables* were concerned, raters gave scores for each taped interaction on three categories, using a 4-point scale (see Appendix 1, Questionnaire):
1) the general competence of the interlocutor in conducting the test
2) the specific competence of the interlocutor in adopting the role of patient or client
3) the rapport established between the two participants.

Table 4a: Frequency of tapes according to interlocutor's level of general competence (N=140)

Category	Frequency	%	Recoded
Very competent	75	54	2 very competent
Adequate	60	43	1 other
Insufficiently competent	3	2	1 other
Not competent	2	1	1 other

Table 4b: Frequency of tapes according to interlocutor's level of competence as a patient (N=141)

Category	Frequency	%	Recoded
Very competent	75	53	2 very competent
Adequate	60	43	1 other
Insufficiently competent	4	3	1 other
Not competent	2	1	1 other

Table 4c: Frequency of tapes according to interlocutor's level of rapport (N=140)

Degree of rapport (4 = good rapport, 1 = no rapport established)	Frequency	%	Recoded	
4	66	47	2	good rapport
3	68	49	1	other
2	4	3	1	other
1	2	1	1	other

For analysis, a similar issue presented itself as with the audibility, namely the small numbers of tapes in some categories (Tables 4a, 4b and 4c).

The data were therefore recoded for each facet, into dichotomous categories, with the highest level of competence in each case contrasted with all the other levels.

A number of Partial Credit model analyses were conducted using these dichotomies (see *Results* below)[2].

Results

Audibility

The first aspect to be considered is the effect of audibility of tapes upon ratings given by raters.

Table 5: Analysis using Partial Credit model, 2 categories of audibility

Table 5: Audibility Measurement Report,
Partial Credit model for 'Audibility' and 'Rater'

Obsvd Score	Obsvd Count	Obsvd Average	Measure Logit	Model Error	Infit MnSq	Infit Std	Audibility	
2153	522	4.1	-0.26	0.07	1.0	0	2	perfect
1299	330	3.9	0.26	0.10	0.9	-1	1	imperfect

Separation 2.90 Reliability of separation 0.89
Fixed (all same) chi-square: 18.80 d.f.: 1 significance: .00

[2]Raters did not agree in their perceptions of the audibility of tapes and the qualities of interlocutors. This meant that ratings of the same candidate under pairs of conditions (where the tape was 'perfectly audible', or where it was 'imperfectly audible', where the interlocutor was 'very competent' or not, etc) were obtained, thus satisfying the requirement of overlap in the design of the analysis. Similarly, because raters did not *always* disagree about their assessments of the facets under investigation, it was possible to get calibrations of the harshness of all raters under each of the conditions (e.g tape is 'perfectly audible', tape is 'imperfectly audible', etc), thus avoiding a confounding of the facet 'rater' with any of the conditions of the variables under examination. FACETS can cope with relatively large amounts of missing data in the data matrix to be analysed, provided that there is some overlap on each of the facets in question (cf Linacre, 1993).

The analysis (Table 5) reveals a significant effect of audibility: imperfectly audible tapes are rated more harshly than perfectly audible tapes; the reliability of the difference in severity associated with audibility is 0.89. There are small errors in the logit values for audibility categories (0.10 and 0.07). The effect, while significant, is not very large.

A bias analysis identified a single rater (rater 7) as biased, although not strongly so (z-score = 2.3, for imperfectly audible tapes). It appears unlikely that the bias of a single rater accounts for the whole effect; on this analysis, audibility seems to affect the raters as a *group*, although with occasionally stronger effects for individual raters.

Interlocutor variables

General competence

Table 7 contains the results of a Partial Credit analysis for this data set.

Table 7: General Competence Measurement Report, Partial Credit model for 'General Competence' and 'Rater'

Obsvd Score	Obsvd Count	Obsvd Average	Measure Model Logit	Error	Infit MnSq	Std	General Competence
1383	312	4.4	0.34	0.09	0.8	-2	2 very competent
770	210	3.7	-0.34	0.12	1.2	1	1 other

Separation 3.01 Reliability of separation 0.90
Fixed (all same) chi-square: 20.08 d.f.: 1 significance: .00

It shows a significant effect for competence (reliability of separation 0.90). Raters appear to compensate for what they perceive to be the relative incompetence of interlocutors. The bias analysis again suggests that the effect is general, and not restricted to any single rater. The effect is relatively large, about .7 of a score point in raw score terms.

Competence as patient

Table 8: Competence as patient Measurement Report, Partial Credit model for 'Competence as patient' and 'Rater'

Obsvd Score	Obsvd Count	Obsvd Average	Measure Model Logit	Error	Infit MnSq	Std	Competence as patient
1345	306	4.4	0.44	0.10	0.8	-2	2 very competent
808	216	3.7	-0.44	0.11	1.1	1	1 other

Separation 4.12 Reliability of separation 0.94
Fixed (all same) chi-square: 36.01 d.f.: 1 significance: .00

The Partial Credit analysis of the dichotomously coded data (Table 9.1) reveals a significant effect of the same kind as with the general perceived competence: candidates interacting with less competent interlocutors appear to be favoured by raters. The bias analysis suggests that this is a general trend, not restricted to any particular raters. The size of the effect is about the same as for 'General competence'.

Rapport

Table 9: Rapport Measurement Report, Partial Credit model for 'Rapport' and 'Rater'

Obsvd Score	Obsvd Count	Obsvd Average	Measure Logit	Model Error	Infit MnSq	Infit Std	Rapport
1175	258	4.6	0.52	0.10	0.9	0	2 good rapport
948	258	3.7	-0.52	0.10	1.0	0	1 other

Separation 4.91 Reliability of separation 0.96
Fixed (all same) chi-square: 50.17 d.f.: 1 significance: .00

The general pattern is again repeated here: Table 9 reveals an effect for this variable, with candidates being favoured by raters if they are interacting with an interlocutor who achieves poor rapport. The size of the effect is greatest for this one of the interlocutor variables (almost a full score point in raw score terms), and the reliability of the effect is high (0.96). The bias analysis found no instance of significant bias.

Discussion

Table 10: Summary of findings

Facet	Level	Ratings given
Audibility of tape	↑	↑
Interlocutor Competence (General)	↑	↓
Interlocutor Competence (as Patient)	↑	↓
Rapport established	↑	↓

Audibility

The general finding of an effect for audibility suggests that care needs to be taken to ensure that recordings are of the highest possible quality and that procedures should be set in train to reduce the number of tapes with audibility problems. Nearly 40% of tapes were perceived as having audibility problems of some sort; this is far too high.

Given that audibility problems cannot be entirely eliminated, then action needs to be taken to sensitize judges to this issue. Ideally, analyses resulting in candidate measures should include raters' perceptions of audibility as a facet, so that the effect can be neutralized in the resulting measures. Alternatively, candidates at borderline decision points should be monitored so that the possible effect of harshness resulting from audibility could be taken into account in making pass/fail decisions. The effect in fact although real is quite small, and it is not clear what numbers of borderline decisions would be changed if this issue were taken into account. To the extent that the effect is not general but particular to individual raters, feedback to individual raters through rater performance reports (Lunz and Stahl, 1992; Wigglesworth, 1993) could be attempted, with a follow-up study to ascertain whether this made a difference to their behaviour.

From a research point of view, the study needs to be repeated with a much larger sample of ratings. A much larger study, using clearly identified problem tapes, would be useful to establish whether or not the effect is true for individual raters or for raters as a whole.

Competence

In general, an effect was found for competence across the three aspects of interlocutor competence studied. The effect was larger than for audibility, and thus a real issue, as it would alter the likely outcome in borderline pass/fail cases. Large numbers of interactions were also involved. On the one hand, more rigorous interlocutor selection, training and monitoring would seem to be called for; but the proportion of interlocutors rated as less than *adequately* competent is in fact very small, so that simply training interlocutors may not necessarily provide a solution.

Curiously, the effect was in the reverse direction than for audibility: perceptions of problems with interlocutor competence led to higher ratings. There are a number of possible explanations for this finding. First, a perception of lack of competence on the part of the interlocutor may have been interpreted as raising an issue of fairness in the mind of the rater, who may then have made a sympathetic compensation to the candidate. Secondly, the effect of the lack of competence may have been that the interlocutor 'hogged' the interaction, giving the candidate too little time to speak, so that the evidence available to the rater may have been restricted, with the result that the rater gave the candidate the 'benefit of the doubt'. It is regularly the case that when 'comprehension' on the part of the candidate is rated in oral interviews, as in the OET, it is rated more leniently than any other aspect of the candidate's performance (McNamara, 1990). This is probably because the evidence for the level of comprehension is hard to interpret in the absence of very obvious difficulty, and the presumption is in favour of the candidate.

A third possible explanation involves a feature of the design of the study. The tapes used in this study were made of interactions both in Australia and overseas; the interlocutors in this study would thus have included many people who were also acting as raters, and who would have been qualified and experienced ESL teachers. In this, the interlocutors used were thus not properly representative of the overseas interlocutor group, who are usually not trained ESL teachers, although they are representative of tapes rated in the test as a whole; second ratings of all tapes, of interactions both in Australia and overseas, are rated from tape. The perceptions of competence may have been a result of the contrast in interlocutor style thus present. However, it is generally felt that training and experience as an ESL teacher makes a person expert at eliciting speech from a non-native speaker; a kind of scaffolding is provided, a supportive atmosphere is created, there is greater accommodation to the level of the non-native speaker, and so on; such skills are generally believed to facilitate the performance of the candidate in the test setting. But the effect found in this study goes in the reverse direction (greater interlocutor skill results in lower ratings for candidates). Is it that the skilled interlocutor, in eliciting the most representative sample of the candidate's speech, is thereby leading the candidate to expose more fully the potential shortcomings of his/her performance? This seems counterintuitive. Clearly, there is a case for further research on the effect of teachers and non-teachers as interlocutors (studied to some extent in the development and first trialling of the OET: McNamara, 1990). It is also worth noting that raters for the OET are given extensive training as raters, but little training as interlocutors, on the assumption that their teaching experience represents an extensive form of training in the required skills. This assumption may not of course be warranted, and the neglect of interlocutor training for raters who also act as interlocutors is a weakness of current procedures which should be remedied.

Conclusion

The greater richness of face-to-face interaction in the assessment if speaking brings with it its own difficulties; the candidate's score is clearly the outcome of an interaction of variables, only one of which is the candidate's ability. It is important that the extent of the influence of these other variables be understood, both for theoretical reasons as part of our ongoing attempt to adequately conceptualize the nature of performance assessment, and for practical reasons in ensuring fairness to candidates. In addressing these questions, the study demonstrates the potential of multi-faceted Rasch measurement in modelling features of the assessment context in performance assessments, allowing investigation of a range of variables of interest in that setting with an ease and precision that has not previously been possible. Applications of this measurement approach can have practical benefits, too, for rater training, and for ensuring fairness in those cases where ability measures are found to be significantly affected by the variables investigated in a study such as this.

References

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

Linacre, J.M. and B. Wright (1992) *Facets: Rasch Measurement Computer Program, version 2.6.*

Lunz, M.E. and J. A. Stahl (1992) *Judge Performance Reports: Media and Message.* Paper presented at American Educational Research Association, San Francisco, 1992.

McNamara, T.F. (1990) *Assessing the second language proficiency of health professionals.* Unpublished Ph.D. thesis, University of Melbourne.

Wigglesworth, G. (1993) Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing* 10,3.

Wigglesworth, G. and K. O'Loughlin (1993) *An investigation into the comparability of direct and semi-direct versions of an oral interaction test.* Paper presented at the 15th Language Testing Research Colloquium, Cambridge, August.

Appendix 1: Questionnaire completed by assessors of tapes from OET Speaking test

1. Audibility of the tape:

- clearly audible	❑	5
- most clearly audible; effort sometimes required	❑	4
- generally audible; effort required	❑	3
- partly audible; difficult to assess	❑	2
- inaudible / not recorded: assessment impossible	❑	1

2. Competence of the interlocutor

(a) in conducting the whole interaction, generally:

- very competent	❑	4
- adequate	❑	3
- insufficiently competent	❑	2
- not competent	❑	1

(b) in adopting the role of patient/client :

- very competent	❑	4
- adequate	❑	3
- insufficiently competent	❑	2
- not competent	❑	1

3. Rapport established between the interlocutor and the candidate:

Good rapport |——|——|——| no rapport established
4 3 2 1