A comparison of analytic and holistic scales in the context of a specific-purpose speaking test

NORIKO IWASHITA and ELISABETH GROVE – The University of Melbourne

ABSTRACT

This paper reports on a study which examined patterns of rating in the Occupational English Test (OET) for health professionals speaking test, a threshold requirement for professional accreditation of overseas-trained health professionals who wish to practise their profession in Australia. Based on score data gathered over an eight-year period, the study investigated whether in using a set of five analytical criteria and one holistic criterion, Overall Communicative Effectiveness, raters tended to favour one criterion more than another, thus advantaging candidates who performed well on certain criteria and penalising those who did not. The results indicated that one analytical criterion, 'Comprehension', was significantly out of line with the others, a finding which has implications for criterion selection in rating scale development.

Introduction

The present study examined the use of the rating scale and patterns of rating in the speaking component of the Occupational English Test (OET) for health professionals. Passing the OET, a proficiency test in four parts (Reading, Listening, Writing and Speaking), is a threshold requirement for the accreditation of overseas-trained health professionals of non English speaking background who wish to practise in Australia. The test is administered both within and outside Australia several times per year to approximately 500 candidates. In view of its high-stakes nature, a number of previous studies have investigated various aspects of the OET speaking sub-test (for example, McNamara 1990; Lumley, Lynch and McNamara 1994; Lumley and McNamara 1995; Brown and Lumley 1997; McNamara and Lumley 1997; Lumley 1998; Jacoby and McNamara 1999), including rater characteristics and rater behaviour; standard setting; context and content validation of the speaking tasks; interlocutor and assessment mode variables; and validation of the rating scales. However, the issue of how each analytic rating criterion functions in relation to the holistic criterion has not yet been researched.

In the speaking sub-test, candidates undertake two role-plays in which they take the part of the health professional in a simulated clinical interaction with a patient. The 'patient' is the interlocutor, who is also usually the rater accredited to act as assessor. The assessment of candidate performance employs a scale consisting of six criteria that make up the construct of language ability being tested: five single-feature analytic criteria, Grammar and Expression, Fluency, Intelligibility, Comprehension, and Appropriateness, and one global assessment criterion, Overall Communicative Effectiveness. The assessment of candidate ability thus draws on multiple sources of information about various features of performance, and the final result represents an average of the scores awarded on all six criteria. However, when multiple sources of information are used, there is concern that in the assessment of candidates' overall communicative ability (that is the global criterion), raters may rely more heavily on only one or two sources of information (such as comprehension, grammar/expression, et cetera) than on others, favouring candidates who perform well on these criteria and penalising those who perform well on others, and thus calling into question the fairness and validity of the assessment process.

Before proceeding, we should clarify the nature of analytic and holistic rating scales. While a holistic rating scales uses a single global numerical rating to assess test-taker performance, an analytic rating scale uses several subscales to assess different aspects of performance separately. Arriving at a judgment of candidate performance with an analytic scale involves considering several aspects of language separately, whereas a holistic scale examines a number of linguistic features at the same time. The advantages and disadvantages of both types of scale have been widely acknowledged in the literature (for example, Hamp-Lyons 1991; Hamp-Lyons and Kroll 1997; Milanovic, Saville and Shuhong 1996). It is generally recognised that holistic rating scales have the practical advantages over analytic scales of speed of marking and lower cost. However, analytical scales are often considered more useful, as the score awarded on each criterion provides diagnostic information on different aspects of learner performance (Carr 2000; Hamp-Lyons 1991). Analytic scales are also generally favoured on the grounds of reliability, in that they are more likely to result in consistency of ratings, because they allow raters to focus on fewer aspects of the language in assigning a score than do holistic scales, which attempt to provide a global measure of a range of performance features (for example, Brown and Bailey 1984; Hamp-Lyons 1991).

The main problems in the use of holistic rating concern validity: what a holistic score actually represents and whether certain aspects outweigh others as assessors form an overall judgment of test-taker performance. Previous research suggests that this is indeed the case. For instance, in validating a holistic scale used for the assessment of compositions by native speakers, Huot (1990) found that raters tended to focus on content and organisation, thus privileging these over other linguistic features. In a study of raters who used a holistic scale to assess second language speaking performance, Brown, Iwashita, McNamara and O'Hagan (2003) identified a range of performance features to which raters were oriented, finding that their comments reflected those aspects which were most salient to them rather than other features which were not well managed by test takers. Other studies have questioned the appropriateness of making holistic judgments of second language learners on grounds related to the learners themselves. For example, in a study of ESL writing, Hamp-Lyons (1991 and 1995) found that the performance of some second language writers could not be encapsulated in a single score (1991: 244) because such learners do not achieve command of all aspects of composition at a uniform rate. This view is supported by Carr (2000), who argues that holistic ratings are problematic in the assessment of non-native speaker writing because variations among non-native speakers' written performances are even greater than those among native speakers, due to the linguistic constraints faced by non-native speakers.

Since the OET speaking sub-test uses both analytic and holistic scales, we considered it important to find out how raters were using the scales, and in particular, whether all aspects of speaking performance contributed equally to the global assessment of Overall Communicative Effectiveness. In an earlier study validating the OET speaking sub-test, McNamara (1990) found Resources of Grammar and Expression to be the most 'difficult', that is the most harshly rated criterion, and Comprehension the 'easiest', the most leniently scored. Grammar and Expression was also identified as the strongest determinant of the score for Overall Communicative Effectiveness. Although the raters in this study had been instructed to consider the communicative aspects of both lexical and grammatical resources, they appeared to focus more on formal accuracy in assigning their score for this criterion. McNamara speculated that the tendency for the OET raters in his study to be more oriented to the linguistic features of structural and lexical accuracy might be attributable to their professional background as language teachers. However, some doubt has been cast on these claims about rater orientation in the recent study by Brown et al (2003), who found that the particular aspects of speaking performance which were most salient for raters were not structural accuracy, but production features such as intelligibility and fluency. In view of the fact that ten years have passed since McNamara's initial validation study, it is important to investigate whether the pattern of relationship between global and single-feature assessments observed in 1989 is supported by further analysis. Based on the review of the literature, the present study addresses the following research question: What is the relationship of the single-feature criteria to the global assessment of speaking performance?

The data

The data consisted of speaking ratings conducted from 1994 to 2002, 14 782 performances (approximately 7400 candidates) in total, most of whom had been rated twice. A total of 70 raters assessed the OET speaking test during this period. However, disparities in the number of ratings carried out by individual assessors reduced to 29 the number of raters whose assessments could be subjected to statistical analysis. Ultimately, 13 488 assessments, consisting of assessments by 29 raters of a total of 7347 candidates, were analysed. Table 1 shows the number of candidates for each year.

Table 1: The number of candidates in each year

Year	Candidates
1994	701
1995	974
1996	1390
1997	1207
1998	681
1999	477
2000	602
2001	714
2002	601
Total	**7347**

Results

All data were submitted to analysis by means of the many-faceted Rasch model program, Facets (Linacre 1990). In this statistical model, each assessor's rating is seen as a function of the interaction of several factors: the ability of the candidate, the difficulty of the item (criterion), the characteristics of individual raters, and other background variables (for example, candidates' first language, their occupation, the assessment mode, and so on). Because it can estimate the interaction among these various factors (facets), the model allows investigation of their impact on test scores, thus bringing all facets together into a single relationship expressed in terms of the effect they are likely to have on a candidate's chance of getting a particular score. In contrast, raw scores (those actually awarded by raters, on a scale of 1 to 6, where 1 is the lowest and 6 the highest possible score, for each assessment criterion) cannot provide information on the impact of particular aspects of the assessment process. Therefore, in the many-faceted Rasch output, all figures are shown as logit values (a logit being a unit of measurement) rather than as raw scores.

Table 2 shows the difficulty of each rating criterion (item) in logits.[1] The results of the analysis show the criterion, Resources of Grammar and Expression, to be the most difficult on which to obtain a high score. In contrast, Comprehension was the easiest on which to receive a high score. The difference in logit scores between these two criteria is very large, amounting to more than two logit units. These findings indicate that candidates are likely to have received a considerably higher score for Comprehension than for Grammar and Expression. However, the global criterion, Overall Communicative Effectiveness, and the single-feature criteria, Intelligibility and Fluency, are all more or less similar in level of difficulty (0.42 and 0.3 logits respectively), while it is relatively easy to receive a high score for Appropriateness.

Table 2: Item difficulty of each rating criterion (in logits)

Assessment criteria	Item difficulty	Error	Fit
Grammar/expression	1.13	.02	1
Fluency	0.56	.02	1
Overall	0.42	.02	0.7
Intelligibility	0.3	.02	1
Appropriateness	-0.4	.02	1
Comprehension	-2.01	.02	1.4

The third column in Table 2 shows the 'fit' of each criterion. Fit concerns the extent to which the general pattern of candidate performance is an acceptable basis for estimating the difficulty of each item (criterion). A common rule of thumb for the acceptable range of 'fit' values is between 0.7 or 0.8 and 1.2 or 1.3 (McNamara 1996; Linacre 1992 and 1999).[2] Any item with a value above 1.3 is regarded as 'misfitting'. Misfitting items are those where an individual's performance on one item could not be predicted from that person's performance on other items. For example, if a candidate received high scores on four of the five criteria but a very low score on the fifth criterion, the fifth criterion would not function well as a measure of the candidate's ability. Another important example of the detrimental effect of a misfitting criterion arises in the situation where, if most candidates receive a very high score on one criterion, that criterion offers no useful measure of each candidate's ability. As shown in Table 2, the fit value of the Comprehension criterion is 1.4. This figure indicates that Comprehension does not accurately estimate candidate ability on the speaking test. According to the raw data, candidates who received a high score on Comprehension did not always receive high scores on other criteria.

An item with a value below 0.7 is an 'overfitting' item, one which is generally considered redundant. That is, an overfitting item yields no additional information to that already provided by assessments on the other criteria: candidate performance on this item is too predictable from the overall pattern of performance on the other items. 'Overfit' may also signal items that have an inbuilt dependency on other items. For example, whether or not a candidate is awarded a score of 6 on criterion A depends on the candidate getting 6 on criterion B. Criterion A is thus overfitting because it does not make an independent contribution to the estimate of the candidate's ability. In our OET speaking data set, the global assessment criterion, Overall Communicative Effectiveness has a fit value of 0.7, and is therefore overfitting. However, this overlap with the assessments awarded on the analytic criteria is to be expected, given that Overall Communicative Effectiveness is a global assessment of a range of features of candidate ability which have already been assessed on the single-feature criteria. Nevertheless, it is important to ensure that no one single-feature criterion weighs more heavily on the global assessment criterion than any other. We therefore conducted a further analysis, Multiple Regression, to examine how well candidates' scores on Overall Communicative Effectiveness could be predicted from the scores on some of the single-feature criteria. The results are shown in Table 3.

Table 3: Results of regression analysis

Single-feature criteria	ß	t	p
Intelligibility	0.19	32.90	0.001
Fluency	0.29	45.35	0.001
Comprehension	0.11	20.32	0.001
Appropriateness	0.18	26.95	0.001
Grammar/Expression	0.24	34.98	0.001

The figures in the second column ß ('Coefficients Beta') in Table 3 provide information on how much each variable (that is each single-feature criterion) contributes to the global assessment criterion, Overall Communicative Effectiveness. The higher the value, the more the influence of the particular criterion on the global assessment score. The Coefficient Beta figure (ß) for Fluency is the largest and for Comprehension, the smallest, but the difference between the two is less than 0.2, indicating that all single-feature criteria contribute significantly to the overall communicative ability criterion.

To summarise our findings on the relationship of the global assessment criterion to single-feature criteria, Table 2 shows that the degrees of difficulty of each single-feature criterion vary from one criterion to another. We predicted originally that one or two particular single-feature criteria might have a stronger impact on the global score than the others. However, the results of the multiple regression analysis indicate that while there are slight differences in the weight contribution of each single-feature criterion to the global assessment score, these differences are not great enough to warrant concern.

Discussion

The present study investigated aspects of ratings of the OET speaking test in terms of rating scale. In one important respect, our findings resemble those of McNamara (1990) on the relative level of difficulty of two of the single-feature criteria, Resources of Grammar and Expression, and Comprehension. As indicated in Table 2, Grammar and Expression is the criterion on which it is hardest for candidates to receive a high score, while Comprehension is by far the easiest, and the difference between them amounts to more than 2.0 logits. This variance is so substantial as to raise questions about the usefulness of Comprehension as a measure of candidate ability. That raters assess this category so much more leniently than the others probably arises from the fact that it can only be assessed indirectly. Comprehension is notoriously hard to define and, because it is not an observable aspect of speaking, is extremely difficult to judge. This criterion is unlike the other criteria which can be directly observed. In using it, therefore, raters must rely on inference, and, so long as there is no obvious communication breakdown, tend to give candidates the benefit of the doubt by awarding a high score. As the data revealed, very few assessors ever assign a score lower than 4 on Comprehension, even to candidates whose performance they judge to be weak on all other criteria.

Despite this similarity, our findings are somewhat different from McNamara's (1990). In particular, no single criterion was found to contribute to the global assessment criterion more significantly than any other. Even though Grammar and Expression was the most difficult, it was not a significant determinant of the score on Overall Communicative Effectiveness. Furthermore, whereas Fluency was the second/third easiest item in McNamara (1990), our study found Fluency to be the second hardest category. These disparate findings may arise from different characteristics in the populations of test-takers considered in both studies, or may indicate that the orientation of raters has shifted over the years, so that grammatical and lexical accuracy is no longer treated as the most important aspect of speaking performance. If this is the case, it may be largely due to the increasing dominance of communicative models of language teaching and assessment in the course of the 15 years which separate the two studies. But direct comparison is difficult, since the data sets in both studies differ greatly in size (that is in McNamara, only data from two test administrations in 1988 and 1989 were investigated, whereas in the present study, data from multiple test administrations from 1993 to 2002 were examined). The numbers of raters involved in the two studies were also different: ours included the ratings of 29 assessors while McNamara included only four.

The fact that Resources of Grammar and Expression was the hardest criterion is also problematic. Considering the relatively advanced proficiency of OET speaking candidates (that is the majority were awarded overall scores above 4 out of a possible maximum of 6), we would not necessarily expect this to be the most difficult aspect of performing the test task. Indeed, a recent study involving detailed analysis of test-taker speaking performance (Iwashita, Brown, McNamara and O'Hagan 2003) found that grammatical accuracy and complexity did not vary in relation to other features of performance across a range of proficiency levels. In other words, the second language learners in their study evidenced no greater difficulty with grammar than with any other aspect of speaking. In our study, it may be that the raters have assessed the Grammar and Expression criterion more harshly than the other criteria for reasons unrelated to test-takers' actual proficiency.

The substantial differences in item difficulty of each criterion also raise questions about the usefulness of analytic scales. As previously mentioned, although analytic scales are often favoured because they are assumed to provide diagnostic information on the strengths and weaknesses of test-taker performance, it is possible that a low score on a particular criterion may not reflect a weakness in the candidate so much as reveal the raters' differential perceptions of each criterion. Further investigation via in-depth analysis of test discourse will be necessary in order to establish whether some criteria were assessed more severely than others. Large differences in degree of item difficulty are of particular concern, since the scores of the single-feature criteria are combined to arrive at a final score for each candidate. For example, the total scores of a hypothetical candidate A, who receives a score of 6 on Comprehension, but 4s on Grammar and Expression, Intelligibility, Fluency and Appropriateness, would be identical to those of another candidate B, awarded a score of 6 on Grammar and Expression but 4s on the rest (that is 22 out of a possible maximum of 30). However, in view of the comparative difficulty of receiving a score of 6 on Grammar and Expression, the quality of the two candidates' performance would have been substantially different: B would be more proficient than A, a difference masked by the high score for Comprehension. Although the comparison is hypothetical, a score of 6 on Comprehension is much easier to obtain than the same score on Grammar and Expression. Scores of 6 on these two criteria therefore indicate significantly different levels of candidate ability. As shown in the fit statistics (that is, column 3 in Table 2), Comprehension was found to be 'misfitting', indicating that this criterion does not provide an accurate estimate of candidate ability. It was therefore recommended that Comprehension be eliminated from the range of analytic criteria used to assess the OET speaking subtest (Grove 2002).

Conclusions

The present study has investigated ratings of the OET speaking test by examining how each rating criterion functioned in relation to the global assessment of test-taker performance. There were considerable differences in the relative harshness and leniency with which raters assessed different aspects of performance. In general, candidates were awarded a higher score on the Comprehension criterion and a lower score on Grammar and Expression, and Comprehension was a poor predictor of overall candidate performance. These findings contrast significantly with those of the original OET validation study carried out by McNamara (1990).

While these findings are valuable in raising questions about the construct of speaking ability measured by the test, the study has several limitations which should be acknowledged. Considering the wide variety of candidate background factors (for example, profession and language) and the two types of assessment mode used in the test (that is, tape versus live), it is possible that score differences among criteria may be attributable to these factors and to the interactions among them. More sophisticated analyses (such as bias analysis; see Wigglesworth 1993 and 1994) would be required in order to examine whether each criterion is assessed differently according to the candidate background and assessment mode; how these separate factors might have interacted in the assessment of candidates; and whether any specific factor had a particular impact on scores. However, the entire data set could not be subjected to such forms of analysis. Not only did the number of ratings by each rater vary substantially, but there was also a large variation in the numbers of candidates in each language and professional group.

Despite these limitations, the study has several implications for assessment in general, from assessment of proficiency to assessment of second language learning in the classroom. In particular, it has revealed the inadequacy and indeterminacy of the Comprehension criterion and been instrumental in its removal from the OET speaking assessment. The examination of rating patterns using both analytic and holistic scales suggests that the variable severity with which assessors judge different aspects of learner performance may mean that overall scores do not accurately reflect candidate ability. These findings also raise the possibility that analytic rating may be overrated, and that using a single holistic criterion may prove a more accurate and efficient gauge of proficiency. However, in order to justify replacing the current combined analytic-holistic assessment scale with a single global assessment, extensive further analysis would have to be undertaken. Such a major change does not seem to us warranted by the findings of this study.

ACKNOWLEDGMENTS

The authors thank Language Australia Ltd for funding the research study, and, in particular, Ann Latchford, Manager OET, for her keen support of the project.

NOTES

1 In the logit scale, the average difficulty of each item is set at zero. The higher the logit score, the harder for candidates to get a higher score on that criterion; the lower the logit score, the easier to obtain a high score on a particular criterion. In other words, items with negative signs are easier than the average, and those with positive signs are more difficult than the average.

2 The range of acceptability (.7 to 1.3) used for this study is conservative compared with that proposed by Myford and Wolfe (2000), who suggest a range of .5 to 1.5. However, for a high stakes test such as the OET, we consider it appropriate to adopt the more conservative range.

REFERENCES

Brown, A., Iwashita, N., McNamara, T., & O'Hagan, S. (2003). *Getting the balance right: Criteria in the integrated speaking test.* Paper presented at the Annual Language Testing Research Colloquium, Hong Kong Polytech University, Hong Kong.

Brown, A., & Lumley, T. (1997). Interviewer variability in specific-purpose language performance tests. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 137–150). Jyväskylä: University of Jyväskylä.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning*, 34(1), 21–42.

Carr, N. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition texts. *Issues in Applied Linguistics*, 11(2), 207–241.

Grove, E. (2002). *Development of level descriptors for the OET speaking test.* Unpublished report submitted to Language Australia. The University of Melbourne.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second-language writing in academic contexts* (pp. 241–276). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1995). Rating non-native writing: The trouble with holistic scoring. *TESOL Quarterly*, 29(4), 759–762.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 – Writing: Composition, community and assessment*. TOEFL Monograph Series MS-5. Princeton: Educational Testing Service.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, 41, 201–213.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2003). *Analysis of test-taker discourse in the development of speaking scale*. Paper presented at the Annual American Applied Linguistics Conference (March 22–25). Sheraton National Hotel, Arlington Virginia, USA.

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes* 18(3), 213–241.

Linacre, J. M. (1990). *FACETS: Computer program for many facetted Rasch measurement*. Chicago, IL: Mesa Press.

Linacre, J. M. (1992). *A user's guide to facets*. Chicago, IL: Mesa Press.

Linacre, J. M. (1999). How much is enough? *Rasch Measurement Transactions*, 12, 653.

Lumley, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes*, 17(4), 347–367.

Lumley, T., Lynch, B., & McNamara, T. F. (1994). A new approach to standard-setting in language assessment. *Melbourne Papers in Language Testing*, 3(2), 19–39.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54–71.

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, 7(1), 52–75.

McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in offshore assessments of speaking skills in occupational settings. *Language Testing*, 14, 140–156.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment*. Studies in Language Testing 3 (pp. 92–114). Cambridge: University of Cambridge Press.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3), 305–335.

Wigglesworth, G. (1994). The patterns of rater behaviour in the assessment of an oral interaction test. *Australian Review of Applied Linguistics*, 17(2), 77–103.