Assessor training in a high-stakes test of speaking: The Hong Kong English language benchmarking initiative

David Coniam, The Chinese University of Hong Kong
Peter Falvey, The University of Hong Kong

Abstract

This paper focuses on the importance of assessor training in producing reliable grades in a high-stakes criterion-referenced test. The test, designed for teachers of English, is a direct performance-based test of speaking skills. The Speaking Test is one of a battery of assessment tools used in the implementation of English language benchmarks for teachers of English in Hong Kong. The paper provides a background to the Hong Kong Government's English language benchmark initiative, discusses current trends in language assessment, describes the Speaking Test, and investigates the effect of an orientation, training and standardisation programme for trainee assessors for the Speaking Test. It demonstrates – through the use of the multi-faceted Rasch measurement computer program *FACETS* - that rigorous training can decrease bias among assessors, narrow initial wide ranges in the award of grades and help reduce inconsistency.

Introduction

In 1995, the Hong Kong Government's Education Commission decided to investigate the establishment of language benchmarks for Chinese, English and Putonghua for all teachers in Hong Kong, not just language teachers (Coniam and Falvey, 1996). It charged the Advisory Committee on Teacher Education and Qualifications (ACTEQ) with investigating, establishing and implementing the benchmarks. ACTEQ subsequently accepted the recommendations of its English Language consultants that English Language Benchmarks be set by creating a battery of tests: involving a performance test of Classroom Language using criterion-referenced assessment methods with appropriate scales and descriptors, traditional pen and paper tests for Reading and Listening, and criterion-referenced tests for the assessment of Speaking and Writing.

The current paper has evolved from a series of validation studies of the different elements of the English Language Benchmark initiative (see, for example, the Speaking Test validation report (Coniam and Falvey, 1998) and the report detailing the training of assessors for the English Language Benchmark Classroom Language Assessment (Falvey and Coniam, 1999)).

The purpose of the current paper is to describe the process of training assessors for the Speaking Test in order to investigate levels of agreement between assessors and the robustness of the Speaking Test scales and descriptors as revealed in the grades awarded by assessors. The paper begins by describing current trends in assessment and the background to the English Language Benchmark Speaking Test. It then moves on to the major segment of the paper which describes the process of training assessors for the Speaking Test in terms of their orientation, standardisation and their application of the Speaking Test scales and descriptors and concludes with comments on the nature and rigour of assessor training.

Background to the English Language Benchmark Speaking Test

The background to the English Language Benchmark initiative and the rationale for the different test types is discussed further in Falvey and Coniam (1997). To be able to view the Speaking Test of the English language benchmark test in perspective, however, it is important to consider current trends in assessment and the rationale for adopting a competency-based, task-based approach to the design and implementation of the Speaking Test. The following section therefore presents a brief overview.

Current trends in assessment

In the past two decades, two trends have emerged in the area of assessment and evaluation. These are criterion-referenced assessment (often linked to task-based curriculum and assessment procedures in English language assessment) and competency-based assessment (often linked to vocational, and, increasingly, professional-based training and assessment). In discussing the latter, Brindley states:

> *Competency-based models of vocational education and training have in recent years dominated the educational landscape in Australia, the*

UK and New Zealand. They have also begun to exert a significant influence in the field of language learning (1995, p. 145-164)

Brindley (1995: 1-2) stresses the need for a theoretical approach to assessment and discusses the necessity for test developers to begin with a clear theoretical conceptualisation of the abilities they are assessing and to 'reality-test' their constructs against data from the target language use situation. McDowell (1995, p. 11-29), in describing the construction of a test for overseas-qualified teachers, begins with an overview of the Bachman (1990) model of language ability which provided a starting point for test development. Bachman, himself, stresses the need for data from a target-language use situation in order to construct valid assessment instruments.

The growth of competency-based models for use in the assessment of 'high stakes' professional competence is evident in assessments such as higher level medical examinations (e.g. for post initial qualification certification), and in the assessment of the competency of accountants, architects and estate agents, (Eraut, 1994).

However, the 1970s model of 'performance-based competency' with a behaviourist underpinning used, initially, for low-level vocational skills assessment is not the only type of competency assessment to be utilised. 'Descriptive Competencies' can be categorised as of the 'performance capability type' as evidenced in typical North American 1970s approaches to vocational assessment or, more recently, 'simply to describe any piece of knowledge or skill that might be construed as relevant' (Eraut, 1994: 179). The approach adopted in the Hong Kong benchmarking initiative has been of the latter, more modern variety.

McNamara (1996) provides a useful discussion of performance tests and, as we shall see below, describes procedures for ensuring that the measurement issues involved in performance assessment and the reliability of assessors are dealt with in as comprehensive a manner as possible.

Making the test fit the situation - issues of validity

While reference has been given above of countries which have some form of language ability benchmarking in place, the manner is which the *assessment* of teachers' language ability for English language teaching is conducted also needs to be taken into account. In this

context, the tests which appears to match closest the demands placed upon English language teachers are the Guam Educators' Test of English Proficiency, described by Stansfield et al. (1990) and the UCLES/RSA Cambridge Examination in English for Language Teachers (CEELT) (see Falvey and Andrews, 1994).

The Guam test assesses the four skills of listening, speaking, reading and writing, with some of the test content based around the subject area of language teaching (e.g., a composition error detection test assessed by multiple choice methods). In the UK, in addition to reading and listening tasks which use topics related to language teaching, the CEELT assessment mechanism contains *direct* tests of speaking and writing. The latter tests are based on tasks which teachers would be expected to carry out as part of their professional duties. While the approaches taken by Stansfield et al. (1990) and by UCLES/RSA for their CEELT examination are to be commended in that they relate assessment tasks to the teaching situation, there is still no examination of a teacher's *language ability* in the language classroom itself. Such assessments are considered essential because they may reveal strengths and weaknesses not in evidence in a formal test situation.

Elder (1994) reports on differences between teacher spoken performance in the classroom and teacher spoken performance with peers. This was one factor in deciding to introduce a Speaking Test into the Pilot Benchmark Assessment (English) (PBAE) as a complement to the benchmarking of teacher spoken language in the classroom. The Speaking Test provides an opportunity for teacher test takers to display their ability to talk about language with peers and talk to an interlocutor in a narrative or expository genre from a prompt arising from the teacher-focused reading aloud test. The Speaking Test also tests teachers' ability to read aloud, a teacher-specific skill which is not always displayed during a visit to a classroom, but a skill which the test developers and the Government felt should be assessed.

Performance tests

The advantages of criterion-referenced assessment in direct tests of teacher language ability have been described elsewhere (Coniam and Falvey, 1998; 2000). There are, of course, disadvantages. The major concern with criterion-referenced assessment, given that the

descriptors are both appropriate and applicable, particularly in a high stakes examination, is whether the grades awarded by the assessor are reliable; i.e., whether assessors are consistent in their application of the descriptors for each scale. The ways of overcoming such criticisms and making the award of grades reliable all involve thorough orientation, standardization and training of assessors. In terms of performance tests generally, the Speaking Test of the English language benchmark initiative would be described by McNamara (1996: 43-45) as a 'weak' version.

Assessor training

The importance of assessor training is any English language examination is an issue which has long been accepted as an essential factor in determining the reliability of a test (see e.g., Webb et al., 1990).

In the case of performance-based assessment, where judgements are even more subjective than in a simulated test environment when more variables are controlled for, see, for example, the discussion in McNamara (1996:97, 110-111) of the Australian English language test for health professionals, the *Occupational English Test.* It is important to attempt to ensure reliability through extensive assessor training and standardisation, including, as McNamara illustrates (1996: 111), sanctioning inconsistent assessors.

It has been argued that it is not possible to achieve very close levels of agreement between assessors (see Lunz and Stahl, 1990). While Lunz and Stahl may perhaps overstate the point, in that, given enough time, very close agreement is possible, these views must be taken into consideration by test developers. As an example, in a determined effort to create reliable assessor judgments, the training programme for the Classroom Language Assessment assessors in Hong Kong (another component of the English Language Benchmark initiative) lasted approximately 35 hours (Falvey and Coniam, 1998)

Webb et al. (1990) discuss the problems associated with assessor stringency, leniency and inconsistency. They state that problems with assessor stringency and leniency can be handled by statistical adjustment. However, they make it clear that assessor training is essential for other problems – specifically, assessor inconsistency. Lumley and McNamara (1995), in discussing inconsistency in

assessors report that training and standardisation is not only essential, but also that further moderation is required shortly before the administration of Speaking Tests because a time gap between the training and the assessment event reveals that inconsistencies re-emerge.

Wigglesworth (1993) explores the use of bias analysis (using multi-faceted Rasch measurement) as a tool for improving rater consistency in assessing oral interaction. She illustrates how assessor leniency or harshness were reduced after feedback to assessors on their performance.

In a recent article, Weigle (1998) illustrates, on a test of writing, how training affects assessor consistency scores. Using multi-faceted Rasch measurement, she illustrates through an analysis of assessors' scores calculated prior to and after training, how assessors became more self-consistent. She illustrates how the spread of scores, in terms of the degree of severity decreased after training. As will be shown below, this effect was also apparent in the current study, and hence the need for thorough assessor training is reinforced.

The current study follows Wigglesworth's (1993) and Weigle's (1998) practice of applying multi-faceted Rasch measurement, using the computer program FACETS (Linacre, 1994). A brief outline of the Rasch measurement model is given below.

The Rasch model

In multi-faceted Rasch analysis, as in the standard Rasch model, the aim is to obtain a unified metric for measurement not unlike measuring length using a ruler. The scale on a ruler should not change when measuring a variety of objects. The measurement scale derived by application of the Rasch model is based on the probability of occurrence of certain phenomena (item difficulty, student ability, different judge severity-leniency levels). Once a common metric is established for measuring different phenomena (in our case, features of the Speaking Test), the different features can be examined and their effects controlled. The result of using a Rasch model of measurement provides independence from situational features in a particular test, students, etc. In other words, the results can have a general meaning. Multi-faceted Rasch analysis is a Rasch-based approach where various situational factors are explicitly taken into consideration in

constructing measurement. The units of measurement in Rasch analysis are *logits*, which are centred at zero; this is the 50% probability represented by an "item" of average difficulty. (For an overview of multi-faceted Rasch analysis, the manner in which it may be conducted, and the results interpreted, the reader is referred to McNamara (1996) Chapters 5-8 for a discussion; also Wigglesworth, (1993: 307).

Principles, Procedures and Processes in Training and Standardising Assessors

As stated earlier, in any examination, but particularly in a high-stakes examination of direct performance that has high validity, it is important that attention is paid to issues of reliability. In assessments of performance which rely wholly on assessor applications of the criteria established for the assessment, reliability can be established through a process of:

- agreement on the validity of assessment constructs
- creation of detailed specifications
- creation of valid, detailed and operationalisable descriptors
- provision of credible assessor training and standardisation

The purpose of standardising assessors is to ensure that strong measures of reliability occur whenever a number of assessors apply grade descriptors to a criterion-referenced assessment instrument. This was the case with the English language benchmark Speaking Test. In criterion-referenced assessment, which depends on the application of assessors' judgements to the criteria described in the descriptors, it is important that two principles are adhered to:

- Judgements by one assessor over time with a number of test takers need to be consistent.
- Different assessors judging an individual test taker should provide assessments that are closely correlated.

There are a number of well-established standard procedures that can be used to train and standardise language assessors (see the procedures and training videos of the University of Cambridge Local Examination Syndicate for its teacher language examinations, as well as its general English oral examinations). These procedures were

applied in the specific training procedures used with trainee assessors for the Speaking Test.

Participants

Fourteen trainee assessors were drawn from a list of provisional assessors provided by the Hong Kong Examinations Authority (HKEA). The majority worked in the field of language teacher education; the remaining trainee assessors were experienced English language tertiary teachers with experience of assessing high-school students for the HKEA.

The test takers were drawn from applicants to the English major programme of the Postgraduate Certificate in Education at the University of Hong Kong.

Procedures for standardising trainee assessors using the performance of videoed test takers

For training and standardisation purposes, videotapes of Speaking Test test takers were used. They consisted of the language performances of videotaped test takers who had given written permission for the videos to be used for training purposes. In addition, simulations of a typical assessment procedure were carried out by the trainee assessors themselves in front of the whole group of trainee assessors in order to familiarise themselves both with standard administration procedures (e.g. standard spoken rubric) and the administration of the assessment instruments themselves.

The major procedures used in the orientation, training and standardisation of trainee assessors are detailed below:

1. awareness raising
2. exposure to Speaking Test test takers on video
3. ranking the videotaped test takers without preconceived criteria
4. comparison of rankings and the provision of justifications by the trainees for their rankings
5. extrapolating, discussing and setting-out criteria from the procedure used in points (2 & 3) above

6. exposure to and discussion of the Speaking Test assessment constructs, criteria and descriptors
7. application of the descriptors to the ranked videoed test takers (see steps 2 & 3 above)
8. comparison and discussion of the application of the descriptors
9. working towards a consensus and standardisation
10. repetition of steps (4) to (9) until consistency of assessments is achieved and the assessors have been standardised

An outline of the training programme which utilised the procedures above is given in Figure 1 below. Reflecting the high-stakes nature of this examination, a whole day was devoted to the assessor training exercise. This procedure can be contrasted with the Hong Kong Examination Authority's (HKEA) normal procedures for training oral assessors for the public Grade 12 oral examination (the *Use of English* Examination). Only two hours are allocated by the HKEA for the training of assessors for their public exam which is used for university entrance purposes.

Figure 1: Format of Training and Standardisation session

09.00	Introduction: Welcome Orientation to English language benchmarks Discussion of purpose of language benchmarks Introduction to different forms of ASSESSMENT The place of constructs in assessment How scales and descriptors are created in criterion-referenced assessment Logistics for Pilot Benchmark Assessment
09.45	Pre-training ranking exercise – whole group ranking of one set of videos
10.15	Discussion
10.30	Assessment 1 followed by discussion – whole group
11.30	Simulation 1 followed by discussion – whole group
1.30	Assessment 2 followed by discussion – two groups
2.15	Simulation 2 followed by discussion – two groups
3.15	Assessment 3 followed by discussion – two groups
4.10	Assessment 4 followed by discussion – whole group
4.30-5.00	Round-up

After the ranking exercise, trainee assessors observed four complete sets of videos of test takers who were teachers attending an in-service postgraduate course. Thus, the trainee assessors assessed a total of 12 teachers as each set of test takers comprised three in-service teachers.

The Speaking Test - Scales

The Speaking Test contains of three tasks. For each task, there are two scales, as laid out in Figure 2.

Figure 2: Speaking Test Tasks and Scales

Task	Task Type	Scales
1	*Reading aloud* (test takers assessed singly)	• Pronunciation, stress and intonation • Reading aloud with meaning
2	*Telling a story / recounting a personal experience / presenting arguments* (test takers assessed singly)	• Grammatical accuracy • Organisation and coherence
3	*Professional oral interaction* (test takers assessed in a group of three discussing an authentic student composition)	• Interacting with peers • Talking about language with peers

Method

First, the 14 trainee assessors were asked to rank a set of test takers and place them on a scale of most to least able. In discussion, trainee assessors were asked to provide reasons for their judgements.

The 14 assessors were then given the Speaking Test scales and descriptors, given time to read and make sense of them, shown the first set of three test takers and asked to rate them without discussion. The purpose of this 'blind' rating was to investigate how much initial variability occurred among the assessors. Subsequently, this variability would be compared with their performance on the final rating exercise at the end of the day. After the initial assessment of teachers in Set 1, the trainee assessors were then taken through three

more full sets of videotaped test takers. In addition, the trainee assessors also role-played being interlocutors and test takers during two simulation exercises in order to familiarise themselves with interlocutor talk and other procedures for the Speaking Test.

As stated above, Weigle (1998) – using multi-faceted Rasch measurement - describes a similar procedure for training and re-training assessors of written scripts. She reports on the importance of training and standardisation, and the degree to which raters may have assimilated a grading scheme through comparative analyses of early and subsequent marking.

The principles upon which multi-faceted Rasch analysis is based are mentioned above. In this study, three facets have been specified for analysis of data from the training programme:

- the 14 trainee assessors
- the test takers
- the six scales

To summarise, all things being equal (i.e. assessors, test taker ability and the six scales), all scales will centre around zero logits. In terms of assessor judgements, a logit score above zero (a *positive* measure) indicates harshness; a logit score below zero (a *negative* measure) indicates leniency. For the test takers, a positive logit value indicates higher language ability, while a score in negative logit values indicates a test taker with lower language ability.

The computer program FACETS provides a number of statistics which give an indication as to how well the data fits the model. One of these is the mean square statistic. For this statistic, acceptable practical limits of fit have been proposed as 0.5 for the lower limit and 1.5 for the upper limit (Lunz and Stahl, 1990; see also Weigle, 1998 for a discussion of this issue). Additionally, FACETS reports on bias between facets; that is whether particular assessors react in apparently unexpected ways to any assessment categories. FACETS produces many detailed analyses from a number of different perspectives. In this paper, the analyses that are germane to the discussion will be explained. (See McNamara, 1996, for a detailed explanation of the use of the program FACETS and the manner in which its results may be interpreted.)

Results and Discussion

Tables 1 and 2 below present analysis of the data produced by the 14 trainee assessors for Set 1 of the videos (which was assessed 'blind', i.e., before discussion of the scales and descriptors).

Table 1 presents an analysis of the 14 assessors in terms of model-data fit. The reader is referred to the infit mean square statistic, which indicates the extent to which the data fits the model. (Practical limits of fit proposed, as mentioned above, are 0.5 for the lower limit and 1.5 for the upper limit.)

Table 1: Assessors' Scores – Results of first session, conducted before training begins

Measure (logits)	Infit mean square	Model error	Assessors
+0.73	1.1	0.30	J
+0.46	1.2	0.30	K
+0.20	0.7	0.30	A
+0.12	1.0	0.30	D
+0.12	0.9	0.30	N
-0.07	0.8	0.31	L
-0.16	0.6	0.31	F
-0.26	0.9	0.32	I
-0.58	2.1	0.33	H
-0.79	1.2	0.33	C
-0.90	1.4	0.32	G
-1.11	1.3	0.31	E
-1.37	0.3	0.29	B
-1.45	1.2	0.29	M
-0.31	1.1	0.31	Mean
+0.65	0.4	0.01	SD

RMSE 0.31 Adj S.D. 0.58 Separation 1.90 Reliability 0.78
Fixed (all same) chi-square: 68.4 d.f.: 13 significance: .00

The logits range between the assessors is very wide. The most severe assessor, Assessor J has a measure value of 0.73 logits, and the most lenient assessor, Assessor M a measure value of –1.45 logits, a range 2.22 logits. This is not unexpected in pre-training analyses of assessor ratings. What is important is whether and how much this range narrows by the end of the training period.

From the infit mean square statistic, it can be seen that two assessors do not fit the model at this stage – Assessor H with a mean square of 2.1 (i.e., greater than 1.5) and Assessor B with a mean square of 0.3 (i.e., smaller than 0.5).

Table 2 below presents the bias report for assessors and scales. It should be noted that a score of ±2 indicates the assessor is unexpectedly lenient on a particular category (-2) or unexpectedly harsh on a particular category (+2). For the purposes of this analysis, however, discussion is based on all the examples of bias that FACETS reports.

Table 2: Bias/Interaction Calibration Report – Assessors and Scales - First session

Bias+ Logit	Model Error	Z-Score	Infit mean square	Assessor	Logit	Traits
1.61	0.67	+2.4	0.1	C	-0.81	Talking
1.29	0.70	+1.8	0.2	M	-1.47	Reading
1.28	0.66	+2.0	0.8	G	-0.92	Talking
-1.21	0.63	-1.9	0.7	C	-0.81	Reading
-1.26	0.63	-2.0	0.7	H	-0.59	Grammar
-1.59	1.31	-1.2	0.1	M	-1.47	Talking
-0.00	0.70	0.0	0.6	Mean		
0.63	0.08	0.9	0.6	SD		

Fixed (all = 0) chi-square: 55.3 d.f.: 70 significance: .90

KEY: Talking - Talking about language with peers
Reading - Reading aloud with meaning
Grammar - Grammatical accuracy

As can be seen from Table 2, there are 6 instances (10% overall) of possible bias reported, three of which occur with the scale *Talking about language with peers.*

Training and standardisation

As described above, after the first 'blind; rating session, the trainee assessors were then given detailed training for three more full sessions and standardisation feedback and follow-up after they had presented the grades they had awarded. Procedurally, it should be noted they were always asked to give grades without conferring, were not allowed to amend their grades, once given, and were

instructed to hand over their grading sheets after each exercise. Table 3 below presents results of the final session in a format similar to that of Table 1.

Table 3: Assessors' Measurement Report – Results of the final session – conducted after training

Measure (logits)	Infit mean square	Model error	Assessors
+0.58	1.0	0.20	C
+0.12	1.0	0.21	H
+0.06	1.9	0.19	D
-0.01	0.7	0.19	A
-0.15	0.6	0.19	F
-0.22	1.1	0.19	K
-0.22	0.6	0.19	J
-0.30	0.7	0.19	M
-0.36	0.7	0.19	G
-0.43	0.9	0.19	L
-0.50	1.2	0.19	E
-0.55	0.8	0.21	N
-0.65	1.1	0.19	I
-0.69	0.9	0.20	B
-0.24	0.9	0.19	Mean
+0.33	0.3	0.01	SD

RMSE 0.19 Adj S.D. 0.27 Separation 1.40 Reliability 0.66
Fixed (all same) chi-square: 39.4 d.f.: 13 significance: .00

As can be seen from Table 3, two significant improvements have occurred. First, the range of scores given has fallen from 2.22 to 1.27 logits. Table 2 shows that there is an unequal bias between lenient and severe assessors (from +0.73 logits at the severe end of the table to –1.45 logits at the lenient end). Table 3, however, in addition to decreasing the range of scores awarded (+0.58 logits at the severe end to –0.69 logits at the lenient end) also demonstrates that the balance of severe and lenient grades awarded is much more equal.

In terms of the two assessors who did not fit the model in Table 2, Assessor B is an interesting case of improvement. Before formal training, Assessor B was very lenient (-1.39 logits) and did not fit the model. After training, Assessor B, while still grading leniently, has become less lenient (-0.69 logits) and now fits the model.

It will also be noted in Table 3 that there is now only one assessor (Assessor D, with a mean square of 1.9) who does not fit the model in contrast to Table 2 where two assessors did not fit the model.

Table 4 presents the bias analysis for the final session in a format similar to that of Table 2.

Table 4: Bias/Interaction Calibration Report – Assessors and Scales - Final session)

Bias+ Logit	Model Error	Z-Score	Infit mean square	Assessor	Logit	Traits	Logit
1.12	0.69	1.6	2.6	C	0.58	Talking	0.19
1.06	0.49	2.2	0.6	E	-0.50	Reading	0.02
-1.26	0.66	-1.9	0.0	B	-0.69	Interacting	0.41
-0.00	0.48	-0.0	0.7	Mean			
0.40	0.04	0.8	0.7	SD			

Fixed (all = 0) chi-square: 50.3 d.f.: 70 significance: 1.00

```
KEY: Talking - Talking about language with peers
Reading - Reading aloud with meaning
Interacting - Interacting with peers
```

Table 4 indicates that the amount of bias detected has dropped substantially. Only 3 instances have now been recorded, and the problems assessors had with the scale "Reading Aloud with Meaning" have been resolved. The incidence of reported bias has therefore decreased from 10% to 4%.

Finally, the data for sessions 2, 3 and 4 were analysed together, since this provides a larger test taker data set for analysis – 9 test takers as opposed to three in each of the individual analyses. In these analyses, the number of assessors who were shown to misfit the model dropped from two to one. It is important that sufficient training be provided to reduce misfits to the minimum possible. This means that, even with a larger sample of test takers, where it might be expected that a greater number of misfits might occur, the misfits decreased significantly. Furthermore, in the assessor/scale bias report, there were no instances at all of bias reported. This indicates that the training has been successful, and that assessors are acceptably close and accurate in their judgements.

Conclusions

This paper has illustrated how the trainee assessors, appointed to implement the Speaking Test of the English language benchmark test in the first major pilot of the test, were orientated, prepared for their role as assessors and standardised. The paper has, in addition, illustrated the value of a whole day's training compared to the current two hours training provided for assessors of speaking tests in other important public examinations offered in Hong Kong.

In the training programme, trainee assessors were first asked to rank one set of three test takers in order to prepare assessors for the subsequent tasks which involve matching test takers to descriptors and awarding grades. They were then asked to rate another set 'blind', without having discussed the descriptors amongst themselves. The purpose of this exercise was to demonstrate that even with the descriptors in front of them it is possible for different trainee assessors to produce different grades. Subsequently, they were taken through a series of simulations as both test takers and assessors, and then, after a series of discussions, completed three full sets of assessment. It has been illustrated above how the initial wide range of marks was reduced, after training, to a much narrower range. Problems associated with misfit on the Rasch model were also greatly reduced and inconsistencies reduced.

Although it can be seen that the training and standardisation exercise worked well, one trainee assessor (Assessor E) was unable, at the end of the day's training, to award grades which adequately fit the model. As a result, it was recommended to the Hong Kong Examination Authority that this assessor be not invited to become an assessor for the Speaking Test.

The success of the training underscores the need for rigorous standardisation and training. It is strongly recommended that any future training sessions for the Speaking Test benchmark assessors should take one full day for the training programme, following the procedures outlines above. It should be stressed that after the initial group of assessors have been trained, both they and subsequent assessors must receive the same thorough training as did the first batch, and, in addition, be moderated shortly before the live administration of the test (see Lumley and McNamara (1995: 69). As Lumley and McNamara (1995) further state, reliability is greatly

enhanced by the use of two assessors. In a high-stakes examination such as that reported upon in this paper, it is essential that two assessors be used in each live administration.

To fail to carry out such rigorous training, standardisation and subsequent moderation procedures means that reliability is likely to be weakened and standards of assessment deteriorate. As stated above, this experience has been contrasted with the two-hour briefing, the standard briefing time, that the HKEA allocates for assessor training for the oral tests in their public examinations. Such training would, in the light of this paper, appear to be *not* adequate preparation for the assessors carrying out assessment in a professional, high-stakes examination such as the language benchmark Speaking Test.

References

Advisory Committee on Teacher Education and Qualifications. 1998. *Syllabus Specifications: Specimen Questions and Notes for Classroom Language Assessment*. Hong Kong: Government Printer.

Bachman, L. F. 1990: *Fundamental consideration in language testing*. Oxford: Oxford University Press.

Brindley, G. (ed.) 1995. *Language assessment in action*. Sydney: NCELTR.

Coniam, D. and P. Falvey. 1996. *Setting language benchmarks for English language teachers in Hong Kong secondary schools*. Advisory Committee on Teacher Education and Qualifications: Hong Kong.

Coniam, D. and P. Falvey. 1998. *Validating the Speaking Test : The Hong Kong English Language Benchmarking Initiative*. Advisory Committee on Teacher Education and Qualifications: Hong Kong.

Coniam, D. and P. Falvey. (2000). The English Language Benchmarking Initiative: A Validation Study of the Classroom Language Assessment Component. *Asia Pacific Journal of Language in Education*, 2, 2.

Elder, C. 1994. Performance testing as a benchmark for LOTE teacher education. *Melbourne Papers in Language Testing, 3*, 1, 1-25.

Eraut, M. 1994. *Developing professional knowledge and ompetence.* London: The Falmer Press.

Falvey, P. and S. Andrews. 1994: The Cambridge Examination in English for Language Teachers (CEELT). In Boyle, J, and Falvey, P. (eds.) *English Language Testing in Hong Kong*. Hong Kong: The Chinese University Press.

Falvey, P. and D. Coniam. 1997. Introducing English language benchmarks for Hong Kong teachers: a preliminary overview. *Curriculum Forum, 6*, 2, 16-35.

Falvey, P. and D. Coniam. 1998. Assessor Training and Standardisation for Classroom Language Assessment: The Hong Kong English Language Benchmarking Initiative. Advisory Committee on Teacher Education and Qualifications: Hong Kong.

Falvey, P. and D. Coniam. 1999. Assessor Training and Standardisation for Classroom Language Assessment: The Hong Kong English Language Benchmarking Initiative. Advisory Committee on Teacher Education and Qualifications: Hong Kong.

Linacre, J. M. 1994. *FACETS*: Rasch Measurement Computer Program. Chicago: MESA Press.

Lumley, T. and T. McNamara. 1995. Rater Characteristics and Rater Bias: Implications for Training. *Language Testing, 12*, 1, 54-71.

Lunz, M. and J. Stahl. 1990. Judge consistency and severity across grading periods. *Evaluation and the Health Profession, 13*, 425-444.

McDowell, C. 1995. Assessing the language proficiency of overseas-qualified teachers: the English Language Skills Assessment (ELSA). In Brindley, G. (ed.) *Language assessment in action.* Sydney: NCELTR, 11-29.

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

Stansfield, C., L. Karl and D. Kenyon. 1990. *The Guam educators' Test of English Proficiency (GETEP).* Final Project Report, Revised. Center for Applied Linguistics, Washington, D.C.

Webb, L., M. Raymond and W. Houston. 1990. *Rater Stringency and Consistency in Performance Assessment.* Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).

Weigle, S. 1998. Using FACETS to model rater training effects. *Language Testing, 15,* 2, 263 - 287.

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