The 37th IAEA Conference 2011

Characteristics of Raters: Reliable and Valid Versus Unreliable and Invalid

CHEUNG Kwai Mun Amy
Hong Kong Examinations and Assessment Authority
Email: kmcheung@hhkeaa.edu.hk

Abstract

This study used the multi-faceted Rasch model, classical statistics and verbal aloud protocols (VAP) to analyse raters' decision-making behaviours on a large-scale oral assessment, involving a stratified sample of 150 Secondary 3 (Grade 9) student performances from Hong Kong. Three different types of raters were used: four Non-native English Teachers, four Native English Speaking Teachers and four Naïve Native English Speakers. It was found that different raters behaved differently in their rating work. Moreover, the quantitative data obtained from Rasch and classical statistics tended to give the same results and supported the data from VAP. For reliable and valid judgement, four seemingly causal factors were identified: 1) Good understanding of test-takers' knowledge in terms of lexical range and common world knowledge; 2) TESOL knowledge; 3) Ability to internalise the rating scale in terms of test takers' performances; 4) Optimal duration of rating (so as to avoid rater fatigue). Three seemingly causal factors contributing to unreliable and invalid judgement were identified: 1) 'Applying partial requirements of the rating criteria' typically involved the inability to separate high and low score points on *stress*, *intonation* and *vocabulary*. 2) 'False markers of ability' i.e., English lexical items commonly 'imported' into Cantonese (students' mother tongue) which seemed 'advanced' (to raters unfamiliar with the cohort), yet were not truly indicative of high student ability. 3) 'Jarring errors' i.e., minor errors like 'go to shopping' which were so obviously 'non-native' in type as to have a disproportionately powerful negative effect on all raters.

Key words: Oral language testing, reliability and validity, verbal aloud protocols

1. Literature Review

In testing students' productive skills (e.g., speaking) through performance tests, raters are essential, since it is their judgement which actualises the rating scale in terms of showing how good a performance is (in criterion-referenced or norm-referenced testing). Rater characteristics are of particular interest as far as the nature and extent of variability in performance assessment is concerned because rating scales by their very nature require raters for their implementation. However, rater behaviour is not uniform. Raters may vary in severity, interpretation of the rating scale, level of alertness, and their selective attention to the various aspects of the performance which they are rating, e.g., grammatical accuracy, pronunciation accuracy and intonation accuracy. Studies indicate that rater training (Alderson, Clapham & Wall, 1995; McNamara, 1996, 2000) and experience (Weigle, 1998) have notable effects. It is also possible that TESOL training may affect the severity and consistency of raters. Studies such as Bonk and Ockey (2003) indicate that rater severity can change over time. Rater fatigue and other factors affecting reliability are also matters of concern in relation to intra-rater reliability. Cho (1999) indicates that raters can become fatigued during a rating session and thereby decline in reliability and Akiyama (2001) found data indicating the effects of rater fatigue within a single rating session. Moreover, Brown (2003) found differences between native and non-native teachers in their assessment of oral proficiency. Not all teachers in Hong Kong are TESOL (Teaching English to Speakers of Other Languages) trained. Even some of those brought in under the Native English Speaking Teachers (NET) Scheme lacked TESOL training. There is a paucity of research on the effect of TESOL training on ratings of oral proficiency. Furthermore, constant exposure to non-native English in daily life may either sensitise or desensitise raters to student errors. In summary, the factors which affect rating include:

- Selective attention to one or more characteristics of the performance being rated
- Degree of rater training
- Raters' background in terms of teacher training (e.g., TESOL/non-TESOL trained)
- Rater status re the language being assessed, i.e., native speaker, near-native speaker, non-native speaker
- Degree of rater experience
- Degree of severity
- Degree of consistency
- Rater fatigue
- Degree of exposure to non-native English in daily life (habituation/desensitisation to errors)

2. Research Question and Methods of Analysis

What are the characteristics of valid and reliable oral raters as well as unreliable and invalid oral raters?

To answer this question, both the qualitative and quantitative data were analysed to identify valid and reliable oral raters. Firstly, the 12 raters[1] were required to rate 115 performances[2] in four batches over a two-week period after they had received distance training. Then verbal aloud protocols (VAP) were performed and raters were required to re-watch eight vide clips which covered the full range of student abilities. To avoid any loss

[1] There were three types of raters (a total of 12) in this study: local English teachers (LETs), native English-speaking teachers (NETs), and naïve native English speakers (NNSs).

[2] A stratified sample of 10 schools from the 452 schools in Hong Kong SAR was obtained from the Hong Kong Examinations and Assessment Authority (Cheung, 2010). Only 115 out of 150 performances were used for raters' rating and the remaining performances for rater training in this study.

of information, they could take notes while completing the task. Immediately after watching each video clip three times, raters started recording and explained aloud their rating of each aspect of the performance in the following order: 'ideas and organisation' (IO), 'vocabulary and language patterns'(VL) as well as 'pronunciation and delivery' (PD) (Rating scale can be found in Appendix 1). Secondly, quantitative methods used as part of triangulation to identify raters' decision behaviours as follows (see Table 1 for figures):

Table 1. Summary of Rater Performance Statistics

Rater	Rater type	Native speaker	Training	Exposure	Infit Mn Sq	Outfit Mn Sq	Severity	Corr w/ EP[1]	IO Indices	VL Indices	PD Indices	VAP % Relevance	D from M (z)[3]
Acceptable range	NA	NA	NA	NA	0.7 – 1.3		-1.5 – 1.5	> 0.7	Corr > 0.7[2]			> 70%	> -2
LET1	LET	NON	TESOL	High	0.76	0.78	0.33	0.79	0.79	0.75	0.75	97%	1.18
LET2	LET	NON	TESOL	High	**0.55**	**0.56**	0.51	0.81	0.79	0.82	0.86	100%	**-3.78**
LET3	LET	NON	TESOL	High	0.74	0.84	0.56	0.80	0.88	0.82	0.85	100%	**-3.07**
LET4	LET	NON	TESOL	High	0.90	0.89	-1.23	0.76	0.78	**0.67**	0.74	100%	**-2.84**
NET1	NET	Native	NTESOL	High	**0.63**	**0.63**	0	0.75	0.75	0.77	0.82	95%	2.6
NET2	NET	Native	NTESOL	High	1.06	1.04	-0.23	0.77	0.79	0.72	0.79	77%	-0.24
NET3	NET	Native	TESOL	High	1.18	1.10	**1.81**	0.74	0.75	0.74	0.72	96%	0.24
NET4	NET	Native	TESOL	High	0.89	0.86	-0.35	0.77	0.82	0.79	0.77	93%	0
NNS1	NNS	Native	NTESOL	Low	**1.49**	**1.43**	-0.60	0.73	0.79	0.77	0.74	84%	3.78
NNS2	NNS	Native	NTESOL	Low	**1.49**	**1.56**	-0.05	**0.69**	0.78	**0.63**	**0.63**	84%	**-2.36**
NNS3	NNS	Native	NTESOL	Low	0.97	0.98	-0.85	0.77	0.75	0.71	**0.68**	94%	7.8
NNS4	NNS	Native	NTESOL	Low	1.37	1.32	0.11	**0.69**	0.80	0.79	0.77	69%	**-3.31**

Remarks:
1. 'Corr w/ EP' refers to 'the correlation of each rater's ratings with the expert panel'.
2. Correlation of each rater's ratings with the combined indices of each assessment criterion (IO: 'ideas & organisation', VL: 'vocabulary & language patterns', PD: 'pronunciation & delivery').
3. 'D from M (z)' refers to 'the distance in terms of z values away from the mean'. For example, if a z score is 2.5, then the sample mean is 2.5 standard deviations above the population mean.

- The fit values of each rater for rater consistency and rater severity in logit values were calculated using FACETS software (Linacre, 1991-2008) to run the multi-faceted Rasch analysis. The acceptable ranges of infit mean square and outfit mean square were the more conservative lower and upper limits of acceptability, 0.7 and 1.3, as employed by McNamara (1996) and Myford and Wolfe (2000).
- An acceptable range of severity from the range of -0.5 to +0.5 was set, based on the overall severity of raters. This acceptable range showed the same results when rater severity was calculated using a 95% probability of the mean of raters' ratings lying within

the mean interval (range of 2.59-2.94). Those who are beyond this range were considered to be either too lenient or too severe.

- Correlation between raters' observed ratings and the ratings by an expert panel was calculated using Spearman 'ρ'. The minimally acceptable level of inter-rater correlation coefficient was taken as 0.7 (as used by Brown (1996), Hughes (1989), and Lado (1961).) According to guidelines for interpreting correlations (Burns, 2000, p. 235), correlation between 0.4 and 0.7 is considered 'moderate'.
- Graphical analysis was conducted to investigate the relationship between minimum, median and maximum values for verifiable quantitative measures 'VQM'[3] (on combined indices of the three assessment criteria) and individual rater's observed ratings. The median value of each combined VQM indices was expected to increase consistently as rating levels rose. Regarding external validity in this study, high levels of correlation (>0.7) were expected between VQM and ratings and those lower than 0.7 were highlighted for caution.
- Raters' verbal aloud protocols (VAP) were coded and quantified. Number of words and percentage of relevant comments were calculated. For the distance away from the mean of number of relevant comments, the acceptable level was where z score > -2. Those who were not at this level were considered as not saying enough to justify their ratings.

3. Results and Discussion

Using both the qualitative and quantitative data, it was found that different raters behaved differently in terms of their rating work. Moreover, the quantitative data obtained from non-VAP methods tended to give the same results and therefore supported the data from VAP. As a consequence, factors contributing to valid and reliable judgement, as well as to invalid and unreliable judgement, were identified and analysed.

3.1 Reliable and Valid Judgement VS Unreliable and Invalid Judgement

3.1.1 Background Knowledge of the Test Takers

Among the three rater types, the naïve native speaker (NNS) raters had the worst performance according to both quantitative and qualitative measures. This is not surprising, since these raters had the least background knowledge of the test-takers (students) and lacked TESOL training. This finding seemed to indicate that such knowledge was an aid in making valid and reliable judgements. According to the results of quantitative measures, the NNS raters' percentage of relevant comments was comparatively low in vocabulary and language patterns (VL) (NNS: 77%; NET: 86%; LET: 97%), which corresponded to the highest fit values in VL (infit: 0.97-1.63; outfit: 0.98-1.78) among the three assessment criteria. The fit values >1.3 indicated that their ratings were not consistent and at times erratic. NNS raters' patterns of minimum, median and maximum values of combined VQM indices were irregular, indicating they had difficulty in giving certain scores or gave their ratings arbitrarily. The correlations between NNS raters' ratings and VQM were only 'moderate', not 'high' like the native English teacher (NET) and local English teacher (LET) raters.

Most NNS raters had difficulty in consistently rating students, especially on VL. Scoring guides such as 'use basic language patterns with possible errors' and 'use familiar vocabulary appropriately' were difficult for NNS raters to interpret. Likely errors and typical Hong Kong student vocabulary were beyond their understanding. VAP from NNS raters also

[3] Verifiable Quantitative Measures (VQM) are counted from observable aspects of transcribed student performances, e.g., grammatical errors, syntactic complexity and pronunciation errors (Cheung, 2010).

indicated that they had problems rating because they lacked background knowledge of the students. One NNS said:

> *"…because I'm a native English speaker, I'm very familiar with the vocabulary and (but) I'm not aware of how much vocabulary the speaker would have studied; so for me to make a judgement is to use familiar vocabulary, you know I didn't feel I'm in a good position to do that…"*

Another NNS added:

> *"I thought the most difficult to rate was the vocabulary and language. This is because I am not trained in this area and I take for granted some of the complex sentence structures English has. I also have no training in [the] Chinese language and therefore the differences in grammar I thought."*

One NNS gave justifications as to why she did not know how to rate students on VL:

> *"I don't know what is familiar to students – what vocabulary they have studied; I am not aware of how much vocabulary the student [has] studied."*

False Markers of Ability

When giving higher ratings for IO, in which expansion of ideas was one of the required components, one of the NNS raters was amazed by the fact that a student could list the names of Christmas carols, and justified a score of 5 for idea expansion. The NNS raters lacked enough cultural background to know that Hong Kong students celebrate Christmas at school and knowing the names of most of the famous Christmas carols was common. This caused them to give high ratings based on 'false markers of ability'. This in turn resulted in very high fit values in VL among the NNS raters.

Jarring Errors

Raters with little exposure to non-native English and without TESOL training may have left them vulnerable to 'jarring' errors as they lack the training and experience to identify errors in the midst of a potentially confusing mess of 'strange' English. However, it is worth noting that 'jarring' errors, such as 'thank you day', 'I go to shopping', 'I very like' and 'the Christmas party is very well', had a disproportionately powerful negative effect on all raters, even experienced LET raters and NET raters.

3.1.2 Knowledge of the Subject Matter

It was found that raters with TESOL training had a higher percentage of relevant comments in VAP and better understood the requirements of the rating scales than their non-TESOL trained counterparts.

Internalising Test Takers' Performances from the Rating Scales

The graphical analysis investigating the relationship between minimum, median and maximum values for verifiable quantitative measures (VQM) and raters' ratings indicated that competent raters were able to internalise marginal performance at each score point. When applying the rating scales, they were able to consider all the composite entities of each level descriptor. The results show that raters with good fit values and high correlations with VQM also demonstrated a good understanding of rating scales in their VAP remarks.

Not Using the Full Range of the Scale When Rating

Student performances were selected according to stratified random sampling which gave a miniature sample of the Secondary 3 (S.3) student population in Hong Kong, so a bell curve distribution of abilities should have been obtained. Such a distribution should have allowed raters to use the full range of the scale. However, some raters, especially LET raters, only gave a limited range of scores. This use of a limited range showed itself on very low fit values which were outside the acceptable range (0.7-1.3), and in extreme values of Z Std't' well outside the normal range. Ratings from raters giving narrow ranges of scores could be up to -7.2 standard deviations (e.g., LET2) from the mean.

Some misconceptions became evident from the VAP of raters without TESOL training.

Confusion of the Rating Criteria

From the findings of this study, raters who were non-TESOL trained were unwilling or unable to separately mark the various rating criteria or confused the criteria. For example, one rater commented:

> *"I found it difficult to distinguish between 'ideas and organisation' and 'vocabulary and language patterns' since there is a lot of overlap in these criteria."*

'Coherence' is one of the components of the 'organisation' construct while 'clarity' relates to the pronunciation construct. However, when one of the raters gave comments on rating PD, she confused these two constructs: she said, *"Coherence and clarity was intertwined with pronunciation"* and she could not distinguish one criterion from the other.

This is the very antithesis of what analytic rating scales set out to do. While there is some commonality of sub-constructs between these criteria, confusion of criteria weakens both the validity and reliability of the ratings.

Incomplete Understanding of the Rating Criteria

Some raters verbalised their judgement according to the scale wordings and they were found to be valid in their judgement. However, when they actually gave scores, these raters gave either lower or higher scores than students deserved based on the Rasch fair average. This was reflected in score distribution (in terms of minimum, median and maximum VQM values). For example, the median VQM value of a 3 rating for some raters was lower than that of a 2 rating for other raters.

A rater indicated that she did not like separating criteria and would have preferred some kind of holistic impression mark. She said,

> *"Some other criteria could be the overall feeling given from the speaker which affects communication."*

Moreover, references to concepts such as 'overall feeling' reflect the very problems that an analytic scale is intended to address.

Using Wrong Criterion for Justification

If raters were internally consistent yet idiosyncratic, they might give a full range of scores to the test-takers but use the wrong criterion for score justification throughout the entire assessment. Such raters may not be identified by Rasch analysis as Weigle (1998)

points out. However, if their correlations with VQM were low, it could be concluded that they were not rating according to the criteria and so were idiosyncratic raters, despite having an acceptable range of fit values according to Rasch analysis.

Using a wrong criterion for rating may lead to problems of validity in rating since other features beyond the criterion are taken into consideration or some features of the criterion are omitted. In other words, test takers (students) were not being assessed on what they were supposed to be assessed on.

Applying Partial Requirements of the Rating Criteria

In this study, each assessment criterion consisted of two to four composite entities. For example, PD included pronunciation, fluency, stress and intonation. Students scoring a 3 attempted native-like stress and intonation, and 'intonation' would not be normally mentioned for the scores below 3 on the rating scale. In some raters' VAP, comments on 'intonation' were missing when raters justified their scores of 3 or above. Therefore, scores of 3 or above given by these raters may not have been valid since it seemed that they had only partially applied the requirements of the rating criteria. This may be a reason why the raters in question had difficulty differentiating a 3 from a 4 in the assessment criteria.

When rating PD, non-TESOL trained raters mainly took 'pronunciation' into consideration. For example, one rater said:

> *"The easiest to rate was the pronunciation as I could judge if it sounded right. It was simply can I understand it and does it sound nice."*

Another such rater assessed PD at word level only:

> *"The easiest criterion to rate is the pronunciation and delivery. The first thing I noticed was that they could not pronounce or deliver words properly."*

These raters only looked at cohesive devices when rating coherence. One example came from a rater, who said:

> *"I found it the easiest to rate 'Ideas and Organisation'. It is measurable, it's easy to tell if the ideas are expanded and if it flows logically and if words are used to connect/organise like 'then', 'first', 'also', etc."*

However, studies such as Widdowson (1983) argue that much of the coherence of a text is inexplicit rather than explicit.

3.1.3 Comparing Students' Performances

Raters were under precise instructions not to compare students, nevertheless some did so. Raters who admitted to comparing student performances in VAP tended to have excessively high fit values according to Rasch analysis, showing that they were not consistent and even erratic at times. For example, NET3 said:

> *"I felt my rating changed as I went through the process [and] I compared students to previous students."*

One NNS rater added:

> *"I know we are not supposed to compare speakers. However, it was quite difficult…because some speakers obviously have [a] better grasp of vocabulary than [the] other speakers did so you know that is problematic as well."*

The above two examples compare sharply with raters who had acceptable fit values. For example, NET2 said:

"I always have to remind myself not to make comparisons."

One of the reasons for making comparisons (even against clear instructions not to do so) may be that raters lacked familiarity with the rating scales; they relied on comparison with earlier student performance rather than on the set standards of the rating scale.

3.1.4 Overgeneralising Student Performances

Raters who overgeneralised student performances usually gave a hasty conclusion without listening to the performances in detail. One reason for this seemed to be that the raters in question had not internalised the student performance as a whole. Instead, they could only identify some discrete chunks from which they overgeneralised. One example illustrates the point: NNS1 gave one student a score of 5 in IO when in fact the correct rating was a 3. She said:

"I thought it's organised well; you know she did use connectives."

This indicated that she thought that connectives were the key factor in coherence. The same rater gave a score of 3 in VL instead of a 2 which she justified by stating:

"However, he does use the[a] word like 'delicious'."

The rater was obviously impressed, but did not realise that 'delicious' was a common vocabulary item for even the lowest level Hong Kong students.

3.1.5 Duration of Rating

Rater fatigue and other factors affecting reliability are also matters of concern in relation to intra-rater reliability. Cho's (1999) study indicates differences in rating between sessions for the same rater, and Akiyama (2001) found data indicating the effects of rater fatigue within a single rating session. In this study, most raters found that rating 30 student performances (maximum two minutes for each) at one time was appropriate. A minority found it boring and only one found it too difficult. It was generally felt that 30 was a good batch size. To quote one rater: *"if raters rate more than 30, they get tired and if less than 30, they may not be consistent"*. There was no empirical evidence to suggest rater fatigue was a problem with 30 students per session.

4. Conclusion

As Huot (1993, p. 203) states, ratings and scales are sets of negotiated principles which raters use as bases for reliable action, rather than valid descriptions of language performances. Therefore, rater training provides a channel where these sets of principles can be negotiated and followed. Rater training should ensure that raters understand the negotiated principles for each assessment criterion and how to put them into practice. Many of the problems exhibited by raters stem either from lack of understanding of the real English proficiency of Hong Kong students or inability to understand rating criteria. Therefore, to achieve adequate consensus among raters in a large-scale assessment context during training, raters should meet the following selection criteria:

- Currently employed in Hong Kong as language teachers, with at least three years' experience in teaching the relevant key stage
- TESOL trained (recognised TESOL training, e.g., CELTA, TESOL components of Diploma of Education, Bachelor of Education or Post-graduate Certificate/Diploma of Education)
- Have at least one year's teaching experience at each band (this would be an advantage but may not be easily available)

References

Akiyama, T. (2001). The application of G-theory and IRT in the analysis of data from speaking tests administered in a classroom context. *Melbourne Papers in Language Testing, 10*, 1-22.

Alderson, J. C., Clapham, C. & Wall. D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of second language group oral discussion task. *Language Testing, 20*, 89-110.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*, 1-25.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Burns, R. B. (2000). *Introduction to research methods* (4th Ed.). French Forest: Pearson Education Australia.

Cheung, K. M. A. (2010). *Reliability and validity in practice: Hong Kong's Key Stage 3 oral assessment.* Macquarie University, Australia, Unpublished PhD thesis.

Cho, D. (1999). A study on ESL writing assessment: Intra-rater reliability of ESL compositions. *Melbourne Papers in Language Testing, 8,* No. 1, 1-24.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations.* Cresskill, N. J.: Hampton Press.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. New York: McGraw-Hill Book

Linacre, J. M. (1991-2008). *A user's guide to FACETS: Rasch-model computer program*. *Version 3.64*. Chicago, IL: Winsteps.

McNamara, T.F. (1996). *Measuring second language performances*. London: Longman.

McNamara, T. F. (2000). *Language testing*. Oxford: Oxford University Press.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the test of spoken English assessment system*. Princeton, NJ: Educational Testing Service.

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

Widdowson, H. G. (1983). *Learning purpose and language use*. Oxford: Oxford University Press.

Appendix 1

6-Point Analytic Rating Scale for each BC Descriptor

Score	Ideas & Organisation	Vocabulary & Language Patterns	Pronunciation & Delivery
5	• Express ideas that are relevant and expanded, when appropriate, with explanations and detail • Organise ideas clearly and coherently	• Use varied and appropriate language patterns • Use a good choice of vocabulary	• Speak clearly, accurately and fluently • Some features supporting communication
4	• Express ideas that are relevant to inform and explain with details • Communicate ideas clearly and coherently	• Use varied and appropriate language patterns • Use appropriate vocabulary	• Speak clearly and fluently, with few or no errors in pronunciation • Use intonation to enhance communication
3	• Express ideas in some detail that are relevant to inform and/or explain • Communicate most ideas clearly and coherently	• Use mostly appropriate language patterns • Use mostly appropriate vocabulary	• Speak clearly with some errors in pronunciation and occasional hesitation • Make occasional attempts to use intonation
2	• Express adequate ideas that are relevant to the topic • Communicate some ideas clearly and coherently	• Use simple language patterns • Use familiar vocabulary appropriately but with errors that may impede communication	• Speak clearly though hesitant with errors in pronunciation that may impede communication OR • Occasional hesitant/stilted speech that may impede communication
1	• Express limited/disjointed ideas that are relevant to the topic	• Use basic language patterns with possible errors • Appropriately use vocabulary drawn from a limited and very familiar range, awkward wording may make understanding unclear	• Speak with frequent errors in pronunciation that impedes communication OR • Hesitant/stilted speech that impedes communication
0	• Do not express any relevant or understandable information OR • Make no attempt at all	• Do not produce any recognizable words or language patterns OR • Make no attempt at all	• Do not produce any comprehensible English speech OR • Make no attempt at all

Remarks: Scores 0 – 4 adapted from the score guide of English Oral Component of S.3 TSA, HKEAA