Investigating Reliability and Validity in Rating Scripts for Standardisation Purposes in Onscreen Marking

CHEUNG Kwai Mun Amy and CHANG Rui
Hong Kong Examinations and Assessment Authority

ABSTRACT

This study investigated the reliability and validity of Assistant Examiners (AEs) in rating the standardised scripts used as benchmarks in onscreen marking (OSM) of the written component of Primary 6 English Language in the Territory-wide System Assessment in the Hong Kong Special Administrative Region. Marking criteria included 'Content' and 'Language.' Standardised scripts were employed for three purposes: 1) training markers, 2) qualifying markers before they started rating and 3) check-marking the markers at random intervals throughout the entire OSM period. Therefore, these standardised scripts played a vital role in monitoring the marking quality even with the cutting edge technology of OSM. Scripts were drawn from a stratified sample (N=250 students) from a total of some 580 participating schools with a student population of 72,000. Having all such scripts marked by all AEs (a total of 250 scripts) would have been time-consuming and induced 'rater fatigue' which was likely to affect rater reliability. Therefore, 'overlapping marking' was adopted where AEs only needed to rate less than 70 scripts each. Each rater had about 20 scripts overlapped another rater thus forming an unbroken chain of overlap. This data enabled correlations between expert panel ratings and AEs' ratings and the Multi-faceted Rasch Model was run to calculate the 'fair average' (FA) for all AEs and 'infit' for each rater. To 'externally' validate the ratings, verifiable quantitative measures (VQM) were used as a check which correlated against both FA and individual ratings. The VQM included 'number of meaningful clauses', 'syntactic complexity', 'lexical variation', 'families of words', etc. The results yielded correlations in the range of 0.6 to 0.9 for FA ($\alpha < 0.05$) and 0.4 to 0.8 for individual raters ($\alpha < 0.05$) showing that the method used in rating scripts for standardisation purposes was in most cases valid and reliable, especially when FA was used.

BACKGROUND

Territory-wide System Assessment (TSA) is a standards-referenced assessment which was conceived of as a 'low-stakes' survey of the performance of student groups. The main purpose of TSA as seen by the Hong Kong Education Commission, was to provide the Hong Kong Special Administrative Region (HKSAR) Government and school management with information on **school** standards in key learning areas for the purposes of school improvement so that the Government would be able to identify schools in need of assistance. TSA is held annually in June (end of academic year) and marking of Primary 6 (Grade 6) Territory-wide System Assessment (TSA) written papers is conducted over a two-week period in July. For TSA 2008, a Chief Examiner (CE) was appointed from the tertiary sector to take charge of conducting the Assistant Examiners' and Markers' training in conjunction with the Manager-in-charge of the subject level from the Hong Kong Examinations and Assessment Authority (HKEAA). Approximately 70 Markers and nine Assistant Examiners (AEs) were recruited. All of such assessment personnel were required to have attained the Language Proficiency Assessment for Teachers (LPAT) qualification in English, (a mandatory requirement for teaching English in the HKSAR school system.) Any teacher who wished to serve as a Marker had to attend a four-hour training session which included discussion of rating criteria and rating of language samples. Teachers were also required to review these ratings focussing on how well they described the samples in question.

LITERATURE REVIEW

In testing students' productive skills (e.g., writing) through performance tests, raters are essential since it is their judgment which actualises the rating scale in terms of showing how good a performance is when compared with either scale criteria (criterion-referenced), or the performances of other students doing the same test (norm-referenced). Hence, when we refer to raters, we focus on the consistency of raters' application of the standards, i.e., rater reliability (McNamara, 1996). Consistent application of the standards requires that rater decisions be 'objective' or at least trans-subjective (Foucault, 1974). Hence, Bachman and Palmer (1996) call for substantial rater training, including rating language samples and that raters review these ratings and discuss how well they describe these samples. Hence, rigorous rater training must have reliability and validity of the selected sample (written scripts in this Study) as a pre-requisite. However, rater fatigue may be a factor weakening the reliability of marking and raters tend to rate more severely over time. Cho's (1999) study showed 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. Therefore, when assigning experienced raters to rate scripts for standardisation purposes (one of which is for rater training), they should not be overloaded.

In the field of testing language proficiency, various 'objective' measures of syntactic complexity have been employed. However, as Foucault (1974, p.94) points out such measures are perhaps better termed trans-subjective. For example, the length of T-units and the number of clauses per T-unit, a measure of syntactic complexity, is found to be the best method to predict learner proficiency (Iwashita, 2006, p.162). A T-unit is a dominant clause and its dependent clauses, as described in Hunt (1965, p.20) who defined it as 'one main clause with all subordinate clauses attached to it'. Syntactic complexity (syntactic maturity or linguistic complexity) is described by Ortega (2003) as 'the range of forms that surface in language production and the degree of sophistication of such forms' (p.492). Syntactic complexity has been extensively investigated in L2 writing studies as well as in L2 speech data (Crookes, 1989; Ortega, 1999; Skehan & Foster, 1999). Iwashita (2006) points out that T-unit length used as an index of syntactic complexity seems to be 'the only measure found by both written and oral language (studies) to discriminate proficiency levels satisfactorily' (p.155) and the findings of his study shows that 'the number of T-units and number of clauses per T-unit is found to be the best way to predict learner proficiency and the measure has a significant linear relation with independent oral proficiency measures' (p.165).

Correlations have traditionally been done using Pearson's r rather than Spearman's ρ despite the fact that ratings do not fit the assumptions for Pearson's r (Burns, 2000), i.e., they are ordinal level data and do not always follow a normal distribution. Having said that, Pearson's r and Spearman's ρ results for the same data set are not always substantially different, as Bonk & Ockey (2003) found in their work. Moreover, 'ρ' can be misleading since it involves converting ratings to ranks which can result in a lot of 'tied' ranks if there are a lot of similar ratings. This in turn can cause a correlation to be under-represented. Conversely, using Pearson's r can over-represent a correlation when used on ordinal level data because such usage involves the plastic interval assumption, i.e., that the difference between any two points on a rating scale (e.g., 1 and 2) is the same as that between any other two points (e.g., 4 and 5). Therefore, it is reasonable to use both 'ρ' and 'r' and look at the differences. If the difference is negligible then it is possible to square the 'r' value and obtain a variance estimate which can show what percentage of the variance is attributable to the variables in the correlation. Suppose, for example, we found that all raters correlated with

each other at an 'r' value of 0.7. We can square this and obtain a figure of 0.49 indicating that the commonality between raters accounted for 49% of the variance in ratings. This procedure is not possible using 'ρ' and therefore constitutes a good reason for using 'r' provided that the 'r' values obtained are not greatly different from the 'ρ' values.

The focus of this study was to investigate the reliability and validity of Assistant Examiners (AEs) in rating the standardised scripts used as benchmarks in onscreen marking (OSM) of the written component of Primary 6 English Language in 2008 TSA in the HKSAR. From this data, the correlation coefficients between expert panel ratings and AEs' ratings were calculated and the Multi-faceted Rasch Model was run to calculate the 'fair average' (FA) for all AEs and 'fit' values for each rater. To validate the ratings, verifiable quantitative measures (VQM) were used as an external validity measure which correlated against both FA and individual ratings. The VQM included 'number of meaningful clauses', 'syntactic complexity / number of T-units', 'lexical variation', 'types of words', 'families of words' and 'tokens'.

METHODOLOGY

1. Assigning Assistant Examiners to Marking

The Primary 6 written component was double marked. Before marking, all markers were trained to ensure familiarity with the marking schemes and were required to demonstrate consistency throughout the entire marking period. To ensure markers' consistency and marking quality, standardisation scripts were used at the three stages of the marking process: 1) Training (~110 scripts); 2) Qualification – to assess markers whether they have met the set requirements before commencing marking (~60 scripts); and 3) Control – to monitor markers' quality during marking; scripts were randomly assigned to each marker throughout the process (~80 scripts).

Scripts to be used for standardisation were drawn from a stratified random sample N=250 from a total of some 580 participating schools with a total student population of 72,000. These scripts were randomly selected from the total population and marked by nine AEs where 'overlapping marking' was adopted. Each AE was only required to rate a maximum of 70 scripts (each of which had about 80 words) in a three-hour session. In other words, AEs were only required to rate about 25% of the total number of scripts. This arrangement was cost effective and helped reduce the chance of rater fatigue which might give rise to low reliability. For each rater, 20 – 30 scripts overlapped with one other rater so that they formed an unbroken chain of overlap.

2. Verification of Ratings by Expert Panel

An expert panel was assembled to verify the scores of the standardisation scripts. This panel consisted of a Chief Examiner (drawn from the tertiary sector), the Manager-in-charge of the level and two subject officers from the HKEAA. 'Fair average' scores (derived from Rasch analysis (Linacre, 1989-2008)) were obtained from ratings by the nine most experienced AE's. The expert panel made reference to the FA scores for verification. Adjustments on scores were made based on members' professional judgement in cases where members did not agree with the Rasch's FA. Only about 6% of the FA ratings required adjustment. Adjustments were made in some scripts with FA scores of 1.5 or 1.6 after rounding them up to '2'. After judging by the expert panel, the scripts in question were adjusted to a score of '1'. However, no adjustments were required when the scripts were rounded down from 2.1 or 2.2 to '2' or rounded up from '1.8' or '1.9' to '2'.

3. Deriving Verifiable Quantitative Measures from the Sub-Sample (N=85)

85 of the 250 performances were 'counted' for all aspects of the assessment criteria. The verifiable quantitative measures (VQM) for 'Content' included 'number of meaningful clauses' and 'syntactic complexity'. 'Number of meaningful clauses' was adopted as a reference for calculating the number of intelligible ideas students used for 'Content'. If an idea was repeated, the idea in question was only counted once. The data for 'Content' VQM were categorised and counted by the Researcher and verified by an expert who also had a strong background in grammar and L2 errors **and** was familiar with the errors typical of Hong Kong students. Syntax is one of the most basic organising principles in language; and syntactic complexity (as a VQM) has been used in many studies of proficiency in both L2 writing and L2 speaking. In this study, T-unit was adopted as a measure for syntactic complexity.

VQM for 'Language' consisted of 'tokens', 'types of words', 'families of words' and 'lexical variation'. The data for 'Language' VQM were machine counted. Tokens were calculated using a computer software package called RANGE (Heatley et al., 2002). RANGE was used to compare a text against vocabulary lists to see what words in the text were and were not in the lists, and to see what percentage of the items in the text were covered by the lists. Types and families of words were also derived from the RANGE programme where lexical variation explored the number of different words (types) produced by the test-takers in relation to the total number of words produced. The following calculations were performed:

$$\text{Lexical Variation} = \frac{\text{Types (No. of different words)}}{\text{Tokens (Total no. of words)}}$$

4. Calculating 'Fit' Values and Rater Reliability

Multi-faceted Rasch analysis was run using FACETS (Linacre, 1989-2008) to calculate the 'fit' mean square values and rater severity. Correlation between the nine AEs' observed ratings and the final ratings agreed by the expert panel on two assessment criteria, i.e. 'Content' and 'Language' were calculated using Pearson's r and Spearman's ρ.

5. Calculating Correlation Using Pearson's r and Spearman's ρ

For the sub-sample of 85 student performances which had been subjected to verifiable quantitative measures (VQM), correlations were done between nine raters' 'fair average' scores (derived from Rasch analysis (Linacre, 1989-2008)) and the VQM derived data using both Pearson's r and Spearman's ρ as cross checks against each other.

LIMITATIONS

Although the verifiable quantitative measures provided a useful external validity check on the raters' ratings, producing VQM for each construct was massively time-consuming and could only be done on a sampling basis (85 out of 250 student performances were selected). Furthermore, some aspects of the rating scale could not be quantified, for example, 'organisation of ideas' where human judgement was required rather than simply calculating the number of explicit cohesive devices. Moreover a written version of the 'home grown' measures of syntactic complexity used in (Cheung, forthcoming) could not be deployed in this study since the time and resources available for designing and calculating such a measure were not available at the time of production of this paper.

FINDINGS

Table 1 shows a detailed measurement report, with rater severity, error and fit statistics among raters (Assistant Examiners). The infit and outfit mean squares were within the acceptance range of (0.7 – 1.3) defined by McNamara (1996) and Myford and Wolfe (2000). The 'infit' mean square indicates the rater's internal consistency while 'outfit' mean square gives indications to their ratings on extreme scores, i.e. '0' and '4' for 'Content and '0' and '3' for 'Language' in this Study. According to Table 1, all AEs were within the range of 'infit' values, meaning that all of them were internally consistent. For the 'outfit' values which show information about the raters giving extreme scores, only two out of nine AEs (i.e., AE2) gave slightly unexpected ratings on both ends (outfit value of 1.32) while AE8 had slightly lower outfit value (0.69) showing she had given slightly limited range of scores on both ends. For rater severity, AE8 was the most severe with 2.31 logits while AE7 most lenient with –3.03 logits. Even though there were differences in severity among raters, the Rasch model could allow for compensation for differences in rater severity since the raters were internally consistent in rating. In general, person reliability (0.98) was high, meaning that the measurement error was low, and all AEs' ratings fitted the Rasch model (-2<ZStd<2).

Table 1. Assistant Examiner Measurement Report

Total Score	Total Count	Obsvd Avg	Fair-M Avg	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N raters
180	106	1.9	1.68	.62	.26	1.10	.7	.95	.0	.89	.93	.60	AE1
200	108	2.0	2.02	-1.23	.26	1.11	.7	1.32	1.1	.83	.92	.63	AE2
285	138	2.0	2.05	-1.40	.26	.94	-.3	1.06	.2	1.04	.96	.67	AE3
268	140	2.0	1.76	.17	.22	.78	-1.7	.71	-1.2	1.23	.93	.60	AE4
289	138	2.1	1.69	.55	.22	1.01	.1	.86	-.5	.95	.93	.57	AE5
239	128	2.0	1.80	-.05	.23	1.02	.1	1.15	.7	1.00	.94	.61	AE6
332	136	2.3	2.34	-3.03	.26	.84	-1.0	.74	-.7	1.18	.96	.64	AE7
215	136	1.5	1.43	2.31	.24	.80	-1.4	.69	-1.1	1.20	.94	.59	AE8
138	76	1.8	1.47	2.06	.30	1.06	.3	.00	.0	.97	.92	.56	AE9
238.4	122.9	2.0	1.81	.00	.25	.96	-.3	.94	-.2		.94		Mean (Count: 9)
57.6	20.6	.2	.27	1.60	.02	.12	.9	.20	.8		.01		S.D.(Population)
61.1	21.9	.2	.29	1.69	.02	.13	.9	.21	.8		.01		S.D.(Sample)

Model, Population: RMSE .25 Adj (True) S.D. 1.58 Separation 6.31 Reliability .98
Model, Sample: RMSE .25 Adj (True) S.D. 1.68 Separation 6.70 Reliability .98
Model, Fixed (all same) chi-square: 342.9 d.f.: 8 significance (probability): .00
Model, Random (normal) chi-square: 7.8 d.f.: 7 significance (probability): .35

In this study, correlation coefficients were used. Several authors have offered guidelines for the interpretation of a correlation coefficient. Burns (2000, p.235), for example, has suggested the following interpretations for correlations in psychological research, which have also become standard for applied linguistics, in Table 2.

Table 2. Interpretations for Correlations in Psychological Research (Burns, 2000)

Correlation	**Correlation**	**Relationship**
0.90 – 1.00	Very high	Very strong
0.70 – 0.90	High	Marked
0.40 – 0.70	Moderate	Substantial
0.20 – 0.40	Low	Weak
<0.20	Slight	Negligible

In order to check whether the Assistant Examiners were rating consistently with the standards, i.e., marking scheme, correlation coefficients between the ratings of the expert panel and each rater were computed. This provided a matrix of correlational data for each rater across the 250 written scripts. According to the data from Table 3, the levels of correlation of AEs' observed ratings with expert panel's ratings in 'Content' ranged from high to very high, showing that all Assistant Examiners showed very high standard of reliability in marking the written scripts. Seven out of nine showed very high correlations with the expert panel ratings, meaning that they showed consistency in their application of the standards in rating 'Content'.

Table 3. Correlations of Raters' Observed Ratings with Expert Panel (EP) Ratings in Content using Pearson's r and Spearman's ρ

Rater	AE1		AE2		AE3		AE4		AE5	
Correlation	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
EP	0.886	0.875	0.896	0.901	0.955	0.954	0.940	0.934	0.925	0.905
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Level of Correlation	High		Very High		Very High		Very High		Very High	
Rater	AE6		AE7		AE8		AE9			
Correlation	r	ρ	r	ρ	r	ρ	r	ρ		
EP	0.957	0.956	0.922	0.909	0.869	0.874	0.912	0.915		
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
Level of Correlation	Very High		Very High		High		Very High			

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

According to the data from Table 4, the levels of correlation of raters observed ratings with expert panel's ratings in 'Language' ranged from high to very high. Five out of nine demonstrated very high correlation with the expert panel ratings, meaning that they showed consistency in their application of the standards in rating 'Language'.

Table 4. Correlations of Raters' Observed Ratings with Expert Panel (EP) Ratings in Language using Pearson's r and Spearman's ρ

Rater	AE1		AE2		AE3		AE4		AE5	
Correlation	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
EP	0.916	0.912	0.858	0.882	0.934	0.927	0.916	0.914	0.907	0.880
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Level of Correlation	Very High		High		Very High		Very High		Very High	
Rater	AE6		AE7		AE8		AE9			
Correlation	r	ρ	r	ρ	r	ρ	r	ρ		
EP	0.926	0.925	0.877	0.857	0.868	0.855	0.888	0.885		
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
Level of Correlation	Very High		High		High		High			

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

To externally validate the ratings, verifiable quantitative measures (VQM) were used as a check which correlated against both FA and individual ratings on two assessment criteria, i.e. 'Content' and 'Language', using Pearson's r and Spearman's ρ. The VQM included 'number of T-units' for syntactic complexity, 'number of meaningful clauses', 'lexical variation', 'tokens', 'types of words' and 'families of words'.

For 'Content', 'number of meaningful clauses' ('r' value of 0.784; with 61.5% of variance explained) had higher correlation than 'syntactic complexity' ('r' value of 0.685 with 46.9% of variance explained). This indicated that 61.5% of the student performances in 'Content' were primarily influenced by 'number of meaningful clauses' while 46.9% influenced by 'syntactic complexity' (as measure by number of T-units).

For 'Language', 'families of words' ('r' value of 0.758, 57.5% of variance explained) had the highest correlation, followed by 'types of words' ('r' value of 0.737, with 54.3% of variance explained). This indicated that around 55% to 58% of the student performances in 'Language' were primarily influenced by 'families of words' and 'types of words'. 'Tokens' ('r' value of 0.556, with 30.9% of variance explained) had the second lowest correlation while 'lexical variation' ('r' value of -0.036, with 0.1% of variance explained) had a negative correlation indicating only a very weak inverse relationship. In other words, 'lexical variation' as measured in this study, did not predict ratings of student performances in 'Language'.

All the VQM of sub-constructs in Table 5 (except for 'lexical variation') showed 'moderate' to 'high' correlations with the fair average of their respective assessment criteria, meaning they were strong predictors of the ratings for their respective constructs. Moreover, 'types of words' and 'families of words' seemed to be good predictors (correlations >0.7) of the ratings for the construct **other** than the one they were supposed to predict. For example, 'families of words' not only had strong influence on its own construct – 'Language' ('r' value of 0.758, 57.5% of variance explained) but also had even stronger influence on 'Content' ('r' value of 0.847, 71.7% of variance explained).

Table 5. Correlations 'r' and 'ρ' of VQM of 85 student performances on 'Content' and 'Language' with raters' fair average scores

Verifiable Quantitative Measures												
Criteria	**Content**				**Language**							
Sub-constructs of VQM	No. of Meaningful Clauses		Syntactic Complexity		Lexical Variation		Tokens		Types of Words		Families of Words	
Correlation	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
FAof Content	**0.784**	**0.755**	0.685	0.673	-0.172	-0.063	**0.733**	**0.761**	**0.834**	**0.864**	**0.847**	**0.872**
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.115	0.567	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
FA of Lang	0.665	0.608	0.543	0.505	-0.036	0.099	0.556	0.561	**0.737**	**0.745**	**0.758**	**0.762**
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.741	0.366	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

** correlation is significant at the 0.01 level (2-tailed).
*correlation is significant at the 0.05 level (2-tailed).

According to Table 6, the levels of correlations between individual observed ratings and VQM for content were similar to fair average scores of all AEs and VQM for content. Among the nine raters, AE2 had the highest correlations between 'Content' observed ratings and VQM sub-constructs for 'Content' while AE1, AE3 and AE9 had comparatively lower correlations. For 'Language', individual AEs' 'observed' ratings had higher correlations with 'families of words' and 'types of words'. Among the nine raters, AE5 had the highest correlations between 'Language' observed ratings and VQM sub-constructs for 'Language' while AE1 and AE9 had comparatively lower correlations. 'Lexical variation' again proved problematic showing only small negative correlations against content indicating that it was not a good predictor of rating for 'Language'. This finding was similar to that for Secondary 3 oral in Cheung (forthcoming).

Table 6. Correlations 'r' and 'ρ' of VQM of student performances on 'Content' and 'Language' with Individual Raters' Respective Observed Scores

Verifiable Quantitative Measures												
Criteria	**Content**				**Language**							
Sub-constructs of VQM	No. of Meaningful Clauses		Syntactic Complexity		Lexical Variation		Tokens		Types of Words		Families of Words	
Correlation	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
AE1	0.662	0.653	0.564	0.525	0.070	0.202	0.339	0.347	0.704	0.699	0.686	0.683
Sig(2-tailed)	0.0001	0.001	0.008	0.014	0.763	0.380	0.133	0.123	0.0001	0.0001	0.001	0.001
AE2	**0.872**	**0.846**	**0.835**	0.795	-0.207	-0.187	0.620	0.738	0.790	0.772	0.775	0.760
Sig(2-tailed)	0.0001	0.0001	0.0001	0.0001	0.369	0.416	0.003	0.0001	0.0001	0.0001	0.0001	0.0001
AE3	0.672	0.649	0.454	0.482	-0.057	0.135	0.663	0.673	0.794	**0.802**	**0.822**	**0.842**
Sig(2-tailed)	0.0001	0.001	0.026	0.017	0.792	0.530	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
AE4	0.792	0.790	0.791	0.783	0.048	0.124	0.693	0.708	0.736	0.732	0.786	0.799
Sig(2-tailed)	0.0001	0.0001	0.0001	0.0001	0.846	0.612	0.001	0.001	0.0001	0.0001	0.0001	0.0001
AE5	0.793	0.721	0.733	0.708	-0.108	0.074	0.596	0.595	**0.806**	**0.823**	**0.802**	**0.804**
Sig(2-tailed)	0.0001	0.0001	0.0001	0.0001	0.530	0.669	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
AE6	0.738	0.702	0.664	0.654	-0.145	0.091	0.447	0.523	0.627	0.645	0.641	0.677
Sig(2-tailed)	0.0001	0.0001	0.001	0.001	0.508	0.681	0.033	0.009	0.001	0.001	0.001	0.0001
AE7	0.746	0.696	0.621	0.596	-0.077	0.109	0.625	0.527	0.735	0.698	0.768	0.731
Sig(2-tailed)	0.0001	0.0001	0.001	0.001	0.708	0.596	0.001	0.006	0.0001	0.0001	0.0001	0.0001
AE8	**0.820**	0.793	0.668	0.653	-0.045	0.016	0.776	**0.809**	0.744	0.737	0.780	0.799
Sig(2-tailed)	0.0001	0.0001	0.007	0.008	0.873	0.954	0.001	0.0001	0.001	0.002	0.001	0.0001
AE9	0.682	0.671	0.509	0.497	0.047	0.047	0.435	0.354	0.532	0.473	0.556	0.482
Sig(2-tailed)	0.007	0.009	0.063	0.071	0.874	0.873	0.120	0.215	0.050	0.088	0.039	0.081

** correlation is significant at the 0.01 level (2-tailed).
*correlation is significant at the 0.05 level (2-tailed).

DISCUSSION AND CONCLUSION

In this study, the Multi-faceted Rasch Model was run to calculate the 'fit' values and severity for each rater as well as 'fair average' (FA) for all AEs. The overall performance of the AEs was very good and they rated consistently throughout the process.

This study then investigated the reliability and validity in rating scripts for standardisation purposes in onscreen marking. For checking the reliability in AEs' ratings, each AE's ratings were correlated against the ratings of the expert panel using Pearson's r and Spearman's ρ. The levels of correlation ranged from high to very high, showing that all Assistant Examiners showed high consistency in marking according to the standards required. The very good performance of these AEs gave indications of recruiting prospective raters since all the AE's with high consistency had: 1) experience in rating large scales of written assessments; 2) knowledge of the examinee, i.e., students' language ability and their background knowledge; 3) substantial relevant teaching experience; and 4) 'subject' (ESL/EFL) training. Also, these AEs were given 70 scripts or less (each script with about 80 words) to rate in a three-hour session. All of theses indicated that they did not show signs of rater fatigue.

For the external validity check of the scripts rated by the AEs, we used VQM through correlating relevant VQM for each rating criterion against the students' fair average (FA) scores for each criterion for all raters as obtained from Rasch analysis. Essentially, FA scores evened out rater differences by iterative measures and gave us interval level data which

should be close to the students' true score assuming the measures were valid and that most raters were not idiosyncratic. All the VQM, except lexical variation (the problems with which were discussed in Cheung (forthcoming) and the following paragraph), produced high correlations against their FA score counterparts (from 0.556 for 'tokens' to 0.784 for 'number of meaningful clauses'). This seemed to indicate that when using the scales for this, the raters were estimating the same things which were counted and calculated by the VQM. Hence, the scores used in the scripts for standardisation were valid and reliable. However, it was important to note that VQM also correlated against FA score figures for rating the criterion other than those that they were supposed to measure. For example 'families of words' for 'Content' showed even higher correlations (0.847) against the FA than for 'Language' (0.758). This finding is not surprising since 'Vocabulary' is an aspect of language which allows us to organise information both in the sense of explicit cohesive ties and in the sense of strategic placement of related lexis i.e., lexical cohesion as described in Halliday and Hassan (1976).The most likely explanation for this phenomenon is that 'Content' and 'Language' constructs in written language at this key stage (Grade 6) was heavily dependant on the number of words students used and how well students could use the words under the same topic. Alternatively, it could simply be evidence that students were acquiring the various components of English aspects of language at roughly equal rates.

When we studied the individual sub-constructs of VQM for 'Content' and 'Language', similar results were found when correlating them with fair average (FA) scores and with individual AEs' observed scores. The only really problematic VQM as was 'lexical variation' which showed a small negative correlation figure of -0.036. The finding on lexical variation echoes the concerns raised by Iwashita et al., (2001), Richards (1987) and Vermeer (2000) on the use of ratio measures for lexical variation such as that used in this study.

'Syntax' is a fundamental organising principles of language; therefore, it is scarcely surprising students who can organise their syntax well (as indicated by high T-unit values) are going to get good ratings for 'Content'. Having said that, it may well be possible to produce a better system i.e. home grown version of 'syntactic complexity' for written English which parallels the system which Cheung (forthcoming) developed for spoken English in Secondary 3 TSA oral presentations. This system produced correlations of 0.87 ($\alpha<0.05$) against FA for 'Vocabulary and Language Patterns', correlations of 0.87 ($\alpha<0.05$) against FA for 'Ideas and Organisation' and correlations of 0.89 ($\alpha<0.05$) against FA for 'Pronunciation and Delivery'. Cheung's (forthcoming) system drew heavily on experience knowledge of local teachers concerning the order in which Hong Kong students acquired syntax and the errors typical of Hong Kong students at various levels of schooling. In the meantime, building up students' power in lexis and encouraging students to use and acquire written English should be major areas of pedagogic concern. On the whole, all AEs were estimating values for constructs highly similar to those measured in VQM, except for 'lexical variation'.

Generally it was concluded that the scores used in the scripts for standardisation were valid and reliable. Hence, we can say that the use of 'overlapping marking' run by Multi-faceted Rasch programme proved successful. This method was cost effective and saved at least 75% of the time which would otherwise have been required in marking. In future, verification of ratings by the expert panel only needs to be done for the scripts with decimal FA scores ending in 0.5, 0.6 and 0.7 which are rounded **up** to the next full digit. Hence, the expert panel can save at least 50% of time on verification of FA ratings.

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.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford 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.

Burns, R. B. (2000). *Introduction to research methods*. (4th Ed.). Longman: Person Education Australia Pty Limited.

Cheung, K.M.A. (forthcoming). *An analysis of reliability and validity in the Secondary 3 oral presentation rating scale.* PhD thesis in progress. Macquarie University, Australia.

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.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in second language acquisition 11*, 367-83.

Foucault, M. (1974). *Power/knowledge: selected interviews and other writings*. New York: Pantheon Books.

Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. Longman Group Ltd: London.

Heatley, A., Nation, I. S. P. & Coxhead, A. (2002). *RANGE and FREQUENCY programs*. http://www.vuw.ac.nz/lals/staff/Paul_Nation

Hunt, K. (1965). Grammatical structures written at three grade levels. *NCTE Research Report No. 3.* Champaign, IL, USA: NCTE.

Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language assessment quarterly 3,* 151-69. Lawrence Erlbaum Associates, Inc.

Iwashita, N., McNamara, T. & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language learning 21*, 401-36.

Linacre, J.M. (1989-2008). *FACETS, Version 3.63* (computer programme).

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

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English Assessment system.* Princeton, N. J.: Educational Testing Service.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in second language acquisition 21*, 109-48.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics 4*, 492-518.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of child language 14*, 201-09.

Skehan, P. & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language learning, 49*, 93-120.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language testing 17,* 65-83.