Hong Kong Teachers' Centre Journal《香港教師中心學報》, *Vol. 8*

What are raters estimating: How much do ratings on scale criteria really reflect the characteristics of student performances in terms of the various components of the criteria

能力評審員如何作出判斷：評審準則的等級如何能在各項評審範疇中反映學生不同的表現

CHEUNG Kwai Mun, Amy

Hong Kong Examinations and Assessment Authority

Abstract

This paper aims to explore rater behaviour in assessing oral presentations using verifiable quantitative measures (VQM) as an external validity check on ratings. Twelve raters from a range of backgrounds were recruited to rate 115 Secondary 3 student oral performances in 'individual presentations'. These performances were drawn from a sample of 10 schools participating in a pretest conducted for the commencement of the oral component of Hong Kong's Territory-wide System Assessment in 2006. About 20 students were drawn from each school in three ability categories: low, medium and high. Students were selected based on their internal examination results. Fifty-eight of the 115 student performances were transcribed and assessed on VQM for 'ideas and organisation', 'vocabulary and language patterns' and 'pronunciation and delivery'. VQM results were correlated (Spearman's ρ and Pearson's 'r') against fair average scores derived from Rasch analysis of ratings. The resultant correlations ranged from 0.6 to 0.9. It was concluded that raters were estimating values for constructs highly similar to those measured in VQM.

Keywords

language assessment, oral testing, verifiable quantitative measures, external validity

摘要

本研究是以可驗證的量化量度方法，探究英文科說話評審員在說話評估「個人短講」中的評審表現。是次研究，邀請了12位具有不同學歷及經驗的人士擔任說話評審員，對115位中學三年級學生進行說話能力測試，評審學生的說話能力。這些學生是來自於參加2006年「全港性系統評估」預試的學校。參加預試的學校共有10所，其中包括不同能力組別的學生。每所學校都是根據校內的英文科成績(高、中、低)，挑選20位學生參加說話能力測試。是次研究是從115位參加測試的學生中，抽取了58位學生進行研究，以文字紀錄了這些學生在個人短講中的內容，再以可驗證的量化量度方法(VQM)，評估學生在「內容和組織」、「詞彙和句式」和「發音和表達」三方面的表現。本研究亦利用史比爾曼「ρ」及皮爾遜「r」計算VQM數據和羅許平均值的相關系數。是次研究所獲得的相關系數為0.6至0.9。總而言之，說話評審員給予評級的建構值和VQM所得的結果甚為相近。

關鍵詞

語文評估，說話測試，可驗證的量化量度方法，外部效度

Background

The assessment of spoken language ability relies heavily on the subjective judgments of raters and their interpretation of the rating scales. To complicate matters further, some rating scale constructs are composite entities, for example, the IELTS 2007 uses a construct called 'Fluency and Coherence' which is clearly composite. As Fulcher (2003, p.12) points out 'the key indicators of fluency are speech rate and speech continuity. The key indicators of coherence are logical sequencing of sentences, clear marking of stages in a discussion, narration or argument, and the use of cohesive devices within and between sentences'. The oral rating scale used in this study also involved a number of composite entities for each construct. For example, one of the constructs, 'Vocabulary and Language Patterns', consisted of four sub-constructs: lexical variation, vocabulary richness/ token index, syntactic complexity and grammatical accuracy. This conflation of constructs into a single rating criterion is common, perhaps because it saves time and makes the rating scale more 'user friendly' for raters. While this kind of conflation may have practical reasoning behind it, we do have to realise that raters may be differentially influenced by the various constructs which reside within one rating criterion. For instance, we might have three raters assessing 'Vocabulary and Language Patterns'. One rater may be primarily influenced by grammar, another by vocabulary range and another by language patterns. Furthermore, some raters may be influenced by constructs belonging to a criterion completely 'other' to the one they are supposed to be rating. Therefore, in order to really examine the validity of ratings it is important to ensure that rating scale criteria consist of sub-constructs with features which can be calculated by verifiable quantitative measures (VQM), for example, grammatical accuracy. These VQM can then be

correlated against the Rasch derived 'fair average' (FA) of ratings of a given criterion which 'iterates out' rater idiosyncrasies. Then, these correlations can be squared to provide variance estimate to show how much the various aspects of the students' verifiable performance affect the raters. This paper thus aims to explore rater behaviour in assessing oral presentations using VQM as an external validity check on raters' ratings. Nowadays, it is somewhat politically incorrect (i.e., non PC) to use the word 'objective' but we need to describe things which go beyond the subjective (and much as it might upset postmodernists there are such things.) Therefore, we are going to borrow a word 'trans-subjective' from Foucault (1974, p.94) and use it where once upon a time we would have used the word objective. In short, we seek to answer the question – 'Are raters really rating some trans-subjective' aspects of student performances?'

Literature review

In the field of testing oral 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. One such measure is the length of T-units and the number of clauses per T-unit, a measure of syntactic complexity. Iwashita (2006, p.162) cites this as the best predictor of learner proficiency. 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).

After viewing the problems encountered in using measures to analyse the fragments and ellipsis found in speaking assessment data, Foster, et al., (2000) suggest that the analysis of speech units (AS-units) should consist of 'an independent clause or sub-clause unit, together with any subordinate clause(s) associated with either' (p.365). The coding of AS-units is complicated and therefore very few studies have used the analysis of AS-units. A number of other VQM have been suggested in Wolfe-Quintero et al's thorough meta-study of fluency, accuracy and complexity measures of L2 writing proficiency (1998, p.119). These include words per T-unit. 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'.

A variety of VQM have also been suggested for spoken data: words per clause, words per error-free T-unit, clauses per T-unit, dependent clauses per clause, word type measure, sophisticated word type measure, error-free T-unit per total T-units and errors per T-unit. Iwashita et al., (2001), Richards (1987) and Vermeer (2000) have expressed concerns that the use of ratio measures is problematic since spoken language is short and that, therefore the difference between (the amount of clauses and T-units produced by) high level learners and lower level learners will be cancelled out. Moreover, Harrington (1986) has stated that the usefulness of T-unit as a measure of oral proficiency is limited. However, 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).

Some of the aforementioned features are investigated in the study by Banerjee, et al., (2007) and used by Hawkey & Barker (2004) in their 'intuitive approach to re-marking', i.e. using 'syntactic complexity' and 'vocabulary richness' to measure 'sophistication of language', 'grammatical accuracy' to measure 'accuracy' and use of 'cohesive devices' to measure 'organisation and cohesion'. Where ratings of student performance concern measurement of coherence, as in Halliday and Hasan (1976) and Kennedy and Thorp (2002), special problems arise in creating 'countable' measures of coherence. These measures have been correlated against judge ratings in a number of studies with mixed results. Halliday and Hasan (1976) argue that coherence within a text depends on five categories of cohesive ties: reference; ellipsis; substitution; conjunction; and lexis. Since then, a common approach to quantification of coherence has been to analyse the number of occurrences in form and context of use of connectors. Kennedy and Thorp (2002) further suggest that test-takers at the lower IELTS band levels have a higher chance of using explicit linking devices than test-takers at higher IELTS band levels. Only a very weak relationship has been found between the overt use of linking words and test-taker performance in recent research (Ghazzoul, in progress). The findings of the study by Banerjee, et al., (2007) show that the nature of the task determines test takers' use of demonstratives and test takers at higher levels of language proficiency seem to use fewer demonstratives and rely more on other types of cohesive ties. Writers at higher IELTS band levels are expected to use lexical ties to create cohesion and so display more lexical variation, which is assumed to indicate higher sophistication.

In terms of fluency, Fulcher (1996) employed correlations against 'countable' measures in his quantitative design by first using discourse analysis and counting the occurrence of a range of fluency features. He followed up by using multiple regressions to identify which fluency features significantly predict examinee scores as given by raters. The situation becomes even more complex when we remember our earlier point about that most language rating criteria are actually composites born of convenience. Finally, there is the issue of combining scores on various rating criteria into a unitary score. Unitary score production is a major question for investigation since as Douglas (1994) points out similar scores may represent qualitatively different performances and as Lumley and Quian (2001) point out that grammatical accuracy has the strongest (perhaps disproportionately so) influence on test scores. The study by McNamara (1990) indicated that even on the design of the Occupational English Test (OET), (an Australian test for professionals, where grammatical accuracy was officially downplayed), the Item Response Theory based analysis indicates that raters' perception of the grammatical and lexical accuracy of candidates' performances played an important part in determining their total scores.

Methodology

1. Recruiting twelve raters from a range of backgrounds

Twelve raters were recruited to rate 115 student performances in four batches over a two-week period. These raters were required to complete oral assessors' training in distance mode. This involved performing trial marking after watching exemplar clips. The raters included four local English teachers, four native English speaking teachers and four naive English speakers (who lived in English speaking environments and seldom had contact with non-native speakers).

2. Collecting sample student performances

The 115 performances were drawn from a sample of 10 schools participating in the pretest which was conducted for the commencement of oral component of the Territory-wide System Assessment in 2006. About 20 students from each school were collected in three categories: low, medium and high. Category assigned was based on students' internal examination results.

3. Deriving verifiable quantitative measures from sub-sample (N=58)

This study made use of the categories developed by Hawkey and Barker (2004) and some features investigated relate directly to these categories, as shown in Table 1.

Table 1. Comparison of Hawey and Barker (2004)/CSW Target Features and Those in the Present Study

Hawkey and Barker (2004)/ CSW features	**Features investigated in the present study**	**Assessment criteria in the present study**
Organisation and Cohesion	- Meaningful Clause - Syntactic Complexity	Ideas and Organisation
Sophisticated of Language	- Token Index - Lexical Variation - Syntactic Complexity	Vocabulary and Language Patterns
Accuracy	- Grammatical accuracy	
	- Pronunciation accuracy - Stress - Intonation - Fluency	Pronunciation and Delivery

Fifty-eight of the 115 performances were 'counted' for all aspects of the criteria. The verifiable quantitative measures (VQM) for 'Ideas and Organisation' (IO) included 'number of meaningful clauses' and 'index of syntactic complexity'. VQM for 'Vocabulary and Language Patterns' (VL) consisted of 'grammatical accuracy index' and 'token index/ vocabulary richness', 'lexical variation' and 'index of syntactic complexity'. VQM for 'Pronunciation and Delivery' (PD) consisted of 'pronunciation accuracy index' and 'fluency index' as well as 'stress accuracy index' and 'intonation accuracy index'. (VQM for the three criteria are discussed in detail in Cheung, forthcoming a.)

The data for all VQM were categorised and counted by the Researcher and verified by experts who also had a strong background in grammar and L2 errors and were familiar with the errors typical of Hong Kong students.

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

For the sub-sample of 58 student performances which had been subjected to verifiable quantitative measures (VQM), correlations were done between 12 raters' 'fair average' scores (derived from Rasch analysis, Linacre, 1991-2007) and the VQM derived indices using both Pearson's 'r' and Spearman's 'ρ' as cross checks against each other.

Limitations

Although the verifiable quantitative measures provide a valuable 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 (58 out of 115 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 used.

Findings

To answer the question: 'Are raters really rating some trans-subjective' aspects of student performances?', correlations (r and ρ,) were done between fair average (FA) scores on all three assessment criteria, i.e. 'Ideas and Organisation' (IO), 'Vocabulary and Language Patterns' (VL) and 'Pronunciation and Delivery' (PD) with verifiable quantitative measures (VQM) relating to these criteria.

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, 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

Using the aforementioned interpretations for correlations, according to Table 3, among the VQM of sub-constructs on IO, 'index of syntactic complexity' ('r' value of 0.895) had very high correlation, followed by 'meaningful clauses' ('r' value of 0.835). 70% to 80% of variance was explained meaning that student performances were mostly influenced by these two VQM.

Table 3. Correlations 'r' and 'ρ' of VQM of 58 Student Performances on 'Ideas and Organisation' with Raters' Fair Average Scores

Criteria	Ideas and Organisation					
VQM	No. of Meaningful Clauses		Index of Syntactic Complexity		Combined Indices of IO	
Correlation	r	ρ	r	ρ	r	ρ
FA of IO	0.835	0.858	0.895	0.923	0.881	0.900
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
FA of VL	0.795	0.810	0.869	0.893	0.847	0.859
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
FA of PD	0.795	0.809	0.877	0.905	0.851	0.866
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Overall FA	0.814	0.837	0.887	0.915	0.866	0.885
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

* ρ<.05, ** ρ<.01, 2-tailed

For 'vocabulary and language patterns' (VL), 'syntactic complexity' ('r' value of 0.869) had the highest correlation among the four VQM, with 76% of variance explained. 'Token index/vocabulary richness' ('r' value of 0.850, with 72% of variance explained) and 'grammatical accuracy index' ('r' value of 0.862, with 74% of variance explained) also had 'high' correlation with the criterion while 'lexical variation' ('r' value of -0.204, with 4% of variance explained) had negative correlation. 72% to 74% of variance was explained meaning that ratios of VL in student performances were primarily influenced by 'token index' and 'grammatical accuracy index'.

Table 4. Correlations 'r' and 'ρ' of VQM of 58 Student Performances on 'Vocabulary and Language Patterns' with Raters' Fair Average Scores

Criteria	**Vocabulary and Language Patterns**									
VQM	Grammar Accuracy Index		Token index		Lexical Variation		Index of Syntactic Complexity		Combined Indices of VL	
Correlation	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
FA of IO	0.902	0.906	0.891	0.897	-0.247	-0.348	0.895	0.923	0.883	0.888
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.061	0.007	0.0001	0.0001	0.0001	0.0001
FA of VL	0.862	0.866	0.850	0.857	-0.204	-0.290	0.869	0.893	0.860	0.858
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.124	0.027	0.0001	0.0001	0.0001	0.0001
FA of PD	0.860	0.867	0.846	0.858	-0.227	-0.302	0.877	0.905	0.852	0.858
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.087	0.021	0.0001	0.0001	0.0001	0.0001
Overall FA	0.880	0.888	0.867	0.877	-0.228	-0.306	0.887	0.915	0.870	0.880
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.085	0.019	0.0001	0.0001	0.0001	0.0001

* $\rho<.05$, ** $\rho<.01$, 2-tailed

Among the VQM of students' performance in PD, the correlation of 'pronunciation index' ('r' value of 0.852) was the highest and it was considered to be 'high', with 73% of variance explained. The second highest in correlation was 'fluency index' (0.843) with 71% of variance explained and the variance could be the errors found in fluency, such as hesitations, repetitions, extra fillers and pauses. For 'stress index', the correlation was '0.748' with about 56% of variance explained. 'Intonation index' had correlation of 0.720 (the lowest compared among VQM), meaning that only 52% of the variance in PD ratings was explained by errors in intonation.

Table 5. Correlations 'r' and 'ρ' of VQM of 58 Student Performances on 'Pronunciation and Delivery' with Raters' Fair Average Scores

Criteria	**Pronunciation and Delivery**									
VQM	Pronunciation Index		Fluency index		Stress Index		Intonation Index		Combined Indices of PD	
Correlation	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
FA of IO	0.896	0.899	0.891	0.895	0.774	0.754	0.690	0.607	0.901	0.913
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
FA of VL	0.853	0.856	0.844	0.849	0.738	0.681	0.722	0.620	0.875	0.863
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
FA of PD	0.852	0.860	0.843	0.853	0.748	0.691	0.720	0.623	0.877	0.869
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Overall FA	0.873	0.879	0.864	0.874	0.759	0.726	0.720	0.629	0.892	0.895
Sig. Level	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

* ρ<.05, ** ρ<.01, 2-tailed

The combined indices of all three assessment criteria in Tables 3 – 5 had high correlations with the fair average (FA) scores from the 12 raters' ratings of the three criteria using both Pearson's 'r' and Spearman's 'ρ'. For example, the correlation of the combined indices of IO and fair average of IO was very high ('r' value of 0.881), followed by the combined indices of PD and fair average of PD ('r' value of 0.877) and then by the combined indices of VL and fair average of VL ('r' value of 0.860). Interestingly, the combined indices of PD had higher correlation with the other two fair averages, i.e. fair average of IO ('r' value of 0.901) and fair average of VL ('r' value of 0.875) than the combined indices for IO ('r' value of 0.881) and VL ('r' value of 0.860). This probably reflects the fact that when the students were unable to make their presentation intelligible because of poor syntax or poor pronunciation and delivery, the raters could not make judgments on their ideas or language patterns. However, it is probably safe to say that the 'high' levels of correlations between combined indices and fair average scores indicated that the ratings on the scale criteria did really reflect the characteristics of student performances in terms of the various components of the criteria.

Among all the VQM of sub-constructs in Tables 3 - 5, except for 'lexical variation', all the other VQM had 'high' correlations with the fair average of their respective assessment criteria, meaning they had strong influence in the ratings of their own construct. Moreover, 'syntactic complexity', 'grammatical accuracy', 'vocabulary richness/token index', 'pronunciation index' and 'fluency index' did not only have strong influence (correlations >0.8) in the ratings of their respective construct but also in the other constructs. For example, 'grammatical accuracy' not

only had strong influence on its own construct – 'Vocabulary and Language Patterns' ('r' value of 0.862, 74% of variance explained) but also on 'Ideas and Organisation' ('r' value of 0.902, 81.3% of variance explained) as well as 'Pronunciation and Delivery' ('r' value of 0.860, 74% of variance explained). While results for raters on the whole show very encouraging correlations against VQM, the results for individual raters showed lower correlations in the range 0.59 to 0.89 (correlations between ratings from individual raters and VQM are discussed in detail in Cheung, forthcoming b).

Discussion and conclusion

We began our investigation of external validity using VQM through correlating relevant 'combined indices' for each rating criterion as obtained from VQM 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. On the face of it, the rating scales seemed to function well. All the VQM derived indices produced the very high correlations against their FA score counterparts (0.881 for IO, 0.860 for VL and 0. 877 for PD). 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 derived indices. However, it was important to note that VQM derived indices also correlated against FA score figures for rating criteria, other than those that they were supposed to measure. For example the combined indices for IO also showed high correlations against the FA for VL and PD (0.847 and 0.851). Moreover, the combined indices for VL showed high correlations against the FA scores for IO and PD (0.883 and 0.852) respectively. This could be seen as evidence that our raters were not focusing on the rating criteria i.e. that they were rating extraneous criteria. However, the most likely explanation for this phenomenon is that grammatical accuracy in spoken language is heavily dependant on pronunciation and that organization of ideas is dependant on both grammatical accuracy and on paralinguistic features such as stress and intonation which are subsumed under pronunciation and delivery. Alternatively, it could simply be evidence that students were acquiring the various components of English aspects of language at roughly equal rates.

Generally it was concluded that the whole raters were estimating values for constructs highly similar to those measured in VQM. 'Syntax' is a fundamental organising principle of language; therefore, it is scarcely surprising students who can organise their syntax well are going to get good ratings for IO. Likewise 'vocabulary' is another aspect of language which allows us to organise information and also another component of VL. It is thus hardly surprising if IO ratings correlated well with that VL index consisting of 'syntactic complexity', 'grammatical accuracy', 'vocabulary richness (i.e. token index)' and 'pronunciation' is a fundamental tool for the realisation of 'vocabulary'. Nor should we be surprised that VL correlated so well against combined indices of PD. Pronunciation, intonation and stress are the tools with which people mark lexical and grammatical distinctions in their speech. In fact, it would be strange

if there was not a high correlation between VQM of PD and ratings of VL.

When we studied the individual components of VQM for VL, we found that the 'grammatical accuracy', the 'token index' and the 'index of syntactic complexity' all correlated with FA scores for VL. This was as it should be, given the foregoing discussion. The only really problematic VQM was 'lexical variation' which showed small negative correlation figures in the range from -0.204 to -0.348 with the FA scores for the measures of the various rating criteria. This comes as no great surprise. The expert raters and teachers who devised our rating sales told us that students only used words they had been taught in school. We were warned that longer students' utterances would result from the recycling of familiar vocabulary. Therefore, since 'more able' students produced longer utterances, their ratio of new vocabulary to total number of words in fact would be smaller than that of the 'less able' students who were able to avoid repetition of vocabulary by virtue of their shorter utterance length. Therefore, the more able the student was, the worse would be his/her result on 'lexical variation'. The finding echoes the concerns raised by Iwashita et al., (2001), Richards (1987) and Vermeer (2000) on the use of ratio measures since the spoken language is short and the difference between (the amount of clauses and T-unit produced by 'high' and 'low' level learners) will be cancelled out. However, the results with 'lexical variation' may also indicate that students at the level investigated (end of key stage 3) did not exhibit much lexical variation. In other words, the students at large had a limited range of vocabulary regardless of level and that the distinction between high and low level learners was simply one of facility within a shared body of lexis. This would scarcely be surprising given that for most students in Hong Kong's local school system, English is a 'foreign' language – the lexis of which they primarily acquire through a standard system of schooling. Further research is needed to see this apparent lack of differentiation between students in terms of lexis acquired is also the case in written English. In the meantime building up students' power in lexis and encouraging students to use and acquire spoken English outside the classroom should be major areas of pedagogic concern.

Banerjee, et al., (2007), have suggested that a more realistic measure of linguistic ability would be to look for the ideal group of measures that, when applied together, produced a learner language profile that could be reliably classified as being at a given level in a predetermined scale (p.246). In this respect, the 'syntactic complexity' used in this study is an area worth exploring and can be further developed so as to indicate the L2 development of students across the three key stages, i.e. Grade 1 to Grade 9. From the findings of this study, 'syntax' had strong influence on raters' ratings in the three assessment criterion although this study did not use T-unit length as an index of syntactic complexity. This is similar to the findings of Iwashita (2006), who noted that syntactic complexity was a good predictor of oral proficiency. In short, we can say that taken as a whole, (despite the problems relating to lexical variation measures), raters really are rating some trans-subjective aspects of student performances. We can further conclude that our VQM (with a few exceptions) seem to have tapped

what the raters are really responding to. The core caveat here is 'taken as a whole'. By this, we mean looking at Rasch fair average for a group of raters. 'r' values were in a very healthy range (0.8 – 0.9) indicating that factors accessible to VQM were explaining most of the variance in ratings. However, when we come down to individual raters, we find a much lower (yet still healthy) correlations against VQM indicating the importance of the 'smoothing' function of Rasch iteration in the production of a fair average. Yet even with regard to individual raters here we find that syntactic complexity was the 'king' of the VL indices and also had a powerful effect on IO ratings. Therefore, those teachers hoping to improve students' oral proficiency ratings, especially 'individual presentations' in TSA need to realise that teaching of spoken syntax and grammar is of prime importance and should be systematically taught in schools so that students can master fundamental skills in order to progress to more complex skills. In other words, schools need to avoid exposing students to advanced structures before they have mastered the simpler structures which underlie them (Pienemann, 1998).

Acknowledgement

The author thanks the Hong Kong Examinations and Assessment Authority for providing student performance video clips to this Study.

References

Banerjee, J., Franceschina, F. & Smith, A. M. (2007). *IELTS Research Reports Volume 7*. IELTS Australia and British Council.

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

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

Cheung, K.M.A. (forthcoming b). *How indicative are verbal aloud protocols? PhD thesis in progress*. Macquarie University, Australia.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition 11*, 367-383.

Douglas, D. (1994) Quantity and quality in speaking test performance. *Language Testing 11*, 125-144.

Foster, P., Tonkyn, A. & Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics 21*, 354 – 374.

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

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing 13*, 208-238.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson Longman.

Ghazzoul, N. (in progress). *Coherence in English academic writing of Arab EFL learners with special reference to Syrian and Emirati university students*. PhD thesis in progress. Lancaster University. U.K.

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

Harrington, M. (1986). The T-unit as a measure of JSL oral proficiency. *Descriptive and Applied Linguistics 19*, 49-56.

Hawkey, R. and Barker, F. (2004). Developing a common scale for the assessment of writing. In *Assessing Writing 9*, 122-159.

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

International English Language Testing System (IELTS). (2007). Handbook 2007. Retrieved 1 September 2008, from http://www.ielts.org/generalpages/article 294.aspx

Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly 3*, 151-169. 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*, 401-436.

Kennedy, C. and Thorp, D. (2002). *A corpus investigation of linguistic responses to an IELTS Academic Writing Tasks*. IELTS British Council Research Program.

Linacre, J. M. (1991-2007). *A user's guide to FACETS: Rasch-model computer program*. Version 3.62. Chicago, IL. Winsteps.

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

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition 21*, 109-148.

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.

Pienemann, M. (1998). Developmental dynamics in L1 and L2 acquisition: Processability theory and generative entrenchment. *Bilingualism: language and cognition 1*, 1-20.

Richards, B. (1987). Type/token rations: What do they really tell us? *Journal of Child Language 14*, 201-209.

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.

Wolfe-Quintero, K., Inagaki, S. and Kim, H-Y. (1998). Second language development in writing: measures of fluency, accuracy and complexity. Technical Report 17. Honolulu: University of Hawaii at Manoa, Second Language Teaching and Curriculum Centre.