English Language Teaching; Vol. 9, No. 2; 2016
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

How Do Raters Judge Spoken Vocabulary?

Hui Li[1]

[1] School of Humanities and Social Science, University of Strathclyde, UK

Correspondence: Hui Li, Faculty of Humanities and Social Sciences, Lord Hope Building (4th Floor), 141 St James Road GLASGOW G4 0LT. E-mail: hui.li@strath.ac.uk

Received: October 13, 2015 Accepted: January 10, 2016 Online Published: January 11, 2016

doi: 10.5539/elt.v9n2p102 URL: http://dx.doi.org/10.5539/elt.v9n2p102

Abstract

The aim of the study was to investigate how raters come to their decisions when judging spoken vocabulary. Segmental rating was introduced to quantify raters' decision-making process. It is hoped that this simulated study brings fresh insight to future methodological considerations with spoken data. Twenty trainee raters assessed five Chinese students' monologic texts on vocabulary in this study. Both segmental rating and overall rating were retrieved from the raters. Rasch analysis suggested variation between raters in their judgment of vocabulary, although consistency was found in general. Besides, there was a mismatch between candidates' vocabulary scores and their lexical statistics. The raters' decision-making process was generally cumulative.

Keywords: raters, vocabulary, decision-making process, and oral examinations

1. Introduction

This study sought to shed light on raters' performance in oral proficiency tests from two perspectives: raters' judgement on vocabulary and raters' decision-making process (how do raters reach their decisions). Past studies have shown a good deal of research on vocabulary, and even much more research on raters' performance in oral examinations. In particular, descrepancy between multiple raters' judgement has always been an area of intest since 1920s. Some interesting findings relevant to vocabulary assessment in oral examinations have been identified in the past studies (Brown, Iwashita, & McNamara, 2005; Lorenzo-Dus & Meara, 2005; Malvern & Richards, 2002; Read, 2000, 2005). To date, however, few studies have focused solely on rater performance in assessing vocabulary. As vocabulary is increasingly recognized as an effective predictor of language proficiency, it is especially important to understand how raters judge it. The first purpose of this simulated study is thus to fill this gap.

This study focuses on both actual scores awarded and rating process. Raters are found to vary in assessing linguistic features (Bejar, 2006; Elder, 1993) and frequently deviate from rating criteria (Eckes, 2005; Erdosy, 2003; Taylor & Jones, 2001). May (2006), for instance, found more than 30% of raters' comments subsumed additional features that were not mentioned in the criteria. With reference to vocabulary, there is no consensus in the existing assessment literature on the construct of vocabulary (Meara, 1996; Read, 2000). The potential difference in interpretations of the construct of vocabulary among raters is likely to lead to inter-rater variation on their judgement. Inter-rater variation refers to discrepancy in judgment between raters due to rater characteristics rather than difference between test-takers' performances (Eckes, 2005; McNamara, 1996). In regard to rating process, many belive that subjective ratings are impressionistic assessments made rapidly by raters (i.e. Meara & Babí, 2001), which may partly due to the fact that rating descriptors include vague and impressionistic terminologies (Knoch, 2011); however, little empirical evidence can be found to back up this intuition. It is expected that this study may shed insight into raters' decision-making process to better our understanding on how oral texts are judged by raters.

Most importantly, this study is in an attempt to seek a new mechanism in tracing raters' decision-making in assessing speaking. As is known, the investigation of raters' decision-making process in assessing oral texts has been largely held back by the complications of assessing oral data: temporal auditory input rather than constant visual one i.e. writing (Read, 2005). Methodological errors oftern factor in the results and till now there is no effective research methodology to use with spoken data. With the aim to investigate raters' decision-making process in assessing spoken vocabulary, it is hoped that this study could bring fresh understanding to future methdological considerations with spoken data.

2. Literature review

Rater reliability is found to be one of the two major sources of measurement error in language assessment, which refers to the extent to which the information collected are consistant and reliable (Bachman, Lynch, & Mason, 1995). The subjective nature of rater judgement, although not being regarded as a threat to the validity of tests, has been viewed as a variable of potential bias. Also, inter-rater reliability has been a concern among scholars and widely studied with respect to gender (O'Loughlin, 2002; Lumley & O'Sullivan, 2005), rater background (Chalhoub-Deville, 1995; Elder, Golombek, Weigle, Boldt, & Valsecchi, 2003; Winke, Gass, & Myford, 2013), and rater training (Elder, Barkhuizen, Knoch, & von Randow, 2007; Lumley & McNamara, 1995; Weigle, 1994). More recently, it is generally accepted that rater effects can be controlled and predicted. However, past studies also show that, even if raters display agreement quantitatively (acceptable significant correlations), they may come to their decisions on the basis of various criteria (Douglas, 1994) and may interpret and apply scoring scales inconsistently in oral assessment (Eckes, 2005; Taylor & Jones, 2001).

Let us therefore next consider what is known about how raters assess spoken vocabulary. Spoken vocabulary was mainly used as an embedded construct in oral examinations. Thus to date, what we know of how vocabulary is assessed comes mainly from studies that focus on oral proficiency in general rather than on vocabulary specifically. Some findings include: (1) Raters' judgements on vocabulary do not correlate with lexical diversity (D) in oral proficiency interviews (Lorenzo-Dus & Meara, 2005; Malvern & Richards, 2002); (2) Raters exhibit idiosyncratic approaches regarding saliency of lexical features in assessing vocabulary in oral interviews. They typically make more negative than positive comments on vocabulary (Brown, 2006); (3) Raters' judgements are sensitive to word types, tokens, and difficult words in OPIs (Brown et al., 2005; Lorenzo-Dus & Meara, 2005); (4) Raters have conflicting views on assessing linguistic aspects vis-à-vis pragmatic aspects of vocabulary in oral examinations (Brown et al., 2005); (5) It is difficult for raters to assess vocabulary at adjacent IELTS band levels (Read, 2005); (6) High correlations have been found between subcategories in oral examinations, such as vocabulary, grammar, fluency, etc. (Brown and Taylor, 2006; Malvern and Richards, 2002; Taylor and Jones, 2001); and (7) Vocabulary and grammar were prone to be rated more harshly than other constructs in oral examinations(Galaczi, 2005).

Vocabulary is nowadays widely used as a subscale on analytic rating scales, whereby raters are asked to give a score each to different components. More and more studies find statistically significant relationship between vocabulary, oral proficiency and raters' marks (Brown et al., 2005; Iwashita et al, 2008; Iwashita, 2010; Sato, 2012; Li & Lorenzo-Dus, 2014), Lorenzo-Dus and Meara (2005) examined test-takers' vocabulary in oral proficiency interviews with a focus on whether the amount of support interviewers' provided was related to candidates' vocabulary production. The interviewer in their study reported that she was particularly sensitive to candidates' use of 'difficult words' in her judgments. The same finding was also echoed in Li and Lorenzo-Dus' (2014) study, in which candidates' use of sophisticated vocabulry was found to be the most salient lexcial feature attended to by raters. Brown et al. (2005) and Iwashita's (2010) studies further confirmed that vocabulary was a predictor of candidates' language proficiency. More recently, Lu (2012) investigated the relationship between candidates' spoken vocabulary and their spaking task performance. A number of lexcial measures were used in his study, i.e. types (the number of different words used in a running text) and lexical sophistication (the number of less frequently used words in a running text). In contrast to Lorenzo-Dus and Meara (2005), and Li and Lorenzo-Dus (2014), he found types correlated more closely with raters' judgments than other lexical indices; whereas lexical sophistication did not have a strong correlation with ratings. So far, past literature underpins the importance of vocabulary as a predictor of candidates' oral proficiency and more importantly, it also reveals that our understanding of how vocabulary is judged by raters is mixed. In particular, little do we know about how raters arrive at their decisions when making judgments.

Meara and Babí's (2001) study proposed a new methodology in investigating raters' decision-making process: segmental rating. As the authors state, their paper 'is not intended as a formal analysis of the way examiners work; however, it does explore a methodology which looks promising in this regard and might be developed for use in further investigations of rapid assessments' (Meara & Babí, 2001). In their experiment, written texts are segmented and are presented word by word automatically on the computer to their raters.

Meara and Babí's attempts in segmenting written texts seem to have worked in assessing writing. Firstly, the data obtained from raters successfully quantify their decision-making points. Secondly, the significant correlations from two groups of raters, to some extent, prove the validity of the methodology. Finally, the findings echo what we have found about raters' cognitive decision-making processes in assessing writing. For example, the authors suggest that even if assessors can arrive at a reliable agreement, their reasons may vary. Hence, this study offers a good example of how this methodology works with raters' data in written contexts.

Meara and Babí's study provides a new approach to exploring raters' decision-making processes. Although using raters' concurrent verbal reports (Cumming, Kantor, & Powers, 2002) is now a prevalent approach to investigating raters' decision-making process when assessing writing, it is difficult to use such a method in assessing speaking. As Brown et al. (2005) notice,

'One limitation of verbal-report studies involving assessments of speaking proficiency rather than writing proficiency is that the real-time nature of the assessment precludes the elicitation of concurrent reports, and limits, therefore, what can be inferred about the process of rating, as opposed to the performance features to which raters attend.'

One primary difference between assessing oral and written data lies in the fact that the former is passing and fleeting; whereas the latter is tangible and solid. The former demands cognitive ability to decode and memorize the audio data, thus invariably distracts raters' attention from verbalising their thoughts, which is equally a highly cognitive demanding task. For that reason, most researchers working with oral data take the approach of retrospective verbal reports. That is to say, verbal reports were collected after a task (May, 2006; 2009). One disadvantage of this approach, obviously, is that the information collected may be not 'complete' or 'real' since it is retrieved from short term memory and it may have been refined after the whole task is finished. Although some studies show the correlation between concurrent verbal reports and retrospective verbal reports is acceptable, the reliability of the latter remains debatable.

Segmental rating offers a workable alternative. Obviously the same method will not apply directly to the oral data introduced in this experiment; nonetheless segmental rating presents a new method of investigating raters' decision-making process. The rationale underlying segmentation of texts is that by tracing the change in segmental scores within a text, it may be possible to outline and quantify raters' step-by-step decision-making. Based on this assumption, and building upon Meara and Babí's (2001) model, segmental rating is adopted in this experiment with adaptations to the segmentation method (see Method). In this way, raters' decision-making process is explored via examining segmental scores to see how their final decisions are formed.

Given the above two aims, and in terms of research questions, the following qestions are addressed in this experiment:

RQ1. Do raters agree with each other in assessing vocabulary? If so, to what extent?

RQ2. How do raters come to their final vocabulary scores?

3. Method

3.1 Texts

Different five Chinese learners, who were studying at a UK university when the experiment was taken, produced five oral narrative texts. The learners were from two proficiency groups: group A candidates of high ability (MA students) and group B candidates of medium ability (undergraduate students). Among the five spoken texts, three were from group A candidates, and two were from group B. The five texts last about 23 minutes in total. Table 1 shows candidates' lexical statistics, together with their vocabulary scores and IELTS scores that were taken shortly before the experiment. Various lexcial measures were used to quantify candiates' vocabulary output in a story-telling task (Table 2). In the story-telling task (a family picnic story), candidates were asked to tell one story based on six pictures that test students' narrative and descriptive vocabulary.

The five texts were segmented (by the researcher) before being rated. Due to their various lengths, they were chunked into a different number of segments. Segmentation was made on the basis of semantic-based T-units, which are defined by Young and Milanovic (1992) as 'an independent clause and any dependent clauses'. Most segments consisted of either one long utterance or two to four short utterances. A pause was inserted between adjacent segments. The segmented Text 1 had 6 segments, Text 2 had 7 segments, Text 3 had 11 segments, Text 4 had 10 segments, and Text 5 had 16 segments.

As segmental rating is a new method in assessing spoken vocabulary, it was essential to ensure its feasibility and was crucial to ensure that raters would feel comfortable in the way that the texts were segmented. In order to do so, once segmented, the texts were presented for rating in two pilot studies. A specific concern of the pilot studies was to examine whether raters could produce vocabulary scores during the pause between segments.

Two raters participated in both of the pilot studies. Both raters had a similar professional background to that of the raters who participated in the experiment. At the end of each pilot test they were asked the following questions:

(1) Are the texts clear enough (RQ1 and 2)?

(2) Are the pre-test instructions clear so that you know what to do in the assessment (RQ1 and 2)?

(3) Do you feel comfortable with the way the test has been segmented (RQ1 and 2)?

(4) Do you think each segment has sufficient vocabulary information for you to give a mark (RQ1 and 2)?

(5) Is the pause between segments long enough for you to come to a judgement? Or is it too long for you (RQ1 and 2)?

Based on the feedback from the two raters in the pilot studies, two changes were made to the initial experimental design. Firstly, the pause between segments was increased from five seconds to eight seconds so that raters would not feel that they had to give a score in haste. Secondly, the written instructions were revised to make it explicit that raters' task was to assess only one construct: vocabulary. Thus, the term vocabulary was changed to bold and was underlined.

The second pilot study used the revised version of the test. The feedback confirmed that the revised version was rater-friendly.

Table 1. Comparison of candidates' lexical measures and their proficiency scores

Text	Group	Tokens	Types	TTR	D	Plex	K1	K2	Offlists	Vocab scores	IELTS
T1	B	140	66	.47	32.47	0.40	93.71%	6.29%	0%	4	5.5
T2	A	246	119	.48	47.82	0.24	92.00%	6.18%	1.91%	5	7.0
T3	A	315	142	.45	56.84	0.77	91.72%	6.37%	1.99%	4	6.5
T4	A	245	110	.45	52.78	0.42	94.17%	3.75%	2.08%	4	6.0
T5	B	553	181	.33	51.77	0.44	92.86%	5.60%	1.54%	5	5.5

Table 2. Lexical measures and their explanations

Lexical measures	Explanations
Tokens	Number of total words in a running text
Types	Number of total different words in a running text
D	Vocabulary variation (also called lexical diversity): a recent sophisticated approach to compute the ratio of total different words in a text
Plex	Vocabualry sophistication: used to indicate how many difficult words are used in a text
K1	The first 1000 most frequently used words
K2	The second 1000 frequently used words
Off-lists	Not on Nation (1990)'s list of the most frequently used words

3.2 Raters

Twenty raters participated in this experiment. They were studying for the course of the Certificate for English Language Teaching to Adults (CELTA course) at a UK university. All of them were native speakers of English. Their training as language teachers was on going when the experiment was conducted. Since the experiment was embedded within their syllabus, the raters were highly motivated to perform the assessment task. The raters in this experiment were asked to rate vocabulary on a 6-point scale (from 1- poor to 6 -good).

3.3 Test Adminstration

3.3. 1 Research Design

The assessment was conducted during a one-hour workshop within a CELTA course of language testing. The workshop was divided up into two sessions: the first half hour was used for the experiment; the second half was for a seminar. During the latter, raters provided their own understanding of vocabulary assessment in oral contexts, and explained their perceptions of how they had assessed vocabulary in the preceding experiment. A trained oral language assessor and the module lecturer jointly led the group discussion. The seminar was video recorded.

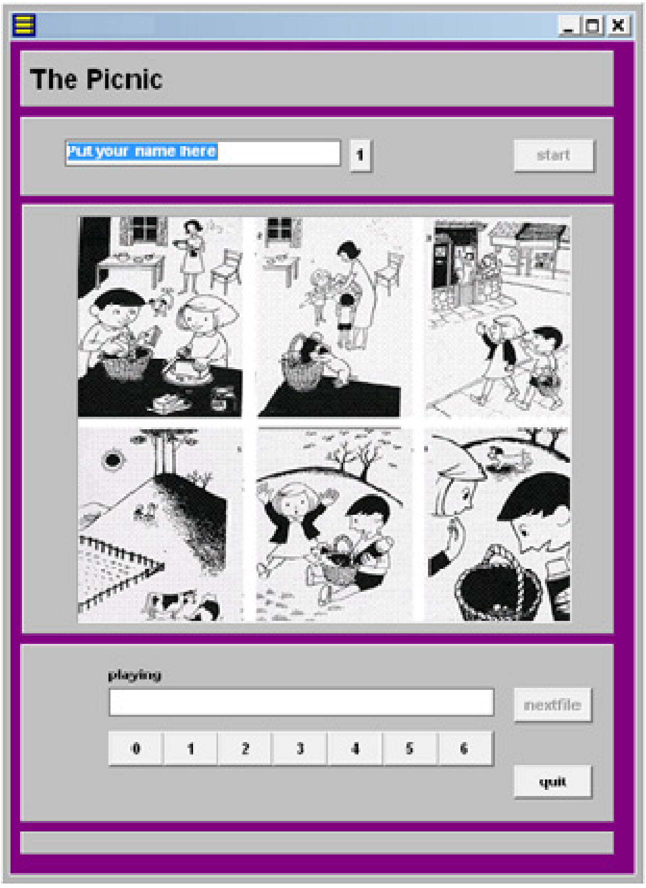

Figure 1. The interface of the computer program in this experiment

The raters assessed the oral texts using a program especially written for the experiment (Note 1). The six pictures in the story-telling task were incorporated into the interface of the program (see Figure 1). Raters listened to the five texts using headphones and marked each segment in turn. Segments in each text were automatically activated after an eight-second pause. During the pause, raters were asked to provide a mark on range of vocabulary for the segment that they had just heard. After the first text had finished, they were required to provide a mark on vocabulary for the whole text. Then, after pressing the button 'nextfile', raters proceeded to the next text. Two sets of vocabulary scores were thus retrieved from raters: segmental scores for every segment within each of the texts and an overall vocabulary score for every text.

Both oral and written instructions were provided to raters. The instructions and marking sheet for overall scores were in the same document issued to raters prior to the study.

3.3.2 Data Analysis

Three statistical approaches were adopted for data analysis. Kendall's coefficient of concordance was calculated to examine the extent of agreement among raters (RQ1). Multifaceted Rasch analysis was used to further explore the degree of rater reliability and rater bias (RQ1). It helps to break down the group performance of raters into individuals so that it is possible to identify 'misfit' raters, and to observe the interactions between various 'facets': texts, segments, and scoring scales (McNamara, 1996). Spearman rank order correlation (one-tailed) was computed between segmental scores and overall scores to investigate raters' decision-making process (RQ2).

Rater reliability coefficient could be caculated as a single number such as Kappa, Spearman correlation, and Kendall's coefficient of concordance as used in this research. This approach helps to reveal the extent to which raters agree with each other, but fails to explain what variables factor in the agreement/disagreement. In order to gauge raters' decision-making process, Rash analysis is also used in data analysis. Rasch analysis has been prevalently used in L2 performance assessment (Bonk & Ockey, 2003; Eckes, 2005; Lumley, 1998; Lumley & McNamara, 1995; McNamara & Knoch, 2012; Weigle & Nelson, 2004; Weigle, 1998). It has been suggested as an effective method to examine raters' performance (Fulcher, 1996). Multifacets Rasch analysis has been seen as showing 'a great deal of promise in finding and accounting for the relative effects of contextual features that we identify' (Hudson, 2005).

Rater agreement was examined by investigating their performance both as a group (Kendall's concordance), and as individuals (Rasch analysis). In addition, the segmental ratings helped to examine raters' consistency in rank-ordering various segments within each text in particular; whereas overall ratings and mean segmental ratings helped to get some idea of their consistency in rank-ordering the 5 texts in general.

Sample size in the analysis was small (five texts). However, Rasch analysis was conducted on segmental scores. Altogether there were 50 observations (segments) and 20 observations per rating-scale category (raters), thus giving 1000 observations for the whole data set, which was well over the restriction of 30 observations per element, and 10 observations per rating-scale category (Linacre, 2006). Rasch analysis was conducted by using FACETS (Minifac facets student evaluation version, Linacre, 2006).

4. Results

4.1 Raters' Assessment on Vocabulary

Statistical analysis was conducted on the basis of two sets of ratings: overall scores and segmental scores. Mean scores of both scores for each text, averaged over raters, are set out in Table 3. The mean ratings ranged from 2.75 (T1) to 4.40 (T3). Apart from large deviation in T5 (sd =1.27), raters seemed to be quite consistent with their overall scores, especially in T1 (sd = .64). Compared with overall scores, the standard deviations were smaller for segmental scores, in particular for T5 (sd=.93).

Table 3. Descriptive statistics of overall and segmental scores from 20 raters for 5 texts

	T 1 (n=6)		T 2 (n=7)		T 3 (n=11)		T4 (n=10)		T5 (n=16)	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
R1	3.2	.98	3.1	.69	3.5	.69	2.9	.87	3.6	.62
R2	3.0	.89	4.1	.69	5.4	.50	5.1	.73	4.9	.80
R3	3.7	.82	3.7	.76	4.2	.60	3.1	.32	3.8	.58
R4	2.2	.75	4.6	.53	4.6	.81	4.8	1.14	4.1	1.15
R5	3.3	.52	3.4	.53	3.9	.30	3.0	0	3.9	.34
R6	3.0	.63	3.7	.49	4.8	.60	3.2	.79	2.0	.52
R7	2.8	.75	3.0	.82	3.4	.67	3.1	.88	2.5	.82
R8	2.5	.55	3.4	.79	5.2	.40	4.8	.42	4.4	.50
R9	2.0	.63	4.1	.69	5.1	.70	3.2	1.03	3.7	1.08
R10	3.5	.84	4.7	.76	4.8	.98	2.7	.95	2.2	1.05
R11	2.5	1.38	1.9	.90	2.8	.98	1.9	.74	1.6	.50
R12	3.0	1.05	3.9	.90	3.3	1.01	2.6	.52	3.1	.57
R13	4.0	1.41	4.7	.76	5.6	.50	3.8	1.23	3.3	.95
R14	2.7	1.11	3.9	.90	3.2	.60	3.0	.47	3.0	.52
R15	1.8	.98	4.4	.98	4.9	1.45	2.1	.88	3.2	1.05
R16	1.8	.75	2.4	.53	2.6	.92	2.5	.97	2.6	.81
R17	3.0	.63	4.3	.49	4.1	1.04	3.9	1.10	4.8	.83
R18	1.8	.75	3.6	.53	4.9	.30	3.1	.57	2.0	.52
R19	3.5	.55	3.1	.69	4.5	.69	3.2	.42	3.1	.77
R20	3.0	.63	4.7	1.11	4.1	.94	3.5	.97	2.8	1.00
MSR	2.82	.65	3.74	.77	4.25	.88	3.28	.85	3.23	.93
MOS	2.75	.64	4.05	.83	4.40	.94	3.10	.91	3.15	1.27

MSR: mean segmental scores; MOS:mean overall scores.

Table 4. Unexpected responses in segmental ratings

Cat	Exp.	Resd	StRes	SEGEMENT	TEXT	RATER
2	3.8	-1.8	-2.0	2a	1	2
3	4.7	-1.7	-2.0	5c	3	17
6	4.2	1.8	2.0	11c	3	15
5	3.2	1.8	2.1	1c	3	9
2	3.8	-1.8	-2.1	8e	5	15
1	2.7	-1.7	-2.1	12e	5	6
5	3.2	1.8	2.1	2e	5	15
6	4.1	1.9	2.1	6c	3	10
5	3.2	1.8	2.1	14e	5	9
1	2.8	-1.8	-2.2	9d	4	15
1	2.8	-1.8	-2.2	1c	3	15
6	4.1	1.9	2.2	5e	5	17
4	2.3	1.7	2.2	4d	4	16
6	4.1	1.9	2.2	2e	5	4
1	2.9	-1.9	-2.2	3a	1	9
1	2.9	-1.9	-2.2	12e	5	10
1	2.9	-1.9	-2.2	15e	5	10
2	4.0	-2.0	-2.3	4e	5	13
6	4.0	2.0	2.3	4e	5	17
5	3.0	2.0	2.3	6d	2	14
6	3.9	2.1	2.3	8d	4	4
2	4.1	-2.1	-2.4	7e	5	13
3	4.9	-1.9	-2.4	6c	3	17
1	3.1	-2.1	-2.4	10e	5	10
1	3.1	-2.1	-2.5	4e	5	10
3	5.0	-2.0	-2.5	2b	2	2
3	1.5	1.5	2.5	16e	5	16
6	3.7	2.3	2.5	2b	2	20
4	2.0	2.0	2.6	16e	5	1
6	3.6	2.4	2.8	1b	2	10
1	3.5	-2.5	-2.8	5a	1	4
1	3.5	-2.5	-2.8	5d	4	15
2	4.5	-2.5	-3.0	3b	2	8
6	3.2	2.8	3.2	2a	1	13
5	2.4	2.6	3.2	2a	1	10
5	1.6	3.4	5.4	5a	1	11

Kendall's coefficient of concordance for both overall ratings and segmental ratings was computed separately. Raters moderately agreed with each other in assessing the five texts (Kendall's W= .464, p<.001), and in assessing the 50 segments in the five texts (Kendall's W=.368, p<.001). In particular, the extreme low level of significance emphasizes the relatedness of the scores awarded by the 20 raters.

Rasch analysis was further conducted to investigate interactions between the various variables in the experiment, namely, segments, texts, and raters for rater variation, which may also result from raters' disagreement on the marking of specific segments. Table 4 displays the unexpected responses from raters on different segments. The unexpected responses listed have standard residuals and raw residuals of more than 2. Raw residual is simply score minus expected value (according to the model); each raw residual is divided by the standard deviation of the complete set of residuals to give a standardized residual (colomn 4).

Table 4 shows that unexpected scores are clustered around a few segments, such as segments 2a, 5a, 2b, 1c, 6c, 4e, 12e, and 16e. Apart from segment 16e, which has the same positive standard residuals from two raters, the rest of the segments observe different positive or negative standard residuals. Let us take segment 5a for example. This segment had the highest standard residuals from rater 11 (std. residual = 3.4). Estimated on his usual harsh style in marking, the expected score for this segment is 1.6. However, surprisingly, the rater marked the segment at 5, which is much higher than the expected score. That means some features in this segment triggered the rater to mark this segment much higher than his normal marking pattern shows. In contrast, rater 4 was very lenient, but he gave the same segment (5a) the lowest score on the scale, 1 out of 6, indicating a negative judgement from this rater on that segment. The disagreement in judgements between the raters is self-evident, revealing that raters might have attended to different lexical features, or different language skills in the same segment, or they varied in judging the same language features. In a similar fashion, segments 2a, 2b, 1c, 6c, 4e and 12e also observe contradicting judgements from raters.

4.2 Raters' Decision-Making Process

Raters' decision-making processes were examined by investigating the relationship between segmental ratings and overall ratings. Spearman rank order correlation (one-tailed) was calculated between the two. Table 5 displays correlations between segmental scores and the overall scores in the five texts separately. The table shows that significant correlations were found for all the segments except for the first one in T2. Also, the highest correlation between segmental score and overall score in each text, highlighted in red, was found either at the fourth (T1, T2, T3, and T5) or the fifth (T4) segment.

Table 5. Correlations between each segmental score and overall score in each text

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16
T1	.649**	.646**	.583**	.743**	.626**	.734**										
T2	.310	.411*	.385*	.725**	.657**	.733**	.671**									
T3	.510*	.604**	.826**	.850**	.646**	.655**	.655**	.757**	.589**	.785**	.462*					
T4	.614**	.773**	.623**	.654**	.832**	.765**	.805**	.710**	.500*	.583**						
T5	.706**	.546**	.646**	.877**	.763**	.529**	.743**	.662**	.569**	.699**	.768**	.825**	.681**	.602**	.818**	.755**

* Correlation is significant at the 0.05 level (one-tailed); S=segment; T=text.

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

Figure 2 further illustrates the correlations between individual segmental scores and the overall score of each text. The graph has been plotted from the first segment in a text to the last segment, demonstrating the relation between raters' assessments at one stage (measured by segments) and their final judgement.

Apart from peaks at about the fourth or the fifth segment for each text, considerable fluctuations were also apparent in the progression of lines in Figure 2. The fluctuations showed that raters kept modifying their decision in the process of listening to the texts. With the exception of T2, the correlations in the rest of the 4 texts, however, displayed little trend with segment numbers despite some fluctuations.

Spearman rank order correlation was further computed between the mean scores averaged over certain number of segments within a text and the overall score of a text to investigate whether the rating process was cumulative. Specifically, overall scores were correlated with mean scores of the first 2 segments (MF2), with mean scores of the first 3 segments (MF3), of the first 4 segments (MF4), and of the first 5 segments (MF5) and so forth in a text (see Table 6).

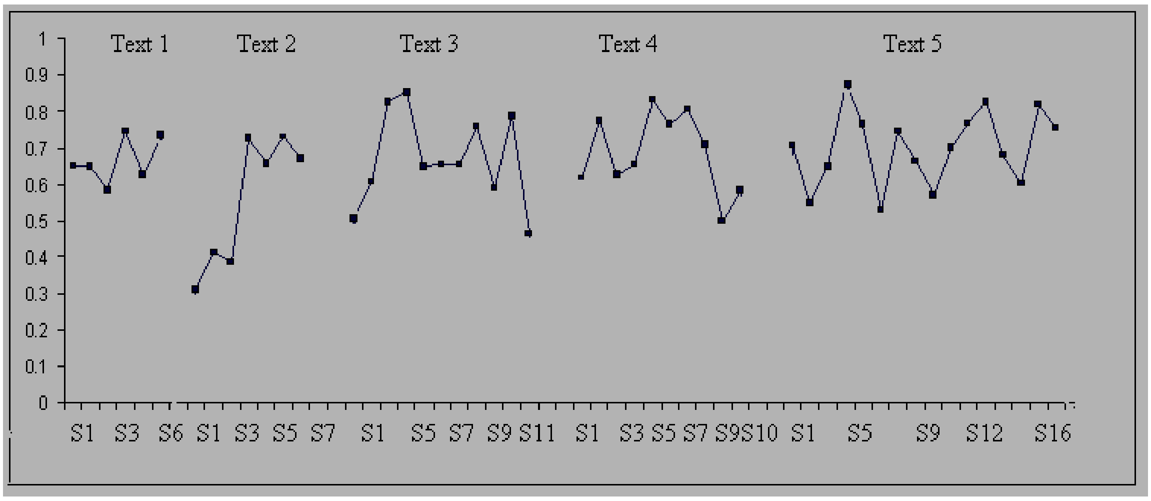

Figure 2. Correlations between segmental and overall score in 5 texts

Table 6. Correlations between mean scores of different numbers of segments and overall scores

	T1 (n=6)	T2 (n=7)	T3 (n=11)	T4 (n=10)	T5 (n=16)
Mean FS2	.760**	.400*	.580**	.729**	.692**
Mean FS3	.809**	.416*	.694**	.727**	.738**
Mean FS4	.832**	.514**	.763**	.760**	.843**
Mean FS5	.840**	.544**	.763**	.807**	.853**
Mean FS6	.884**	.658**	.775**	.825**	.831**
Mean FS7		.682**	.778**	.819**	.845**
Mean FS8			.781**	.831**	.863**
Mean FS9			.768**	.818**	.854**
Mean FS10			.768**	.818**	.848**
Mean FS11			.767**		.845**
Mean FS12					.852**
Mean FS13					.849**
Mean FS14					.837**
Mean FS15					.837**
Mean FS16					.841**

Similar to the results in Table 5 and Figure 2, Table 6 shows that after the first five or six segments, the correlations in most texts increased and/or reduced around a central point. The correlation values in each text continued to improve until they reached a threshold point. The threshold point corresponded to the 6th segment in T1, the 7th segment in T2, the 3rd segment in T3, the 4th segment in T4, and the 3rd segment in T5 (segments are highlighted in Table 6). After that, the correlations dropped a little, and then fluctuated around the correlations between the overall score and the mean segmental score. This suggests that raters' judgements came to be relatively stable after these first four or five segments. It should also be noted that the correlations at the 5th or the 6th segments were closest to the overall scores in all of the five texts (T1 only had six segments).

So far, the correlations between segmental score and overall score for each text revealed the importance of first impressions. The examination of step-by-step mean segmental scores, also, disclosed a tendency for mean vocabulary scores in general to be stable after four or five segments.

5. Discussion

5.1 Inter-rater Agreement

The 20 raters in this experiment showed significant consistency in assessing vocabulary. Raters seemed to be able to reliably discriminate candidates' proficiency levels, which is in line with past studies that vocabulary was a predictor of candidates' proficiency levels (Brown et al. 2008; Lu, 2011). However, Kendall's concordance values in both overall ratings and segmental ratings were just moderate. Rasch analysis further revealed that there were quite a few contradictory judgements on some segments between raters (Table 4). So, despite of the observed agreement between raters, there was unmistakable variation in their judgements.

Firstly, the finding of significant agreement among raters in assessing spoken vocabulary should not be discounted due to moderate correlation values. In particular, statistically significant correlations for segmental scores are very encouraging. The raters were able to rank order the lexical properties of various segments in a relatively systematic way: this shows that 20 raters agreed with each other to some extent on the strength or weakness of various segments produced by the same candidate. Also, we need to remind ourselves that the raters in this experiment had limited rating experiences prior to this experiment. It is sensible to assume that accredited raters, guided with common rating criteria in real life examinations, would have achieved a higher correlation value than that in this experiment. Optimistically speaking, therefore, there is potential for raters to achieve statistically satisfactory significance values in agreement when assessing vocabulary in oral examinations.

Secondly, a close examination of vocabulary scores in each text again points out a possibility of holistic approach in assessing vocabulary. In three out of the five texts (T2, T3, and T5) there was a mismatch between vocabulary scores and lexical statistics (Table 1). For example, T2 had comparatively low lexical statistics, it was rated, however, high in this experiment; specifically, this text had on average the second highest vocabulary score of the five texts. The raters in the experiment, who had little knowledge of the candidate's language level prior to the assessment, provided similar high vocabulary scores to this text as the examiner in the story-telling task did. Vocabulary scores, in this case, were more closely related to the overall language proficiency of the candidate than the lexical statistics of this spoken text. Raters either attended to lexical features that were not examined in this experiment or their judgements on vocabulary were not on the lexical performance of this text per se. Along the similar line, Iwashita et al. (2008) also cautioned that raters' judgments might be from a holistic approach and questioned how distinct the scores of various components were on analytic scales. So, factors other than vocabulary may have affected the raters' judgements on vocabulary, which merits further research.

Finally, vocabulary scores in general were slightly low in my experiment in relation to the candidates' language proficiency levels (Table 3). The average score from the 20 raters for T3 (the strongest high ability candidate) was only 4.4 out of 6. Likewise, Galaczi's (2005) study of the speaking tests for Upper Main Suite examinations (FCE, CAE and CPE) found that vocabulary was harshly rated. Of the four constructs examined, grammar and vocabulary had the lowest mean scores. Galaczi's explanation of this finding was vague: 'these two elements (grammar and vocabulary) in the mark scheme were more noticeable and measurable and as a result marked more harshly by oral examiners' (2005: 16). Similarly, O'Sullivan (2002) found that grammar and vocabulary were most harshly interpreted in writing assessment. Low vocabulary score in this experiment seems to be in line with past literature. However, considering the finding that new raters tend to mark more harshly in general and to produce more extreme scores (Weigle, 1998), it is hard to conclude whether the slight strictness from my raters is due to their general pattern of assessing spoken vocabulary or because they were new raters.

5.2 The Rating Process

In this experiment, evidence of rating processes was mainly retrieved from the correlations between segmental and overall scores in each text from the 20 raters. Two interrelated findings explain raters' decision-making on the assessments of vocabulary of the five texts that they were presented with: (1) they based their overall scores on a cumulative or average impression of the candidates, and (2) they arrived at reliably stable decisions on vocabulary scores quickly (after the 5th or the 6th segment).

The raters' cumulative rating process was demonstrated by the strong correlations found between mean segmental scores and overall scores. This finding seems to be in contrast to the belief that examiners form their decisions on 'specific features of the text to inform their judgements, not an overall "feel"' in assessing writing (Meara & Babí, 2001: 82). Note that Meara and Babí's (2001) study and the present experiment differ in at least two important aspects. Firstly, Meara and Babí examined assessment of written texts, while the present study has used spoken data. Secondly, their study sought to find that whether native and non-native assessors could differentiate between written texts produced by native and non-native speakers, and how much time was needed to arrive at a decision. Raters in their experiment did not assess any specific linguistic aspect of the texts, nor did they have to give scores on the quality of the writing. In contrast, in the present experiment, raters assessed vocabulary on a six-point rating scale.

The examination of individual raters' segmental scores showed that, although raters on average came to a decision on vocabulary scores in a cumulative fashion, their approaches to making a decision were different. There were three patterns in assessing segments within a text: giving identical segmental scores, changing segmental scores drastically, and varying segmental scores within a reasonable range. Most raters belonged to the third group. Nevertheless, when some raters were very consistent in their assessments of segments, with standard deviations less than .50 (Table 3); others displayed a wilder approach, with standard deviations larger than 1.0. For example, the scores from rater 5 were very similar within each text. The standard deviations of his ratings on the 5 texts were all below .53. His raw scores show that all the 10 segments in T4 were rated at 3. He did not change his decision at all after the first segment. Quite the opposite, other raters displayed a large range of variance in giving scores, such as rater 15, whose standard deviations for the 5 texts were either approaching 1.00 or above 1.00. When he rated T3, his ratings ranged from the lowest score of 1 to the highest score of 6 on the scale. This disparity in the range of segmental scores within texts caustions that although raters generally follow a cumulative approach, they may to some extent vary in decision-making process.

The second important finding in relation to the rating process is that raters formed their first impressions fast, just after a few utterances. As shown in Figure 2, after the first two segments, there was no considerable change of scores in all the five texts. This implies that after the first two segments, raters' first impressions on candidates' performance was formed (this equal to approximately 30 words). However, raters' first impressions were not the most reliable rating. The correlations between segmental scores and overall scores became stronger as more segments were heard (Figure 2). After 5 or 6 segments, raters' judgements on vocabulary became stable. Their judgements at this stage were more indicative of their final decisions on candidates' vocabulary. One segment contains one idea unit (see methodology), which normally lasts about 20 seconds. In other words, if 5 or 6 segments were counted by the minute, this was about one and a half minutes into the texts. If counted by tokens, it was after about 120 tokens.

Two implications arise from the above findings. On the one hand, 'minimal' texts (such as texts shorter than one minute) might not be sufficient to generate reliable vocabulary scores. On the other hand, if raters come to their final decisions just after a few utterances, their judgement may not be reliable. Note that T1 and T2 contained only 6 and 7 segments respectively, and the generalization that raters came to stable judgement quickly was mainly based on the observations of Texts 3, 4 and 5. Obviously, more sample texts are needed for a firm conclusion.

6. Conclusion

Raters' agreement in assessing vocabulary has been explored together with their decision-making process in this paper. The segmental rating approach followed in this experiment has been found beneficial not only in exploring the rating process, but in investigating agreement between multiple raters. For example, some raters were found to attend to some common lexical features when assessing spoken vocabulary.

For raters' performance in this experiment, the present study shows that they agreed moderately on the judgements of vocabulary. It seems possible that raters may assess vocabulary reliably in oral examinations. However, there was unmistakable variation between raters. Rasch analysis revealed rater disagreement on lexcial property of some segments. More study is needed to get a clearer idea of where discrepencies lie between raters.

Segmental ratings from raters suggested that raters' decision-making process is cumulative. They might form their judgement fast, but their first impressions might not be reliable evaluations on candidates' vocabulary output. Consequently, spoken texts shorter than one minute may not be able to generate valid judgements from raters.

Vocabulary scores provided in this experiment were found to relate more to candidates' overall language proficiency than to their lexical performance (as measured by types, tokens, D, PLex, etc.). That means raters' judgements on vocabulary might be made holistically. If so, there comes a question of how to interpret vocabulary scores on analytic rating scales. More qualitative data may help to get a more mature conclusion in this aspect.

References

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238. http://dx.doi.org/10.1177/026553229501200206

Bejar, I. I. (1985). A preliminary study of raters for the Test of Spoken English. *ETS Research Report Series, 1985*(1), i-28. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00090.x

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110. http://dx.doi.org/10.1191/0265532203lt245oa

Brown, A. (2006). An examination of the rating process in the revised IELTS Speaking Test. *IELTS Research Reports Volume, 3,* 41-70. Retrieved from http://www.ielts.org/researchers/research.aspx

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test taker performance on English for academic purposes speaking tasks.* Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2005.tb01982.x

Brown, A., & Taylor, L. (2006). A worldwide survey of examiners' views and experience of the revised IELTS Speaking Test. *Cambridge ESOL Research Notes,* 26, 14-18. Retrieved from http://www.ielts.org/researchers/research.aspx

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33. http://dx.doi.org/10.1177/026553229501200102

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision Making While Rating ESL/EFL Writing Tasks: A Descriptive Framework. *The Modern Language Journal, 86*(1), 67-96. http://dx.doi.org/10.1111/1540-4781.00137

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*(2), 125-144. http://dx.doi.org/10.1177/026553229401100203

Eckes, T. (2005). Examining Rater Effects in TestDaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis. *Language Assessment Quarterly, 2*(3), 197-221. http://dx.doi.org/10.1207/s15434311laq0203_2

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing, 10(*3), 235-254. http://dx.doi.org/10.1177/026553229301000303

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37. http://dx.doi.org/10.1177/0265532207071511

Elder, C., Golombek, P., Weigle, S. C., Boldt, H., & Valsecchi, M. (2003). Effects of task and rater background on the evaluation of ESL student writing: A pilot study. *TESOL Quarterly, 37*(2), 345-354. http://dx.doi.org/10.2307/3588510

Erdosy, M. (2003). *Exploring variability in judging writing ability in a second language: a study of four experienced raters of ESL compositions.* TOEFL Monograph Series. http://dx.doi.org/10.1002/j.2333-8504.2003.tb01909.x

Fulcher, G. (1996). Invalidating validity claims for the ACTFL oral rating scale. *System, 24*(2), 163-172. http://dx.doi.org/10.1016/0346-251X(96)00001-2

Galaczi, E., D. (2005). Upper Main Suite speaking assessment: towards an understanding of assessment criteria and oral examiner behaviour. *Research Notes, 20*, 16-19. Retrieved from http://www.ielts.org/researchers/research.aspx

Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics, 25*(1), 205-227. http://dx.doi.org/10.1017/s0267190505000115

Iwashita, N. (2010). Features of Oral Proficiency in Task Performance by EFL and JFL Learners. In M. T. Prior, Y. Watanabe & S.-K. Lee (Eds.), *Selected Proceedings of the 2008 Second Language Research Forum: Exploring SLA Perspectives, Positions, and Practices* (pp. 32-47). Somerville, MA: Cascadilla Proceedings Project.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: how distinct? *Applied Linguistics, 29*(1), 24-49. http://dx.doi.org/10.1093/applin/amm017

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*(2), 81-96. http://dx.doi.org/10.1016/j.asw.2011.02.003

Linacre, J. M. (2006). *A user's guide to Facets: Rasch measurement computer program*. Chicago, IL: MESA Press.

Lorenzo-Dus, N., & Meara, P. (2005). Examiner support strategies and test-taker vocabulary. *IRAL, 43(*3), 239-258. http://dx.doi.org/10.1515/iral.2005.43.3.239

Li, H., & Lorenzo-Dus, N. (2014). Investigating how vocabulary is assessed in a narrative task through raters' verbal protocols*. System, 46,* 1-13. http://dx.doi.org/10.1016/j.system.2014.06.006

Lu, X. (2012). The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal, 96*(2), 190-208. http://dx.doi.org/10.1111/j.1540-4781.2011.01232_1.x

Lumley, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language profficiency. *English for Specific Purposes, 17*(4), 347-367. http://dx.doi.org/10.1016/S0889-4906(97)00016-1

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*(1), 54-71. http://dx.doi.org/10.1177/026553229501200104

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing, 22*(4), 415-437. http://dx.doi.org/10.1191/0265532205lt303oa

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85-104. http://dx.doi.org/10.1191/0265532202lt221oa

May, L. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing*, 29-51. Retrieved from http://pandora.nla.gov.au/tep/80462

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*(3), 397-421. http://dx.doi.org/10.1177/0265532209104668

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

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555-576. http://dx.doi.org/10.1177/0265532211430367

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35-53). Cambridge: Cambridge University Press. Retrieved from http://www.lognostics.co.uk/vlibrary/meara1996a.pdf

Meara, P., & Babí, A. (2001). Just a few words: how assessors evaluate minimal texts. *IRAL,* 39, 75-83. http://dx.doi.org/10.1515/iral.39.1.75

Nation, P. (1990). *Teaching and Learning Vocabulary*. New York: Newbury House.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192. http://dx.doi.org/10.1191/0265532202lt226oa

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3). http://dx.doi.org/10.1191/0265532202lt205oa

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511732942

Read, J. (2005). Applying lexical statistics to the IELTS speaking test. *Research Notes, 20*, 12-16. Retrieved from http://www.ielts.org/researchers/research.aspx

Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing, 29*(2), 223-241. http://dx.doi.org/10.1177/0265532211421162

Taylor, L., & Jones, N. (2001). Revising the IELTS speaking test. *Research Notes, 4,* 9-11. Retrieved from www.cambridgeenglish.org/images/22644-research-notes-5.pdf

Weigle, S., & Nelson, G. (2004). Novice tutors and their ESL tutees: Three case studies of tutor roles and perceptions of tutorial success. *Journal of Second Language Writing, 13*(3), 203-225.

http://dx.doi.org/10.1016/j.jslw.2004.04.011

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223. http://dx.doi.org/10.1177/026553229401100206

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263. http://dx.doi.org/10.1191/026553298670883954

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231-252. http://dx.doi.org/10.1177/0265532212456968

Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition, 14,* 403-424. http://dx.doi.org/10.1017/S0272263100011207

Note 1. Thanks Prof. Paul Meara at Swansea University to help write the programe.

Copyrights