Title	Understanding discrepancies in rater judgement on national-level oral examination tasks
Author(s)	Hui Teng Ang-Aw and Christine Goh Chuen Meng
Source	*RELC Journal, 42*(1), 31-51
Published by	SAGE Publications

Understanding discrepancies in rater judgement on national-level oral examination tasks

Abstract

The oral examination is an important component of the high-stakes 'O' level examination in Singapore taken by 16-17 years olds whose first language may or may not be English. In spite of this, there has been sparse research into the examination. This paper reports findings of an exploratory study which attempted to determine whether there were any discrepancies in rater judgements and thereafter, explore the nature and scope of the discrepancies identified. Five audio recordings were obtained from a simulated oral examination of five candidates conducted by a trained 'O' level oral examiner. Seven other trained 'O' level oral examiners were asked to rate four of the recordings individually and provide concurrent verbal reports. Questionnaires were also given to the raters for data triangulation after the verbalisation. The data were analysed through Verbal Protocol Analysis and descriptive statistics. Rater discrepancies detected in the scores were qualitatively determined to be due to four differences: emphases on factors assessed, constructs of oral proficiency, rater interpretations and approaches in assessment. These findings provide valuable insights into raters' perceptions of the construct of speaking and offer implications for rater training and the development of rating scales.

Keywords: Assessment; Oral proficiency; Rater judgement; Rater perception; Speaking; Testing

I Introduction

Language assessment is a complex process whereby raters are often, especially in the assessment of writing and speaking, required to carry out subjective assessment of a person's language ability. In assessing speaking, it is common to see the use of conversation language proficiency interviews in many countries although rater variability has been widely acknowledged (Lumley & McNamara, 1995).

1 Rater Variability

Over the past decades, several studies support the findings that raters differ in the severity of their judgements of candidate's speaking proficiency and can produce a wide range of scores (A. Brown, 1995, 2000; Lumley & McNamara, 1995; Orr, 2002; Wigglesworth, 1993). Raters were also found awarding the same scores to different performance or different scores to the same performance (A. Brown, 1995; Douglas, 1994; Meiron & Schick, 2000; Orr, 2002). A. Brown (2000) also showed that raters behave either in a performance-oriented manner or in an inference-oriented manner. She also suggested that newly trained raters tend to focus more on the descriptors while the more experienced raters focus more on making inferences about candidates' speaking proficiency.

Raters with a teaching background tend to emphasize more on linguistic factors such as grammar though they might become desensitized to grammatical errors (Hadden, 1991). Chalhoub-Deville (1995), on the contrary, suggested that teachers might emphasize more on creativity than linguistic factors. Lumley (1998), McNamara (1996) and A. Brown (1995) also revealed that teachers may be more tolerant and willing to award marks for effort and at the same time, more reluctant to award extreme marks, producing a narrow range of marks.

Research has also highlighted raters' inconsistent use of rating scales. Wigglesworth's studies (1993, 1994) highlighted raters showing individual pattern of response to the criteria because of their focus on different criteria. Raters also based their decisions on a wide range of

non-criterion information which they deemed important or on different descriptors in the rating scales (A. Brown, 2000; Douglas, 1994; Pollitt & Murray, 1996; Orr, 2002). Therefore, raters seemed to lack understanding of the construct on which the rating scales are based (A. Brown 2000; Orr, 2002; Wigglesworth, 1994).

Furthermore, it was strongly argued that rater differences would still exist after training (A. Brown, 1995; Douglas, 1994; Lumley, 1998). It appears that raters have their own construct of a good performance and at times, this construct may even vary for candidates of varying abilities (Lumley, 1998; Pollitt & Murray, 1996). Raters also have their own distinctive style of examining, which could result in them offering different degrees of support to candidates (A. Brown, 2003; A. Brown & Hill 1998; Lazaraton, 1996b).

In assessing speaking, raters may also differ in their approaches. Pollitt and Murray (1996) highlighted two contrasting approaches:

(i) A *synthetic* process where raters first formed a holistic image of the candidate's performance (from a few first impressions) based on their preconceived understanding of language learners and then comparing it with subsequent observed performance.

(ii) A more objective, less natural process where raters "'scored' the candidates intuitively for each observed utterance, and somehow added these up" (p. 87), limiting their scoring to observed behaviour.

A. Brown (2000) further revealed a third approach which is a combination of both approaches:

(iii) Raters not only sought evidence to support or refute their initial judgements of candidates by candidates' fluency, vocabulary, structures and discourse before retaining or revising their initial judgements but also rated sections of the test independently before weighing them up.

In conclusion, research has revealed the extensiveness of rater variability and the necessity of using qualitative research to gain insights into the assessment process that statistical analysis cannot provide (A. Brown, 1995, 2000, 2004; Chalhoub-Deville, 1995,

1995b; Douglas, 1994; Lazaraton, 1996b, 2000; Nakatsuhara, 2007; Orr, 2002; Pollitt & Murray, 1996).

II Method

1 The testing context

The O' level English examination is a high-stakes national examination for secondary students in Singapore. It comprises three papers testing writing, reading and oral communication. In order to gain admission to any tertiary institution in Singapore, students must achieve a pass grade in this placement test. This study focuses on examining the reliability of the conversation component of the oral examination as it offers the greatest room for rater subjectivity and has the greatest weighting in the paper (refer to Appendix A for details about the oral examination).

Despite its importance, there are only three studies (Eng, 1997, 1999; Koh, 2003) on high-stakes oral examinations in Singapore. Nevertheless, all three studies illustrated concerns regarding the objectivity of rater judgements and supported incorporating a qualitative approach into studies on rater variability. Thus, this study seeks to address the gap in research on the high-stakes 'O' level oral examination in Singapore. It drew on A. Brown's (2000) study on the rating process of the IELTS oral interview, Koh's (2003) study on raters' decision-making process on the picture conversation task used for primary six candidates in Singapore and Hadden's (1991) study on teachers' perceptions of second language communication.

2 Aim

This study aims to explore rater variability in the 'O' level oral examination conversation task through both quantitative and qualitative approaches so that "different method would compensate for each other's inadequacies" (Milanovic, et al., 1996, p. 95). The research

question addressed is “What are the considerations involved in a rater’s decision-making process in the ‘O’ level conversation task?”

3 Participants

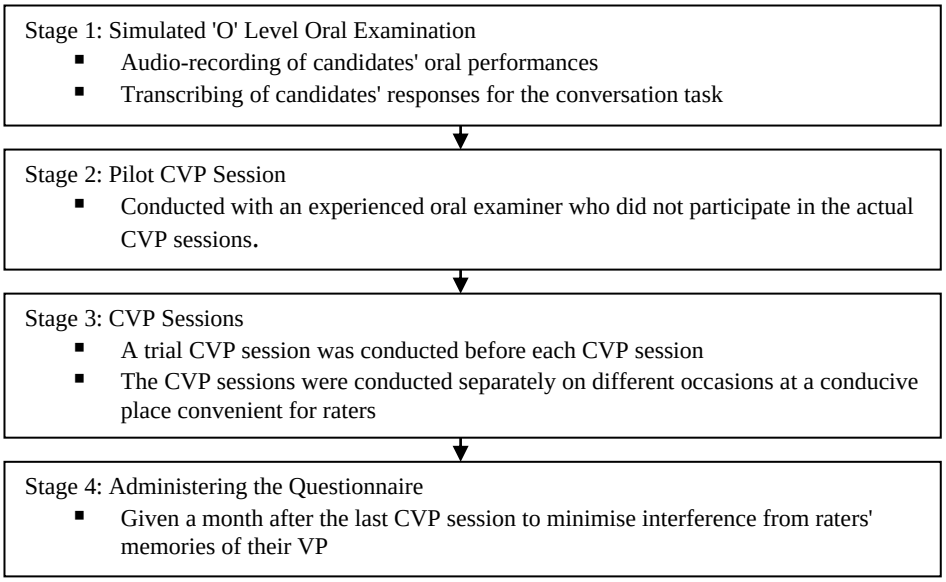
The seven teacher participants were from three neighbourhood secondary schools. They were all trained ‘O’ level oral examiners, having between five to nine years of experience, except for R4 and R5 who had slightly less than five years of experience. The four student participants – one male and three female candidates - were all secondary five students from a neighbourhood secondary school. They were students of low, middle and high ability (as judged by their teacher).

4 Data Collection

This study uses data from Concurrent Verbal Protocols (CVP), questionnaires and scores. The use of CVP help to avoid contamination of data by raters who try to tidy up their thoughts, rationalize their decisions or give extra information retrospectively.

The flow chart (Figure 1) below provides an illustration of the different stages of the data collection process. A questionnaire (Appendix B) was designed specifically for this research to supplement the CVP data. The first six categories (Phonology, Accuracy, Fluency, Strategic Capacity, Topical Knowledge and Personal Characteristics) and items were adapted from Fulcher’s (2003) framework for describing the speaking construct while the seventh category (Examination Criteria) was based on the ‘O’ level oral examination marking scheme. They sought to gather raters’ beliefs about oral proficiency and identify the level of importance they place on the exam criteria. Using a 5-point Likert scale, ranging from 1 “Strongly disagree” to 5 “Strongly agree”, raters were asked to indicate the level of importance they placed on different aspects of speaking in assessing the conversation task. After each category, the raters were also invited to explain the rationale for their response.

Figure 1
Stages in data collection



5 Data Analysis

Data analysis involved initially analyzing the scores using simple descriptive statistics to determine the extent of quantitative similarities and/or differences among the ratings. The performances were then qualitatively studied using Verbal Protocol Analysis (VPA) so as to understand which aspects of candidates' performances were driving the scores, the reasons behind raters' scores and the cognitive and pyscholinguistic processes underlying raters' performances (Kormos, 1998).

All the 28 sets of VP were transcribed fully by the researcher. They were transcribed, segmented and codified on the basis of the categories naturally appearing in the data and in the 'O' level oral examination marking scheme. In order to be able to make more specific inferences, the coding was done as comprehensively as possible (Green, 1998). Inter-coder reliability was established by asking an independent coder to help to code and segment the VP. Due to the vast amount of data, the independent coder was only asked to code R1's VP. The level of agreement was about 85% and is expected with a more comprehensive coding scheme (Green, 1998). Nevertheless, actions were taken to resolve the differences.

III Results

Despite previous training received by the raters, differences existed in four areas: emphases on factors considered, perceived constructs of oral proficiency, interpretations of performances and scores, and approaches in assessment. Raters are indicated as R1, R2 and so on while candidates are indicated as C1, C2 and so on. Table 1 reveals the quantitative differences in rater judgements.

Table 1
Scores awarded to four candidates by raters

Raters	Candidates			
	C1	C2	C3	C4
1	10	11	9	9
2	10	11	8	9
3	11	13	11	11
4	9	10	8	9
5	10	12	9	9
6	11	13	10	10
7	11	14	8	9
Mean	10.29	12.00	9.00	9.43
Band (marking scheme)	7 Band 2	3 Band 1, 4 Band 2	4 Band 2, 3 Band 3	7 Band 2
Range of scores	9-11	10-14	8-11	9-11
Std. Dev.	0.76	1.41	1.15	0.79

Out of a highest possible score of 16 marks, C2 was the best-performing candidate (mean=12) and C3 the worst-performing candidate (mean=9). Most significantly, the biggest range of scores (across two bands) was also awarded for C2 and C3 suggesting that discrepancy is most evident for the high-ability and low-ability candidate. This paper therefore focuses on C2 and C3.

1 Factors Considered

a Criterion Factors

All raters took into account criterion factors during assessment (Table 2). Criterion factors refer to the band descriptors in the holistic marking scheme which focuses on three factors:

- personal response: the quality (how intelligent) and the quantity (the extent of the development) of candidates' responses (No. 1).
- clarity of expression: candidates' ability to express themselves with clarity and in a succinct manner using appropriate vocabulary and structures (No. 2-5).
- engagement in conversation: candidates' ability to respond to examiners' prompts with initiative and effort (No. 6-8).

Table 2
Raters' criterion factors in assessing candidates' performances

No.	**Criterion Factors**	**Raters**							**%**
		R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	
1	Elaboration of Response	■	■	■	■	■	■	■	100
2	Clarity of Expression	■	■	■	■	■	■	■	100
3	Organisation of Idea	■	■	■	■	■	■	■	100
4	Grammatical Accuracy	■	■	■	■	■	■	■	100
5	Use of Vocabulary		■	■	■	■	■	■	85
6	Initiating Discussion of Issues	■		■		■		■	57
7	Response to Prompt	■	■	■	■	■	■	■	100
8	Effort of Candidate	■		■		■	■	■	71

Note: Shaded areas denote raters' focuses

Significantly, R2 and R4 did not focus on both *initiating discussion of issues* or *effort of candidate* which are both aspects that contribute to engagement in conversation. This suggests that they were focusing more on two out of three descriptors and this could impact on candidates' scores, because other raters were awarding scores to candidates based on their effort, for example,

> R3-C3
> Uhm: but he does try hard in trying to respond to whatever the: examiner was prompting. (.) Uh, and to his abili- as far as he is concerned, once he understands the- (.5) question, he will try his best- to: elaborate. (.) Uh, definitely have very good effort. (1.0)

The impact on scores could be significant especially for candidates like C3 who, despite being linguistically weak, put in great effort to response. In high-stakes placement tests, this could

mean a pass or fail and affect candidates' chances of enrolling into higher institutes of education where a pass in English is a prerequisite.

Furthermore, R5 was unwilling to award C2 a band one score because out of the three criteria, C2 did not achieve clarity of expression. This suggests that all three criteria were important to R5. However, R3, R6 and R7 were willing to award a band one score to C2 based on her *intelligent personal response* even though she did not fit all three descriptors.

A. Brown (2000) has attributed such behaviour to either some descriptors being more salient to some raters or that the descriptors are too vague. In addition, we would suggest that raters' display of such behaviour indicates the inadequacy of training since during the compulsory training session it was specifically shown that candidates should be awarded marks for effort.

b Non-criterion Factors

Rater reliability is dependent on raters applying the same standards without being affected by human error, subjectivity or bias (H. Brown, 2004). However, A. Brown (2000), Douglas (1994), Orr (2002) and Pollitt and Murray (1996) have shown that raters referred to a wide range of non-criterion information which they deemed important. Table 3 highlights the non-criterion factors that raters paid attention to. It appears that raters focused on a wide range of non-criterion factors and *novelty of idea* was the only common non-criterion factor among all the raters. While five factors (No. 2-6) were salient to the majority (more than 50%) of the raters, seven other factors (No. 7-13) were salient only to a minority (less than 50% of the raters). This implies that raters were judging candidates' performances using different yardsticks and were not adhering strictly to the marking scheme. Such behaviour threatens the construct validity of the assessment.

Table 3
Raters' non-criterion factors in assessing candidates' performances

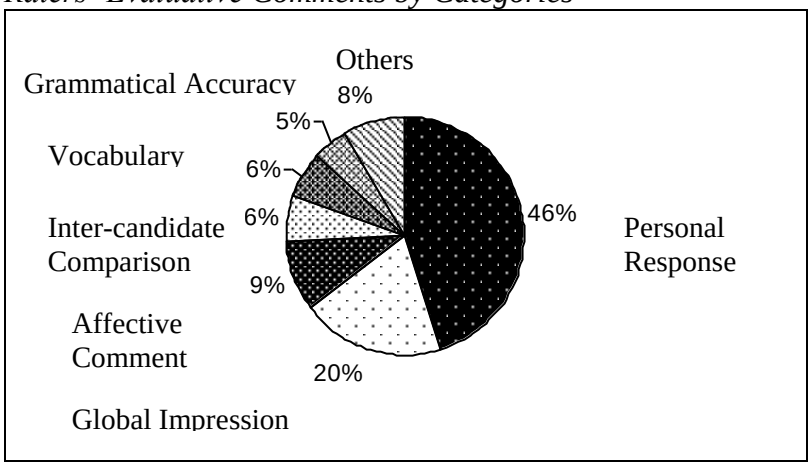
No.	Non-Criterion Factors	Raters							%
		R1	R2	R3	R4	R5	R6	R7	
1	Novelty of Idea	■	■	■	■	■	■	■	100
2	Range of Vocabulary	■		■	■	■	■	■	85
3	Inter-candidate Comparison	■	■		■	■	■	■	85
4	Delivery	■	■		■	■		■	71
5	Interest level of Candidate		■	■		■	■	■	71
6	Personality of Candidate		■	■	■		■	■	71
7	Confidence level of Candidate		■	■		■			42
8	Tone of Candidate	■			■	■			42
9	Test-taking Strategy			■			■	■	42
10	Pronunciation	■						■	28
11	Voice of Candidate					■			14
12	Interactiveness							■	14
13	Moderation	■							14

Note: Shaded areas denote raters' focuses

2 Perceived Constructs of Oral Proficiency

Raters did not have the same understanding of the construct of oral proficiency for the conversation task. Section 1(a) and 1(b) show that the raters' decision-making process was guided by different constructs of oral proficiency which may be related to their beliefs about oral proficiency. Furthermore, these beliefs did not always correspond with the descriptors in the 'O' level marking scheme. There were only six common factors among all the raters and some factors were of importance only to particular raters (see Table 2: no. 1-4,7; Table 3: no. 1). For example, only R7 commented on *interactiveness*. Figure 2 shows the number of raters' evaluative comments according to the various categories..

Figure 2
Raters' Evaluative Comments by Categories



Candidates' ability to give personal responses received the most attention from the raters (46%) followed by candidates' overall performance assessed impressionistically (20%). In comparison, data from the questionnaire showed that raters believed all the all features to be important. Therefore, on balance we can see that while the raters had an understanding of what constituted oral proficiency, their final decisions when assessing candidates could have been influenced by one or two criteria that they felt were most important. Such differences in the way the construct of oral proficiency appeared to have been operationalised during real-time assessment would have serious implications on the overall reliability of a wide-scale oral examination.

3 Interpretations

a Interpretations of Performances

Besides differences in construct, raters may have different interpretations of candidates' performances. In table 4, while other raters felt that C3's personal response was simple (No. 1), R3 felt that C3 gave some very good personal responses (No. 2). This could be because R3 were more able to comprehend what C3 was saying as she commented that C3 gave generally

clear responses (No. 9). A. Brown (2000), in fact, has shown that some raters were more able to follow learner's speech than others.

Table 4
Raters' comments on C3's performance

No	Comments	Raters						
		R1	R2	R3	R4	R5	R6	R7
1	Some relevant simple responses	■	■		■	■	■	■
2	Some very good responses			■				
3	Poor grammar	■		■	■	■	■	■
4	Uninteresting ideas				■		■	
5	Poor organisation of ideas	■		■				■
6	Poor vocabulary	■	■	■	■		■	■
7	Effective use of circumlocution			■				
8	Long-winded responses							■
9	Generally clear responses			■		■		
10	Difficulty comprehending at times				■		■	
11	Did not really answer the prompts	■	■	■	■		■	
12	Displayed a lack of confidence	■						
13	Good effort/initiative			■		■		■

Note: Shaded areas denote raters' comments

b Interpretations of Scores

Besides differences in interpretation of performances, raters may also have different interpretations of candidates' scores. Raters may award different scores to the same performance or the same score to different performances. This could be due to raters possessing different interpretations of abstract terms used in the holistic scales and the scores they represent or their ability to comprehend candidates' utterances (A. Brown's, 2000).

In table 4, raters may have similar interpretation of performance but yet different notions of which score the performance represents (Orr, 2002). R4 and R6 both felt that C3 should just pass. They commented that C3 had poor grammar (No. 3), uninteresting ideas (No. 4) and they had a problem comprehending him at times (No. 10) though he had some relevant simple responses (No. 1). However, R6 awarded him a band two score of ten marks while R4 only awarded him a band three score of eight marks.

On the other, in table 5, R3 and R6 might have awarded the same score to C2 but they were awarding them based on contrasting perceptions of C2's performance. Both agreed that C2 gave intelligent response (No. 2) and that she made some language errors. However, while R6 felt that C2 used interesting vocabulary (No. 7) and displayed a good use of connectors (No. 9), R3 felt that "many of C2's vocabulary need(ed) much thought" (No. 8) and that C3 had problems with grammatical structures (No. 8).

Table 5
Raters' comments on C2's performance

No	**Comments**	**Raters**						
		R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**
1	Gave good responses	■	■			■		
2	Gave intelligent responses			■			■	■
3	Gave some responses				■			
4	Good organisation		■			■		
5	Gave limited responses	■	■					
6	Long-winded responses	■						■
7	Good vocabulary						■	■
8	Weak vocabulary			■				
9	Satisfactory grammar						■	
10	Weak grammar			■				■
11	Hesitant delivery		■		■			
12	Displays a lack of confidence		■					
13	Displays confidence			■				
14	Good effort/initiative					■		■
15	Did not interact with the examiner							■
16	Engaging personality			■				

Note: Shaded areas denote raters' comments

Within raters, R1, R3, R5 and R6 also gave the same score to both C3 and C4 though with different reasons. Table 6 shows R1's, R5's and R6's comments on C3 and C4's performances.

Table 6
Rater 1's, 5's and 6's comments on C3's and C4's performances

No	Comments	Raters					
		R1		R5		R6	
		C3	C4	C3	C4	C3	C4
1	Some relevant simple responses	■	■	■		■	■
2	Some responses				■		
3	Poor grammar	■	■	■	■	■	
4	Uninteresting ideas					■	
5	Poor organisation of ideas	■					■
6	Poor vocabulary	■				■	■
7	Generally clear responses			■			
8	Difficulty comprehending at times					■	
9	Did not really answer the prompts	■				■	
10	Displayed a lack of confidence	■					
11	Good effort/initiative	■	■	■	■		
12	In appropriate tone		■		■		

Note: Shaded areas denote raters' comments

R1 focused on three common criteria (no 1, 3 and 11) for awarding the same scores to C3 and C4 but for C3, she also focused on four other criteria (No. 5, 6, 9 and 10) while for C4, she also focused on one other criterion (No. 12). Such differences in focus highlight that even in awarding the same score, raters could be basing that score on different criteria. The different focus could be due to different emphases on particular factors as in R1's comment on C4:

> R1-C4
> Uhm: (2.8) okay, I: will: (.2) give (.2) her a: (5) nine and nothing more
> R1C4-CR-2
> because (1.8) she's (.5) taking the conversa- she's taking this- (.2) oral exam a little too casually.

or the constrains of the marking scheme:

> R5-C4
> Okay, I'll give her a nine. (1.5) Because I can't justify her meeting all the criteria (.2) for the band five to eight. (.8) Though I personally think she's poorer (.5) than the previous student. (3.0) Okay, nine.

4 Assessment Approaches

a Synthetic, Objective and Mixed Approach

In looking at raters' concluding remarks where raters revealed how they arrived at candidates' scores, we could first clearly identify the *synthetic* approach (Pollitt & Murray, 1996) adopted by R1, R4, R5 and R6. For example, R5 first formed an initial impression of C3's performance and decided on a score. Following that, she proceeded to evaluate and weigh different aspects of C3 performance before finally confirming that C3 indeed deserved nine marks.

The next approach identified is the more objective approach (Pollitt & Murray, 1996), herein termed the *objective* approach, adopted by R2. For example, R2 scored C1's response for the first prompt. Following that, she evaluated C1's response for her second prompt before arriving at a final score.

In the *mixed* approach (A. Brown, 2000), R3 and R7 assessed candidates' responses for both prompts and also individual aspects of their performances. R3 evaluated different aspects of C1's performance before tentatively deciding that C1 deserved a band two grade. At the same time, she was evaluating C1's responses for both prompts. In fact, she subsequently elaborated on the unsatisfactory aspect of C1's response to the second prompt before reaffirming her decision to band C1 in band two. The different approaches can be summarized in table 7.

Table 7
Raters' approaches to assessment

Candidates	Raters						
	R1	*R2*	*R3*	*R4*	*R5*	*R6*	*R7*
C1	synthetic	objective	mixed	synthetic	synthetic	synthetic	synthetic
C2	synthetic	objective	mixed	synthetic	synthetic	synthetic	mixed
C3	synthetic	objective	mixed	synthetic	synthetic	synthetic	synthetic
C4	synthetic	objective	objective	synthetic	synthetic	synthetic	synthetic

Note: Shaded areas denote a different approach from the usual approach adopted

It is noteworthy that raters R3 and R7 did not consistently apply the *mixed* approach. It is possible that they had resorted to a new approach as their usual approach was unable to help them decide on candidates' scores. Also, since both were deciding on whether to award the

candidates with a band one score by looking at candidates' responses to both prompt, it can be further suggested that the two raters believed that a satisfactory answer to both prompts was a necessary condition for candidates to be awarded a band one score. As raters were not instructed on how to weigh the responses for the two prompts or the band descriptors during the training sessions, it is unsurprising that raters adopted the different approaches.

b Different Levels of Severity/Leniency

Raters also approached the assessment with different levels of severity/leniency. Three types of raters in terms of level of severity/leniency were identified: a consistently lenient rater, a consistently strict rater and a rater who is lenient towards the stronger candidate but strict towards the weaker candidate.

R3 was significantly more lenient than all the other raters. She consistently referred to candidates' effort and one of the reasons why she was willing to award C3 a high band two score was because he tried very hard to respond. R3 could also be more generous because she was more inference-oriented (A. Brown, 2000). For example, she did not penalize C1 because she inferred that C1 was not able to hold the discussion because of her language ability and not because of the lack of ideas.

R4 was harsher than all the other raters, even for C2. Although R4 commented that C2 was generally proficient, she commented that C2 (and also all the candidates) had "appalling" grammar and was disorganized. She was clearly also less tolerant with the use of Singlish (Singapore Colloquial English), which was most often spoken by C3 who frequently left out the –en/-ed/s inflections in his sentences. Since R4 is more inexperienced as compared to the other raters, she might have unrealistically high expectations from candidates and these, together with her personal beliefs, could have led to her negative judgements of candidates' performances.

R7 was lenient on C2 but harsh on C3 although she found both of them having poor grammar and being long-winded. This could be because she found greater coherence in C2's response and most importantly, she was willing to award marks for a response that demonstrated knowledge about general affairs (seen in C2). Furthermore, unlike other raters who assessed C3's effort favourably, she found C3 "draggy".

Therefore, it can be inferred that raters' levels of severity/leniency were influenced, though by different extents, by their inbuilt perceptions of what was acceptable to them (A. Brown, 1995, 2000; McNamara, 1996) when rater bias was involved, ratings would tend to diverge (Douglas, 1994).

c Inter-candidate Comparison

Though raters were not expected to make comparisons between candidates, all raters, except R3, did compare candidates' performances. However, the number of comparisons greatly differs. Out of a total of 38 comments, R1 and R2 only made one comparison each (2%) while R6 made 18 comparisons (47%). In addition, in making inter-candidate comparison, raters focused on both criterion factors, such as *elaboration of response*, *organisation of idea* and *sentence structure,* and non-criterion factors such as *range of vocabular*y.

The impact of inter-candidate comparison on the scores awarded can be most evidently seen in R6's case when she resorted to adjusting marks after comparing C2's and C1's performances:

> R6-C2
> If it's 13, it's- (.) then the first person is too high. (3.0) 13, if it's 14, it's a bit high also. (2.5) I just give thir-teen and then downgrade the first one to eleven.

Such comparisons risk changing the test from a criterion-referenced test to a norm-referenced test, where candidates' scores would be affected by the relative oral proficiency of those who are taking the exam in the same session. Comparisons should have been avoided in a criterion-

referenced test, where raters are expected to score criterion-referenced interpretation of candidates' performances using the band descriptors without being influenced by other candidates' performances.

Koh (2003) and Orr (2002) suggested that such a focus on inter-candidate comparison could have been caused by the ambiguity of the descriptors as a whole. When raters were not able to clearly match candidates' performances to the descriptors, they would tend to fall back on their own internal criteria and such a criterion could be the relative performance of other candidates. In addition, the ambiguity could be causing raters difficulty in distinguishing the constructs of different levels of performances and therefore, as in R6's example, raters could end up "feeling" that a particular score is too high or too low for a candidate.

IV Conclusion and Implication

1 Conclusion

This study shows differences in raters' perceptions of the constructs of oral proficiency, their interpretations of the marking scheme and candidates' performances, and assessment approaches. Although the raters had undergone similar training and were experienced 'O' level oral examiners, they adhered to the marking scheme to varying extents, had different preoccupations with different aspects of candidates' performances and assessed in a dissimilar manner. However, they generally placed greater focus on content than language accuracy.

Nonetheless, this study acknowledges three limitations. Firstly, the non face-to-face assessment of candidates limits the discussion to only assessment of verbal data. Secondly, we cannot assume that a candidate has performed unsatisfactorily on a particular aspect of the candidate's performance when a rater does not comment on that aspect (A. Brown (2000). The rater could simply be focusing on noticing mistakes, taking them as a sign of a lack of

proficiency or focusing on other salient factors. Thirdly, Verbal Protocol Analysis has its limitations in understanding raters' decision-making process because some raters might verbalise their thoughts but were unable to "report on the perceptual and retrieval processes that determine which thoughts or patterns will reach their attention" (Ericsson & Simon, 1993, p. l).

2 Implications

This study has highlighted inconsistency in rater behaviour that may affect reliability and the construct validity of the high-stakes 'O' level oral examination. It also questions the usefulness of the marking scheme and the effectiveness of the training session in preparing raters for assessing this high-stakes national examination reliably. Firstly, raters need to assess characteristics of candidates' performances which correspond to the descriptors in the marking scheme (Pollitt & Murray, 1996). Secondly, raters need to be clear about the aspects of candidates' performances that they should and should not focus on. Thirdly, also supported by Eng (1997, 1999), Lazaraton (1996b) and Wigglesworth (1993), raters' feedback should be elicited and at the same time qualitative feedback be given to them during training. Besides this, raters could also observe 'expert raters' decision-making process (Koh, 2003). Fourthly, statistical adjustments for rater characteristics (Bonk & Ockey, 2003) or double ratings (A. Brown, 2000) could be carried out for raters who are identified as more unreliable during the training sessions.

3 Recommendations For Research

This is a pioneer study on the 'O' level conversation task and replications are needed before definite conclusions can be drawn. Firstly, a bigger and more representative dataset is needed in order to validate the findings from this research and to draw stronger conclusions. This could include students from the express stream and less experienced raters. Secondly,

structured or semi-structured interviews with raters after the concurrent verbal protocol sessions could help to elicit more or clearer explanations of the basis of raters' decisions. Thirdly, with a bigger dataset, a more systematic analysis using quantitative tools such as FACETS could be carried out to study the relationship between the various facets. Fourthly, studies could also be carried out to study interviewer, task and scale effects on scores and exploring the use of different prompt types, raters' styles and perceptions, pairing of raters or the use of an analytical scale. Lastly, similar research focusing on the read-aloud task or the picture discussion task could be carried out. This would provide a more holistic understanding of the oral examination and its reliability.

References

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.

A. Brown. A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(3), 1-15.

A. Brown. A. (2000). An investigation of the rating process in the IELTS oral interview. *IELTS Research Reports, 3*, 49–84.

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

A. Brown, A. (2004). An examination of the rating process in the revised IELTS speaking test. *IELTS Research Reports 6*, 41-70.

A. Brown, A., & Hill, K. (1998). Interviewer style and candidate performance in the IELTS oral interview. *IELTS Research Reports, 1*, 1–19.

A. Brown, H. D. (2001). *Language Assessment: Principles and Classroom Practices.* Longman: Pearson

Chalhoub-Deville, M. (1993). Performance assessment and the components of the oral construct across different tasks and rater groups. In Milanovic, M. & Saville, N. (Eds), *Studies in Language Testing 3* (pp. 55-73). Cambridge: Cambridge University Press.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33.

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgement and English language speaking proficiency. *World Englishes*, 24(3), 383-391.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*(2), 125-144.

Eng, P. S. (1997). An evaluation of the picture description task in the GCE 'O' level English language examination. Unpublished MA dissertation. National University of Singapore.

Eng, P. S. (1999). Picture description: Investigating its reliability and validity. *Language & Comunication Review, 8*(2), 33-42.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis*: MIT Press Cambridge, Mass.

Fulcher, G. (2003). *Testing Second Language Speaking*: Pearson Education.

Green, A. (1998). *Verbal protocol analysis in language testing research*: Cambridge University Pr.

Hadden, B. L. (1991). Teacher and Nonteacher Perceptions of Second-Language Communication. *Language Learning, 41*(1), 1-20.

Koh, C. H. C. (2003). *An exploratory study of three raters' decision-making process of the picture conversation task used for primary six candidates in Singapore.* Unpublished honours thesis, National Institute of Education, Nanyang Technological University, Singapore.

Kormos, J. (1998). The Use of Verbal Reports in L2 Research: Verbal Reports in L2 Speech Production Research. *TESOL Quarterly, 32*(2), 353-358.

Lazaraton, A. (1996a). Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing, 13*(2), 151-172.

Lazaraton, A. (1996b). A qualitative approach to monitoring examiner conduct in the Cambridge assessment of spoken English (CASE). In Milanovic, M. & Saville, N. (Eds), *Studies in Language Testing 3* (pp. 18-33). Cambridge: Cambridge University Press.

Lazaraton, A. (2002). *A Qualitative Approach to the Validation of Oral Language Tests*: Cambridge University Press.

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.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*(1), 54-71.

Luoma, S. (2004). *Assessing speaking.* Cambridge: Cambridge University Press.

McNamara, & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational I settings. *Language Testing, 14*(2), 140.

Meiron, B., & Schick, L. (2000). Ratings, raters and test performance: An exploratory study. In Milanovic, M. & Kunnan, A. J. (Eds), *Studies in language testing 9* (pp. 153–176). Cambridge, Cambridge University Press.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In Milanovic, M. & Saville, N. (Eds), *Studies in Language Testing 3* (pp. 92-111). Cambridge: Cambridge University Press.

Nakatsuhara, F. (2007). Inter-interviewer variation in oral interview tests. *ELT Journal*, 62(3), 266-275.

Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores. *System, 30*(2), 143-154.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In Milanovic, M. & Kunnan, A. J. (Eds), *Studies in language testing 3* (pp. 74–91). Cambridge, Cambridge University Press.

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

Wigglesworth, G. (1994). Patterns of Rater Behaviour in the Assessment of an Oral Interaction Test. *Australian Review of Applied Linguistics, 17*(2), 77-103.

APPENDIX A

Details of GCE 'O' level Oral Examination

The oral examination is made up of three sections:

1. Reading aloud (30%)
2. Picture discussion (30%)
3. Conversation (40%).

In the conversation task, raters have to use the two main prompts given and an additional four sub-prompts are provided for extra assistance. Holistic scoring is carried out with the scale being divided into four bands with a range of four marks for each band. The descriptors in each band follow seek to assess candidates on three aspects:

(ii) Personal Response: give a personal response to the theme of the picture and the passage;

(iii) Clarity of Expression: express himself/herself clearly and succinctly in a conversation using appropriate vocabulary and structures and

(iv) Engagement in Conversation: discuss issues that arise with the Examiner stemming from the picture and the passage

The marking scheme, defined by Bachman's and Palmer's model of Communicative Language Ability (CLA) (as cited in Luoma, 2004, pp. 97-101), revealed the conversation task to be a test of communicative competence based on candidates' linguistic ability, use of communicative strategies and the ability to deal with the situational and interpersonal aspects of communication. In terms of Bygate's model of speech as a process (as cited in Luoma, 2004, pp. 103-107), the task requires candidates to have both language knowledge and skill to use that knowledge. Emphasis is placed explicitly on vocabulary and candidates' ability to express themselves succinctly. Interactional skills and agenda management are also tested as candidates interact with raters, trying to engage them in conversation.

APPENDIX B

Questionnaire

Name: _____

Use the following scale (numbers 1 through 5) to describe how you feel about each of the statements below. For each statement, circle/highlight the number that gives the best description of how you feel.

Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
1	2	3	4	5

The following factors are <u>important</u> to me when I assess a candidate's oral proficiency at the <u>conversation</u> component:

<u>Phonology</u>

1) Stress 1 2 3 4 5

2) Rhythm 1 2 3 4 5

3) Intonation 1 2 3 4 5

4) Pronunciation 1 2 3 4 5

5) What is the basis for your choice in this category?

_____

_____

<u>Accuracy</u>

6) Grammar 1 2 3 4 5

7) Vocabulary 1 2 3 4 5

8) Cohesion 1 2 3 4 5

9) Use of standard English 1 2 3 4 5

10) What is the basis for your choice in this category?

_____

_____

Fluency

11) Hesitation 1 2 3 4 5

12) Repetition 1 2 3 4 5

13) Cohesion 1 2 3 4 5

14) Re-structuring sentences 1 2 3 4 5

15) Re-selecting inappropriate words 1 2 3 4 5

16) What is the basis for your choice in this category?

_____

_____

Strategic Capacity

17) The candidate is able to use achievement strategies to overcome lack of knowledge. (E.g. paraphrasing, non-linguistic strategies) 1 2 3 4 5

18) The candidate tends to use avoidance strategies to avoid using language that he/she has no control. (E.g. use of words such as 'thing' or 'stuff'.) 1 2 3 4 5

19) What is the basis for your choice in this category?

_____

_____

Topical knowledge

20)	The candidate gives a personal response.	1 2 3 4 5
21)	The candidate elaborates on his ideas.	1 2 3 4 5
22)	The candidate displays maturity in his ideas.	1 2 3 4 5
23)	The candidate displays a wide range of knowledge.	1 2 3 4 5
24)	The candidate displays a depth of knowledge.	1 2 3 4 5
25)	The candidate uses a wide range of vocabulary.	1 2 3 4 5
26)	The candidate uses a wide range of sentence structures.	1 2 3 4 5
27)	The candidate takes the initiative to engage the examiner.	1 2 3 4 5
28)	The candidate shares interesting ideas.	1 2 3 4 5
29)	The candidate expresses his/her idea cohesively and coherently.	1 2 3 4 5
30)	The candidate expresses his/her idea clearly.	1 2 3 4 5

31) What is the basis for your choice in this category?

_____

_____

Personal Characteristics

32)	The candidate interacts easily with the examiner.	1 2 3 4 5
33)	The candidate has a pleasant voice.	1 2 3 4 5

34) The candidate displays good grooming. 1 2 3 4 5

35) The candidate shows enthusiasm in what he is saying. 1 2 3 4 5

36) The candidate responds enthusiastically to the prompts. 1 2 3 4 5

37) The candidate is relaxed. 1 2 3 4 5

38) The candidate is confident. 1 2 3 4 5

39) The candidate displays the right level of formality. 1 2 3 4 5

40) The candidate is polite. 1 2 3 4 5

41) What is the basis for your choice in this category?

_____

_____

Exam Criteria

42) The candidate gives a personal response to the theme of the picture and the passage. 1 2 3 4 5

43) The candidate expresses himself/herself clearly and succinctly in the conversation, using appropriate vocabulary and structures. 1 2 3 4 5

44) The candidate discusses issues with the examiner stemming from the picture and the passage. 1 2 3 4 5

45) What is the basis for your choice in this category?

_____

_____

46) **Others (please state):**

_____	1 2 3 4 5
_____	1 2 3 4 5
_____	1 2 3 4 5

47) What is the basis for your choice in this category?

_____

_____

The End
Thank You