

Anchor rating in performance-based speaking tests

Gary J. Ockey

International University of Japan

Abstract

With the increasing utilization of performance-based assessments in English language programs, procedures must be enhanced for making the judged performances more reliable. This study investigates the potential of anchor rating through the utilization of many-facet Rasch measurement techniques in order to make ratings more reliable and more consistent within a given test administration as well as from one test administration to another. Five raters, two with a great deal of experience, one with some experience, and two with no experience in rating tests of speaking ability, assigned ratings to sixty graduate students ranging from intermediate to advanced levels of proficiency. The two experienced raters also re-rated videotaped tests of students that they had rated in two previous administrations of the test. Results showed that the two highly experienced raters stayed generally consistent in their ratings over time. A second finding of the study was that the pool of raters in the study varied significantly in their relative severity, but became less diverse with experience and training. New raters also showed marked changes in level of severity from one test administration to the next. It is argued that anchor rating may be a viable means of improving the reliability and comparability of performance-based scores when anchor raters are highly experienced and have a lot of stake in the test.

Key Words: performance assessment, many-facet Rasch measurement

INTRODUCTION

One important issue regarding the viability of performance-based tests is the reliability of the scores assigned by the raters. One limitation of such tests is that they cannot be considered valid measures of a student's proficiency if different raters in a test administration assign scores which are not at a uniform level of severity. Although scores assigned in performance-based tests are purported to represent a level on a defined metric, due to the nature of proficiency measures, these metrics are subject to various interpretations. For example, one rater could define the metric in a rather lenient manner whereas another rater could define the scale in a rather harsh manner. A second limitation of performance-based tests is that a student's measure may not be valid beyond the pool of raters in any given test administration. Just as different raters could differ in how they interpret a scoring metric, a group of raters in a given test administration might interpret a metric differently than a group of raters in another administration of the test. Interpretations of the metric could thus vary not only from one rater to another within a test administration but from one group of raters to another in different test administrations.

The traditional means of dealing with the problem of inconsistent interpretations of a scoring metric in a test administration has been to conduct rater norming sessions to help raters interpret the scale in the same way. In norming sessions, raters compare their interpretations of the scale with other raters in the group and try to form a consensus about the interpretation of the metric. Unfortunately, while rater norming sessions may help to lessen the differences in the

severity of raters, it has been shown that significant differences in rater severity remain after norming (Bonk & Ockey, Forthcoming; Ockey, 2001; Lynch & McNamara, 1998; Weigle, 1998; Fulcher, 1996; Myford, Marr, & Linacre, 1996; McNamara & Lumley, 1995; Raymond & Viswesvaran, 1991; Raymond, Webb, & Houston, 1991; Lunz, Wright, & Linacre, 1990).

As a further means of combating the inconsistency in scores assigned by different raters, many-facet Rasch measurement techniques have been employed after ratings have been assigned. Many-facet Rasch measurement techniques can be utilized to compare the relative severity of raters in a rater pool. After the severity of raters has been determined, the general application of these measurement procedures has been to utilize an average severity of the raters in establishing an interpretation of the level of severity of the scoring metric. Scores are then assigned to examinees so that each rater exhibits an equal level of severity in the assigned scores. Coupled with rater norming sessions, numerous researchers suggest many-facet Rasch measurement techniques be utilized to make scores assigned by different raters more reliable (Bonk & Ockey, Forthcoming; Wolfe, Bradley, & Myford, 2001; Lynch & McNamara, 1998; McNamara & Lumley, 1997; McNamara, 1996; Lunz & Stahl, 1990; Lunz, Wright, & Linacre, 1990) within a test administration. This employment of Rasch measurement techniques may help to diminish the problem of differing rater severity within a given administration of a test.

One means of alleviating the problem of a changing interpretation of the metric from one administration of a test to another is employing rater-training sessions conducted by experienced raters. In rater training sessions, experienced raters teach new raters to interpret the scoring metric in the same way that they interpret the scale by showing examples of prior tests, and then showing and explaining to the new raters the scores assigned to the examinees. The difference between a rater training session and a rater norming session is that in a rater training session, experienced raters teach new raters how to interpret the rating scale, while in rater norming sessions, old and new raters discuss how to interpret the scale and try to reach a consensus in their rating behavior. Thus, in rater training sessions, rather than various interpretations of the metric by each rater contributing to the interpretation of the scale, one or more experienced raters pass along their interpretation to the other raters. This procedure may therefore help to limit the differences in a changing rater pool impacting the interpretation of the scoring metric. However, just as differences in rater severity exist after rater norming sessions, it has been shown that they exist after rater training sessions (Ockey, 2001).

To further deal with the problem of a changing interpretation of the metric from one administration to the next, experienced raters can be utilized as anchor raters when employing Rasch measurement techniques. Such a procedure is analogous to conducting rater training sessions rather than rater norming sessions. Rather than utilizing the mean severity of the raters to

determine the amount that each score is adjusted by the many-facet Rasch measurement technique, ratings assigned by new raters can be anchored to scores assigned by raters that have experience rating students from similar populations of students. In addition to possibly increasing the reliability of the scores in the test administration, this might make the comparison of students' scores from one administration of the test to students' scores from other administrations of the test more accurate.

In order for anchor rating to be most effective, anchor raters would need to stay somewhat consistent in their ratings over time. Research in this area, unfortunately, suggests that raters may not stay consistent over time (Bonk & Ockey, Forthcoming; McNamara & Lumley, 1995). However, Lunz and Stahl (1990) obtained mixed results in their study which considered whether raters stay consistent over grading periods. They compared rater severity at different times of the day and different days of a test administration and found that raters stayed consistent when rating oral tests, but did not exhibit the same amount of rater severity when rating written exams. Other research may suggest that experience is a factor in maintaining a similar level of severity over time. For instance, Myford, Marr & Linacre (1996) found in their large scale study involving the Test of Written English readers that as the number of years of experience as a rater increased, the stability of rater severity level increased. Other researchers (Weigle, 1998; Wigglesworth, 1994; Wigglesworth, 1993) have found that raters become more consistent in their rating behavior with experience; in other words, they are better raters, but whether or not their level of severity remains the same over time was not considered in the studies. Thus, while some studies indicate that raters do not stay consistent over time, other studies that have considered rater experience as a factor seem to suggest that highly experienced raters may stay consistent in their ratings over time. The purpose of this study therefore is to consider the question of whether experienced raters can effectively act as anchor raters to make the scores more reliable within a given test administration and therefore more comparable from one test administration to another.

STUDY DESIGN

Raters

There were five raters in the study, all of whom had post-graduate training in applied linguistics or a related field. Two of the raters, one American male and one British male, were experienced raters that had designed the tests and the scoring metric and had worked with and tested similar populations of students for a number of years. They had conducted numerous rater training sessions and acted as raters for more than 300 tests of students in the examinee population prior to the tests in the study. They were also responsible for convincing administrators and students that the testing achieved its purpose of obtaining accurate measures of the students'

spoken English proficiency. As a result, they were chosen as the anchor raters in this study (designated as E1 and E2, respectively). One of the raters, an American woman, had acted as a rater in the two previous programs and was thus considered an experienced rater, but was not utilized as an anchor rater (designated as R3). The other two raters, a male from New Zealand and an American female, had no prior experience conducting oral proficiency tests nor working with a similar population of students before the program began. They were therefore considered new raters (designated as N4 and N5, respectively).

Rater training

Rater training took place before each test. In each training session, raters were provided with the scoring metric (Appendix 1) and shown videos of tests conducted in previous years, along with the scores assigned to those students. After the first test, raters were retrained with the same procedures along with feedback on their performance on the first test based on Facets measures rulers (see Figure 1). Raters were provided with feedback on their rating behavior in each of the five categories of the scoring metric regarding their severity and degree of utilization of the scoring metric as compared to the anchor raters.

Examinees

All sixty students from the International University of Japan's (IUJ's) 2001 Intensive English Program (IEP) were considered in the study.ⁱ These students were preparing to study in a postgraduate program in IUJ's English medium university. The students' ages ranged from middle twenties to late thirties. Fifty of the students were men and 10 of the students were women. The students were from the following countries: Japan, 30; Indonesia, 13; Cambodia, 4; Laos, 3; Malaysia, 2; Korea, 2; Kenya, 2; and one student from Thailand, Ghana, Tanzania, and Ethiopia. TOEFL scores obtained a couple of days before the first test in the study ranged from 657 to 447, with a mean of 534, and a standard deviation of 40. The majority of the scores clustered around the mean with only a few scores below 500 and above 600.ⁱⁱ TOEFL scores may not have reflected the speaking proficiencies of the students, however, since TOEFL does not directly measure speaking ability, and all five of the African speakers were very proficient speakers but did not all earn high TOEFL scores.

Tests

To consider whether raters stay consistent over time, videos of tests given in the 1999 and 2000 administrations of the test were re-rated in 2001, a couple of weeks before the administration of the 2001 test. To compare anchored and non-anchored tests, the two tests in the 2001

administration were utilized; the two tests in the study were separated by about three weeks. The purpose of the tests is to place students into English classes at the end of the program and provide students with information about their weaknesses and strengths as the program progresses.

The tests in the study were designed as group discussion tests with three members in each group. Test procedures require that students test twice in each test administration,ⁱⁱⁱ grouped with different students for each test. The examiner sits outside of the group and does not participate in the discussion. After brief introductions of examinees and examiner, the test begins with the examiner providing the participants with a written prompt and asking them to begin the discussion. After 12 minutes of discussion, the examiner dismisses the students. The examiner's only task during the test is to assign ratings to students based on the five categories in the scoring metric.^{iv} All tests are videotaped so that after the test, raters can assign additional ratings to the students.^v Thus, each student receives ratings from two different raters in real time (in separate test administrations) and at least a third (usually a fourth and in a few cases a fifth) rating based on a video of his test.

Model for analysis

The many-facet Rasch model via the Facets 3.0 (Linacre 1996) program was utilized to aid in the analysis of the data. The rating scale model (Andrich, 1978) was chosen for this purpose, and the five categories of the scale (see Appendix 1) were combined in the analysis since raters reported that they do not rate categories separately; rather, they tend to compensate for relative severity in one category with relative leniency in another. For instance, if they feel an examinee is between an eight and a nine in grammar, they may err on the side of leniency by assigning a score of nine for grammar and then make up for it by erring on the severe side on another category such as vocabulary when they are not sure which of two scores to assign. A model such as the partial credit model may not make sense in such an analysis. The model for the analysis was:

$$\text{Log } (P_{njk}/P_{njk-1}) = B_n - C_j - F_k$$

P_{njk} is the probability of student n being awarded a rating of k by rater j .

P_{njk-1} is the probability of student n being given a rating of $k-1$ by rater j .

B_n is the ability of student n .

C_j is the severity of rater j .

F_k is the difficulty of the step from category $k-1$ to k .

(Linacre, 1994)

Figure 1 shows an example of the *all Facets vertical ruler*. Facets output places the variables on a common logit scale which allows each measure to be compared. The far left column, titled *measure*, provides the logit measure for each variable used in the study. This scale is

Figure 1 Example of all Facets vertical ruler

[Measure +Examinee -Rater		[Scale]
+ 2 +		+(49) +

	*	
		43

		42

	*	41

+ 1 + *	+	+ 40 +
		39
	*	38
	*	37
	**	36
		35
	*	N5 34
	*	33
	***	32
	****	31
* 0 * ***	* N4	* 30 *
	*****	E2 R3 29
	***	E1 28
	*****	27
	*****	26
	*****	25
	*****	23
	*****	22
	*	21
		20
+ -1 + **	+	+ --- +
	**	19
		18
	*	---
		17

		16

+ -2 +	+	+(12) +

calibrated to have a mean of 0.^{vi} The second column presents the scores of examinees in the study. Each asterisk represents a student in the study. Students with high logit values are more proficient than students with low logit values. Thus a student at the top of the scale is more able than a student at the lower part of the scale. The third column represents the raters in the study. E1 and E2 represent the experienced raters (the anchor raters), R3 the returning rater, and N4 and N5 the new raters. All raters would be at the mean of 0 logits if they are equally severe in their assignment of ratings; this, of course, is rarely the case. Raters with high logit values (positive numbers) are considered severe raters while those with low logit values (negative numbers) are considered lenient raters. The fourth column, labeled *scale*, represents the scale in the study. It is based on a combination of the five categories of the rating scale (see Appendix 1). This scale shows the score a student would receive if his rater were of average severity. For example, an average examinee with an average rater would receive a score of 30 on the 50-point rating scale.

RESULTS

Consistency of experienced raters over time

To determine whether the experienced raters in this study stayed consistent over time, the anchor raters (designated as E1 and E2) rated videotapes of students that they had rated in previous years, 1999 and 2000. There was a little less than two years from the first 1999 rating to the second 1999 rating, and a little less than one year from the first 2000 rating to the second 2000 rating. Since each rater had only rated a small number of examinees in each year, it was only possible for each rater to provide a second rating for 9 examinees from each year of the study. Because of the small number of ratings, the intra-rater correlations were combined for the two years. The results can be seen in Table 1.

Table 1 Intra-rating correlations of anchor raters

Intra-rating correlation (1999 and 2000 compared to 2001)	Mean		Standard Deviation	
0.89	1999/2000	33.9	1999/2000	6.1
	2001	32.6	2001	6.4

N = 36, 18 for each rater, 9 from the 1999 and 9 from the 2000 administrations.

As can be seen in Table 1, the raters show a high degree of intra-rater reliability when rating examinees which they have previously rated. When the ratings given by the raters in 1999 and 2000 are compared to the rating assigned to the same examinees based on a video of the first test in 2001, there is a 0.89 intra-rater reliability. This could be considered a rather high intra-rater correlation suggesting that the raters stay quite consistent in their ratings over time. When the mean rating in 1999/2000 is compared to the mean rating in 2001, there is little difference, 33.9 to

32.6. This difference of slightly more than one point on the 50-point metric may suggest little change in raters from one time period to another.^{vii} Considering the small N size in this study, the findings are very tentative. While these findings do not appear to support the findings of some researchers (Bonk & Ockey, Forthcoming; McNamara & Lumley, 1995) who found that raters do not stay consistent over time, this may be because these studies dealt with raters that were not highly experienced, unlike the two raters considered in the present study. The findings are consistent with the findings of Myford, Marr & Linacre (1996) who found that there is a correlation between consistency and (essay) rater experience and the findings of Weigle (1998), Wigglesworth (1994), and Wigglesworth (1993) who found that raters become better with practice. Thus, it appears that at least some highly experienced raters do stay generally consistent over time, suggesting that it may be possible to use these raters as anchors in order to be able to make scores more reliable and more comparable from one test administration to another.

Anchored versus non-anchored rating on the first test

After anchor raters were shown to be generally consistent in the scores they assigned over time, test scores in the 2001 administration of the test were compared when being anchored by the experienced raters and when not being anchored by the experienced raters. This was done to determine if there was further evidence to suggest that anchor rating is a viable technique for improving the reliability of a test. As would be expected based on previous findings (Bonk & Ockey, Forthcoming; Ockey, 2001; Lynch & McNamara, 1998; Fulcher, 1996; Myford, Marr, & Linacre, 1996; McNamara & Lumley, 1995; Lunz, Wright, & Linacre, 1990), raters varied significantly in their degree of severity in assigning ratings to examinees (displayed numerically in Table 2 and graphically in Figure 2). The error-corrected standard deviation of the raters (separation) as measured by Facets was 3.94 and the likelihood of this being a real difference (reliability) was .94. The chi-squared statistic asserts that this represents a significant difference. It can be seen that new rater N5 is the most severe rater in the study followed by new rater N4. The two experienced raters, E1 and E2 are the most lenient raters, although E2 is only slightly more severe than the returning rater, R3. It is clear that if raw scores were utilized, ratings assigned to students would not be fair; there is about a six-point difference between the most severe rater (N5) and the most lenient rater (E1). Even if two scores were assigned to each examinee, and the average score was taken, it would still mean that examinees rated by raters N4 and N5 would be treated more harshly than examinees rated by raters E1 and E2.

Figure 2 Comparison of Rater anchored and non-anchored scores on the first test

First Test Non-anchored				First Test Anchored			
[Measure]	+Examinee	-Rater	[Scale]	[Measure]	+Examinee	-Rater	[Scale]
+ 2 +		+	+(49) +	+ 2 + *		+	+(49) +
			---				---
	*						
			43				
							43
			---				---
			42		*		42
			---				---
	*		41				41
			---		*		---
+ 1 + *		+	+ 40 +	+ 1 + *		+	+ 40 +
			39		*		39
	*		38		*		38
	*		37		*		37
	**		36				36
			35		**	N5	35
	*	N5	34		**		34
	*		33		****		33
	***		32		****	N4	32
	****		31		****	E2 R3	31
* 0 * ****		* N4	* 30 *	* 0 * ****		* *	* 30 *
	*****	E2 R3	29		*****	E1	29
	***	E1	28		*****		28
	*****		27		*****		27
	*****		26		****		26
	*****		25		*****		25
	*****		23		*		23
	****		22		*		22
	*		21		**		21
			20		*		20
+ -1 + **		+	+ --- +	+ -1 + *		+	+ --- +
	**		19				19
			18		*		18
	*		---				---
			17				17
			---				---
			16				16
			---				---
+ -2 +		+	+(12) +	+ -2 +		+	+(12) +

Table 2 Comparison of raters in non-anchored and anchored first test

Rater	Non-anchored (logit value)	Anchored (logit value)
E1 (Anchor rater)	-.25	-.08
E2 (Anchor rater)	-.08	.07
R3	-.07	.10
N4	.02	.18
N5	.38	.54

RMSE (Model) .05, Adjusted SD .20, Separation 3.94, Reliability .94.
Chi-squared 76.5, Degrees of Freedom 4, Significant at .01 level.

When no anchor rating takes place, and instead, an average severity is utilized, the two new raters will cause the scores given by the experienced raters to be shifted down. Considering the fact that the experienced raters (E1 and E2) were shown to exhibit a large degree of intra-rating reliability over time, it is likely that their ratings represent a more consistent interpretation of the metric than that of the new raters. Thus allowing the experienced raters' scores to be shifted downward by the new raters probably would result in making a score on this test unequal to a score on a test in the 1999 or 2000 administrations of the test; students on the 2001 test would be scored more harshly than students who took an earlier administration of the test. When the scores are anchored half way between E1 and E2, a more probable picture of the raters can be seen (Figure 2 and Table 2). N5 is very severe, and N4 is somewhat severe. Due to the anchoring, E1 and E2 cluster around the mean of 0 logits. The returning rater, R3 is also near the mean of 0 logits, further suggesting that it is the new raters exhibiting severe rating behavior rather than the experienced raters being somewhat lenient and the new raters being somewhat severe.

After the first test, raters were given feedback on their rating performance based in part on the Facets output shown in Figure 2 and Table 2. In addition, they were given further training by rating videos of previous tests. This additional training took place immediately before the second test was conducted three weeks later.

Anchored versus non-anchored rating on the second test

The results of the second test can be seen in Figure 3 and Table 3. Not surprisingly, even with some experience and further training, the raters are still quite different in the severity of the ratings they assign. The error-corrected standard deviation of the raters (separation) was 2.37 times the root mean-square estimation error. The reliability of this separation was 0.85. The chi-squared statistic confirms that this is a significant difference. It should be noted, however, that the group as a whole did show less deviation than in the first test; the error-corrected standard deviation on the first test was 3.94 as compared to 2.37 on the second test.

Figure 3 Comparison of rater anchored and non-anchored scores on the second test

Second Test (Non-Anchored)				Second Test (Anchored)			
Measure	Examinee	Rater	Scale	Measure	Examinee	Rater	Scale
+	3	+	+(48)	+	3	+	+(48)
			47				47
	*		---		*		---
			46				46
			---				---
+	2	+	45	+	2	+	45
	*		---				---
	*		44		*		44
	*		---		*		---
			43		*		43
			---				---
	**		42		**		42
			---				---
			41		**		41
+	1	+	40	+	1	+	40
	*		---				---
			39				39
			---		*		---
	*		38		*		38
			---				---
			37		*		37
			---				---
	**	E2	36			E2	36
	*****	E1	35		*		35
*	*****	R3	34	*	****		34
*	*	N4	33	*	*****	E1	33
	**	N5	32		****	N4 R3	32
	*****		31		****	N5	31
	*****		30		****		30
	*****		---		**		---
	**		29		*****		29
	*		28		***		28
	*		---		*		---
	***		27		*		27
+	-1	+	26	+	-1	+	26
	***		---		***		---
	*****		25		*		25
	*		---		*****		---
	****		24		**		24
			23		*****		23
			22				22
			---				---
			21				21
			20				20
+	-2	+	+(18)	+	-2	+	+(18)

Table 3 Comparison of values of raters in non-anchored and anchored second test

Raters	Non-anchored (logit scale)	Anchored (logit scale)
E1 (Anchor rater)	.09	-.07
E2 (Anchor rater)	.23	.07
R3	-.04	-.19
N4	-.07	-.23
N5	-.21	-.37

RMSE (Model) .06, Adj. S.D. .14, Separation 2.37, Reliability .85.
Chi-squared 31.6, degrees of freedom 4, significant at .01 level.

Interestingly, the non-anchored scores show that the raters made almost a complete reversal in their level of severity. N5 and N4 became the most lenient raters while E1 and E2 became the most severe raters. R3 stayed in the middle. This may have been due to the new raters overcompensating for their severity on the first test because they were told that they were too severe in their rating behavior. It should be noted, however, that rater N5, the most deviant rater on the first test, moved much closer to the mean of 0 logits, suggesting that rater training and experience had a positive effect on his ability to interpret the metric in the same way as the experienced raters. It is also important to point out that on the anchored tests, E1 exhibits an almost identical logit difference from E2. In both cases, E1 is more lenient than E2, and the difference is almost identical, .15 logits on the first test and .14 logits on the second test. This is further evidence that these two raters are exhibiting consistency in their rating behavior.

DISCUSSION

Consistency of experienced raters

This study suggests that utilizing the technique of anchor rating with experienced raters may help to make scores more reliable within a test administration and more comparable from one test administration to another. The findings suggest that experienced raters with a great deal of stake in a test may be generally consistent in their ratings from one test administration to another. In this study, both experienced raters showed a great deal of intra-rater reliability over a two year period. These findings are consistent with the large-scale findings of Myford, Marr, & Linacre (1996) who reported that rating experience and intra-rater reliability are significantly correlated. Other studies that have shown that raters do not stay consistent in their rating behavior over time (Bonk & Ockey, Forthcoming; McNamara & Lumley, 1995; Lunz & Stahl, 1990) give no indication that they dealt with experienced raters with a great deal of stake in the test. It thus seems that rating experience and stake in a test might be important in determining whether raters change over time. It is also likely that rater behavior is based on various other factors as well, suggesting that each rater should be tested to see if he displays consistent rating behavior over time before being selected as an anchor rater. It should not be assumed that all experienced raters with a lot of stake in a test will necessarily exhibit consistent rating behavior over time, but based on these findings, it is a possibility that should be considered. There is a great need to conduct studies to test whether or not experienced raters with high stake in a test are likely to stay consistent from one test administration to another and can thus be utilized as anchor raters. While this study suggests such a case, the results of this study are based on small N sizes. It should be pointed out that rater behavior is likely affected by personality, mood, and other factors which cannot be controlled for in a test. However, factors, such as teaching experience, professional training, experience as a rater

in performance-based tests, and experience with the given population of examinees which can be controlled for in a test should be further studied.

Rating experience and consistency

The study also suggests that rating behavior within a group of raters becomes more consistent with rating experience. These findings are consistent with the growing body of research in this area (Weigle, 1998; Wigglesworth, 1994; Wigglesworth, 1993). Ideally, the same pool of raters could be utilized repeatedly after they have reached a general level of reliability. Unfortunately, many programs (like ours) deal with new inexperienced raters almost every test administration. In addition to rater training, implementing procedures to cope with the problem of new raters and the inconsistencies that they are likely to exhibit when they begin is clearly important.

Rasch measurement techniques

The study also adds to the growing body of research which suggests that many-facet Rasch measurement techniques (or other such techniques) be utilized to increase the stability, and thus interpretability, of test scores (Bonk & Ockey, Forthcoming; Ockey, 2001; Wolfe, Bradley, & Myford, 2001; Lynch & McNamara, 1998; McNamara & Lumley, 1997; McNamara, 1996; Lunz, Wright, & Linacre, 1990). The large difference observed in scores given by different raters in this study suggests that it would be irresponsible to assign examinees raw scores on performance-based tests.

Anchor rating

A further implication of the study is that anchor rating with experienced raters may result in more reliable scores, both in a given test administration and from one test administration to another. This is supported in the study by the two experienced (anchor) raters displaying almost identical levels of severity (compared to each other) from one test administration to the next while the two new raters, especially N5, moved dramatically (from severity to leniency) from one test administration to the next. This movement to greater leniency by the new raters is also not surprising based on previous research by Ockey (2001), who reported that raters tended to inflate scores as students progressed in the program so that students felt like they were improving. It would seem that experienced raters with high stake in the tests would have a better understanding of the importance of the accuracy of the scores and would not be as likely to inflate scores as a program progresses.^{viii} The tendency of raters to inflate scores at the end of a program, coupled with the feedback that the new raters received suggesting that they were too severe in their ratings

on the first test, probably led to their shift to leniency on the second test. Whatever the reason for the shift in the assignment of scores of the new raters, it is quite clear that their rating behavior did not remain stable from one administration of the test to another, suggesting a need for procedures, such as anchor rating, to help control for such a shift.

Conclusion

The results suggest that at least some experienced raters with high stakes in a test stay generally consistent over time and that raters become better with rating experience. This study also makes a case for the value of anchor rating with experienced raters. Finally, aside from the evidence presented here, intuition suggests that anchor rating can be an effective means of making scores more reliable. It seems that utilizing raters that have a great deal of stake in the test and who have rating experience as anchor raters makes more sense than allowing various interpretations of the scoring metric by a changing pool of raters who may not all have a great amount invested in the test. Considering the fact that many programs have a revolving door when it comes to teachers and that new raters have been shown to be inconsistent in rating behavior, it seems that utilizing procedures, such as anchor rating, to stabilize the scores is a necessity.

ⁱ One Indonesian male did not participate in the first administration of the test.

ⁱⁱ See Ockey, 1999 for further details on the proficiency of students in the IEP.

ⁱⁱⁱ Thus, it is assumed that a student's proficiency is measured by the test, rather than his performance on a particular test.

^{iv} Considerable efforts are made before the test is given to assure that students are familiar with the test design and scoring metric, including having the students observe a video of a test and practicing such a situation. See Ockey 2001 for a complete description of the test.

^v Previous research found that there was no significant difference in ratings given in real time and ratings assigned based on viewing a video of the test (Ockey, 2001).

^{vi} For an explanation of the logit scale see Linacre (1994).

^{vii} However, it should be noted that a slight change occurred (toward severity).

^{viii} However, it might be that the small difference in mean average of the anchor raters from the 1999/2000 tests (which were based on test results at the end of the program) to the 2001 test (rated at the beginning of the next year) may also be explained by this rating behavior. It could be that even highly experienced raters with a great deal of stake in a test have a tendency to rate higher on a final test than on a test earlier in the program. Unfortunately, the small N sizes in this study make it impossible to answer this question.

REFERENCES

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika* 43, 561-573.
- Bonk, W. & Ockey, G. (Forthcoming). A many-facet Rasch analysis of the group oral discussion task. *Language Testing*.
- Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing*, 13, 23-51.
- Linacre, J. (1994). Many-Facet Rasch Measurement. Chicago: Mesa Press.
- Linacre, J. (1996). Facets 3.0 [Computer Program]. Chicago: Mesa Press.
- Lunz, M., & Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, 13, 425-444.
- Lunz, M., Wright, B., & Linacre, J. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3, 331-345.
- Lynch, B. & McNamara, T. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15, 158-180.
- McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Addison Wesley Longman Limited.
- McNamara, T. F. & Lumley, T. (1995). Rater Characteristics and rater bias: Implications for training. *Language Testing*, 12, 54-71.
- McNamara, T. F. & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing*, 14, 140-156.
- Myford, C., Marr, D., & Linacre, M. (1996, May). Reader calibration and its potential role in equating for the Test of Written English. *Educational Testing Service, TOEFL Research Reports*, 52.
- Ockey, G. (1999). The use of TOEFL to measure a change in proficiency. *Working Papers on Language Acquisition and Education: International University of Japan*, 10, 1-12.
- Ockey, G. (2001). Is the oral interview superior to the group oral? *Working Papers on Language Acquisition and Education: International University of Japan*, 11, 22-41.
- Raymond, M. & Viswesvaran, C. (1991, December). Least-squares models to correct for rater effects in performance assessment. *ACT Research Report Series*, 91-8.
- Raymond, M., Webb, L., & Houston, W. (1991). Correcting performance-rating errors in oral examinations. *Evaluation & the Health Professions*, 14, 100-122.

- Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263-287.
- Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10, 305-335.
- Wigglesworth, G. (1994, July). The investigation of rater and task variability using multi-faceted measurement. A report for the National Centre of English Language Teaching and Research. Macquarie University, New South Wales.
- Wolfe, E., Bradley, M., & Myford, C. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement*, 2, 256-280.

Appendix 1

Descriptor bands for speaking tests

	Comprehensibility	Fluency	Grammar	Vocabulary usage	Communicative Skills
10	Rarely mispronounces, able to speak with nearly perfect pronunciation, intonation, and rhythm, little or no foreign accent	Fluent speech, speaks confidently and effortlessly, speech is smooth and natural	Uses high level discourse with near perfect accuracy, shows an ability to use the full range of grammatical structures effortlessly and accurately which are needed to achieve the task	Confidently uses wide range of technical and general vocabulary precisely and effectively	Shows confidence and naturalness, shows ability to negotiate meaning, shows how ideas or opinions are related, may initiate conversation, completes task effectively
9	Pronunciation is clear, occasionally mispronounces or has non-perfect intonation or rhythm, articulation is clear, has mastered all sounds, accent may sound foreign, but does not interfere with understanding	Speaks with confidence, but has a few unnatural pauses, occasionally gropes for words unnaturally	Shows ability to use nearly the full range of grammatical structures, but may make some errors when using some complex sentence types, errors do not interfere with meaning	Shows range of technical vocabulary which is sufficient for task, but fine shades of meaning are occasionally inappropriate	Generally confident, responds appropriately to an opinion, shows ability to negotiate meaning, shows how ideas are related, completes task effectively
8					
7	Pronunciation is not perfect but can be understood without concentrated listening, articulation is generally clear, may not have mastered all sounds	Speech is a little hesitant, has some unnatural rephrasing and groping for words	May not have mastered full range of structures, but uses complex and simple sentences, may make a few global errors, has no trouble completing task	Has sufficient vocabulary to complete task, but may not use it appropriately, may use technical vocabulary, but not always effectively	Somewhat confident, responds appropriately when asked for opinion, completes task somewhat effectively
6					
5	Sometimes mispronounces, may require concentrated listening, but is completely understandable, may not articulate clearly, may not have mastered some sounds	Speech is often hesitant, frequent unnatural rephrasing and groping for words,	May use simple (but generally accurate) sentences to express meaning, complex sentences are used but often inaccurate, can express desired meaning, errors may occasionally interfere with meaning, is able to complete task	Vocabulary is adequate for achieving task, but often used inappropriately. Does not accurately use technical terms used in the field	Not confident, shows agreement or disagreement to opinions at the surface level but not at the discourse level, completes task but not effectively
4					
3	Frequently mispronounces, accent impedes comprehensibility, requires concentrated listening but is generally comprehensible	Strained speech, often groping for words, some long unnatural pauses (except for routine phrases)	Relies mostly on simple sentences which are often inaccurate, cannot control complex sentences, mistakes often impede meaning, has difficulty completing task	Lacks the necessary vocabulary to discuss the topic with any sophistication	May use simple phrases to show agreement or disagreement, but does not relate ideas at discourse level, task may not be completed
2					
1	Frequently mispronounces, heavy accent, even with concentrated listening often incomprehensible	Fragmented speech that is so halting that conversation is virtually impossible	Cannot control even simple sentences, grammar is not sufficient to complete task	Vocabulary is inadequate to achieve the task	May require prompting, produces monologues which are unrelated, does not complete task
0					

A score in the lower part of the box indicates that a student has not completely mastered the level.