Rater's Subjectivity in an Oral Test

Ilhong Kim
(Seojeong College)

Kim, Ilhong. 2004. Rater's Subjectivity in an Oral Test. *The Linguistic Association of Korea Journal, 12*(3), 151-170. This experimental study investigated the seriousness of rater's subjectivity in an oral test. The researcher analyzed the verbal test results of English native speakers, who are assumed to be reliable and competent raters only because they are from English speaking countries. The researcher correlated the raters' and inter-raters' scores to those of the examinees' ranks, and investigated the English native speakers' scoring tendencies in the process of their evaluation. From comparing the statistics for the discrepancies between the examinees' scores and rankings, this study admonishes untrained evaluators to be aware of subjectivity, and also suggests that subjective tests should not be used for examinees' rank order even though the results are statistically very significant.

Key Words: rater, examinee, rater's subjectivity, rank order, oral test, subjective test, evaluation, rater training

1. Introduction

It is needless to say that evaluations should be carried out fairly and unbiasedly since their results can affect examinees seriously. Fair and unbiased evaluations are, however, not simple to carry out. Biased and unfair evaluations can weaken test reliability, validity, and could be a disadvantage to an examinee. Moreover, it may cause even social and ethic problems if a subjective test result is used as a cutoff point with a rank order. Therefore, raters should try to design tests that offset the subjectivity that makes them less reliable and valid.

Of the four language skills, verbal evaluations are the most difficult

to do and are most vulnerable to rater's subjectivity. If evaluation criteria are classified too detailedly, an evaluation can give the impression of being a rather mechanical process, and if the evaluation criteria are vague, they can weaken the test reliability because of the rater's subjectivity. The following quote explains a rater's unavoidable subjectivity in oral tests.

> An oral test is a personal encounter between two human beings; it is designed by humans, administered by humans, taken by humans, and made by humans, and it would be a surrender of the test designer's responsibility to allow the evaluation and development of this wholly human activity to be dictated by the statistical savage-machine (Underhill, 1987, p. 105).

A rater should prove an examinee's communicative competence with quantifiable data through evaluation by using his/her spontaneous subjective judgement. The most reliable way to get objective quantifiable data by reducing a rater's subjectivity is by using effective evaluation criteria. Effective evaluation criteria can only reduce a rater's subjectivity. A rater's subjectivity is very crucial because it is relevant to a test's reliability and validity. A reliable test should have test reliability and content reliability as well as score reliability (Harris, 1969). A test's reliability should be consistent with a credible rater's reliability as demonstrated through repeated evaluations, and also should have credible inter-rater's reliability. The fewer discrepancies of standard deviation among raters and their scores, the more credible the test.

Theoretically, a rater should give a score similar to that of other raters. Giving exactly the same score in verbal evaluation as other raters, however, is totally impossible, as in gymnastic games or fine art contests. Verbal evaluations are unavoidable considering the trend of putting emphasis on communicative language skills and its backwash effect (Hughes, 1989) in English education these days. If a perfect verbal evaluation is not feasible, an effective verbal evaluation method

whose inter-subjectivity is credible can be an alternative solution, especially in English education in Korea.

In this study, the researcher attempts to analyze the verbal test results of English native speakers, who are assumed to be reliable and competent raters only because they are from English speaking countries. The researcher wants to correlate the raters' and inter-raters' scores to those of the examinees' ranks, and investigate the English native speakers' scoring tendencies in the process of their evaluation. By comparing the statistics for the discrepancies between the examinees' scores and rankings, this study admonishes untrained evaluators to be aware of subjectivity, and also suggests reputedly effective methods of verbal evaluation that reduce a rater's subjective influence.

2. Background - Rating and Rater's Subjectivity

The verbal evaluation method developed as communicative language skills began to be emphasized in cognitive approaches. As communicative language skills became the EFL focus in the 1970s (Chastin, 1971; Carroll & Freedle, 1972; Curran, 1976; Terrell, 1977), the focus of language evaluation changed from 'language usage' to 'language use.' Since the 1980s, applied linguists have published research papers (Buck, 1990; Flowerdew, 1994; Hansen & Jensen, 1994) criticizing the unidimensional approach, which had been predominant (Shin, 2003).

In general, language competence tests can be roughly classified into two areas. One is the integrative test, which evaluates an examinee's language skill as a whole; and the other is the discrete-point one, which analyzes structural factors such as phonemes, syntax, etc. The integrative test's weak point is that a total score is affected by only one or two facets of the examinees' language abilities, so that it is rather a subjective test (Weir, 1990). Since discrete-point tests are not free from rater's subjectivity, integrative tests are more reliable and valid if, that is, raters are attentive (Jacobs et al., 1981; Charney, 1984; Milanovic et al., 1993).

Discrete-point test methods are better for unexperienced raters since the rater can measure the examinee's ability in accordance with guided criteria relatively easy, while an integrative test is more suitable for professional raters. According to the Korean Educational Development Institute (KEDI), communicative language ability tests in Korea put more emphasis on discrete-point tests without checking an examinee's four language skills equally, relying more upon accuracy than on fluency (KEDI, 1997, pp. 107-110).

Interview tests are problematic because they are done not with natural communication, but by one-sided questions whereby examinees feel uncomfortable because they have no choice but to answer the question. When a rater evaluates an examinee in an interview, the rater judges the examinee's ability in accordance with the evaluation criteria. The evaluation is, then, influenced by subjectivity, which is, by definition, subject to change. The following quote summarizes the importance of using criteria for reducing a rater's subjectivity

> The use of an analytic scale that comprises a manageable list of criteria, each weighted appropriately for programme objectives, goes far to reduce subjectivity (Pino, 1989, p. 488).

The crucial fact of successful oral tests is how well a rater designs the criteria and how much consistency the rater keeps in applying the criteria. A rater, consequently, is required to design evaluation criteria so as to meet the character and purpose of the evaluation case by case. It is not a simple job to describe an evaluation's criteria in detail. Professional language evaluators usually set the evaluation criteria for a test. Even credible tests like ACTFL (American Council of Teachers of Foreign Languages) or FSI (Foreign Service Institute) (Johnson, 1992; Oller, 1976; Spolsky, 1968; Stubbs & Tucker, 1974) are criticized for being defective (Stansfield & Kenyon, 1992).

Following are samples of criteria for aural tests currently being used in Korea: Accuracy of language use, appropriateness, fluency, duration of speech (Kim, 1996, p. 344); Accuracy, range, appropriateness, fluency,

interaction, construction, pronunciation, overall impression (Lee, 1996, pp. 665-669). Carroll determines audibility, complexity, range, speed, fluency, accuracy, appropriateness, repetition, hesitation, and overall impression as an oral test's criteria. Most major English proficiency tests use similar criteria.

The scoring systems of figure skating or heavy gymnastics in the Olympics are designed to diminish subjectivity. The scoring system of heavy gymnastics requires two groups of raters: group A (two persons) and group B (six persons). The two persons of group A suggest an adjusted score by mutual consent only for the performance, and each member of group B suggests individual scores for deduction. The highest and lowest points of the group B are excluded and the remaining four raters' average score is deducted from the suggested score of group A. This type of scoring system is effectively reduces rater's subjectivity.

A rater is apt to evaluate by using his or her own rating strategy or intuition rather than a required evaluation guideline (Shin, 2001a; Shin & Jang, 2002). Due to his subjectivity, a rater cannot consistently maintain the standards of even well-constructed test criteria. Therefore, seeking a method for making a rater's subjectivity consistent is indispensable. Special training is a must for raters to reduce their and other inter-raters' errors, to keep consistent subjectivity. The researcher analyzed the oral test results scored by English native speakers untrained in evaluation to find how seriously a rater's subjectivity affects an examinee's score.

3. Methodology

3.1. Subjects: The subjects in this study consists two different groups; one group is being examined and the other group is rating the examinees.

3.1.1. Examinees: The examinees selected in this study were incumbent secondary English teachers (n=32; 15 males, 17 females; the

majority ranging in age from 30 to 50) who participated in an intensive English program in Korea. Among them, eight had experiences of having given oral tests. The majority of the examinees responded to their English listening ability and pronunciation questions with 'disagree (2)' or 'average (3)'; to speaking ability and accuracy with 'average (3)' or 'agree (4).' More than half of the participants perceived that accurate pronunciation contributes to fluent communication in English by about 70-80%; speakers' gestures affects about 30%. Though all the examinee subjects were experienced secondary English teachers, for efficiency, they were requested to acknowledge the evaluation criteria before the interview started.

3.1.2. Raters: Eight English native speakers participated as raters in this study. Three raters (R1, R2, R3) conducted the first evaluation. They are all Canadians who have taught as EFL instructors in Korea for at least a year, and each holds either a TEFL or a TESOL certificate. The second evaluation was conducted by another five English native speakers (three Americans: R4, R5, R6; two Canadians: R7, R8). R6 holds a TESOL certificate but the others didn't major in English education. All the five raters have at least 3-4 years of teaching experience in Korea, but none of them has been specially trained for scoring.

3.2. Interview Questions: Generally a test should be valid in five different areas: construction, content, concurrent, predictive, and face. The most important factor a test designer should consider is that content should represent speaking ability and be relevant to the examinee's needs and concerns (Cho & Lee, 1991). The questions used for this study were prepared considering the examinees' occupations. The questions were comprised of four different hypothetical situations the examinees may face in a foreign country.

3.2.1. Type 1 (3 questions): Situations in need

Q : Imagine this: A travel agent reserved a hotel for you in

Switzerland. You're scheduled to stay there for three days. But when you arrive, the front desk clerk says, "We have no room for you here, I'm sorry." What would you do?
Q : Imagine this: Some friends invited you to their new apartment for a celebration or a party. They gave you the address, but when you arrive, the lights are off, the door is locked, and your mobile phone has no battery power. What would you do?
Q : Imagine this: Your travel agent reserved an airport bus to drive you from the Toronto International Airport to Toronto University. But when you arrive at the bus, the bus driver says, "Your ticket is not valid." What would you do?

3.2.2. Type 2 (1 question): Food culture

Q : Americans are famous for hamburgers, Italians are famous for pizza, and Japanese for Sushi. Which of these foods is your favorite, and why?

3.2.3. Type 3 (1 question): Korean culture and gift

Q : You must buy a gift for your home-stay family in New Zealand. You're in New Zealand, you're staying with a home-stay family. You could buy a traditional Korean fan, or some Boseong green tea or some Korean red ginseng. Which one would you choose for a gift, and why?

3.2.4. Type 4 (1 question): Teaching methodology

Q : You're teaching an English workshop, and you must choose a final activity. You could choose a dictation activity, or a song activity, or a sentence scramble activity. Which one would you choose, and why?

The contents of the questions above are sufficiently relevant and familiar to secondary teachers' needs and concerns, and relatively easy to answer, which was found through in-person interviews with the test-takers after the interview was over. No time limit was given for answering so as to get enough speech for evaluation. No examinee took more than 5 minutes to finish answering the questions; the majority

took about 3-4 minutes to answer. The length of the obtained speeches did not conform to the assertion that a examinee's speech be about 15-30 minutes (Hughes, 1989).

The researcher, however, had no choice but to accept this limitation, since the examinees' answers to the questions were relatively short; but the duration of the answers were enough to measure the examinees' abilities. The face validity of a test is a non-scientific intuitive standard held by the impressions of its raters and the test takers involved. The face validity of the test content used in this study proved to be a good model of an oral test by the examinees and raters involved. Predictive validity and concurrent validity were excluded in this study since it is impossible to measure them in a short period, and they are not relevant to the purpose of the study.

3.3. Instrument: The evaluation criteria used in this study (Appendix 1) were prepared by a training institute to measure the trainees' English speaking proficiency. The evaluation criteria comprise 5 areas: pronunciation, audibility, fluency, response time/detail/content, comprehension of content. The reliability coefficient of the evaluation criteria was alpha .890, which is higher than the generally accepted coefficient level as a reliable one, which is alpha .5. The evaluation criteria used in the study are very reliable consequently.

3.4. Procedure: The evaluation was conducted by English native raters in accordance with the evaluation criteria on the examinees' answers to the three randomly selected questions. Though Lee (1998) asserts that a 5-point-scale-scoring is more effective than a 20-point-scale-one, the researcher adopted a 20-point-scale-scoring system hoping to get more accurate figures. The scoring system used in the study is 100 points in total: 20 points each in pronunciation, audibility, fluency / response time/detail/ content, and comprehension of content.

The first evaluation was made by two raters (R2 & R3) by using a discrete-point scoring system. In this study R2 rated three criteria; pronunciation, audibility, and fluency/response time; R3 rated two

different criteria, they were detail/content, comprehension of content. The total score of R2 and R3 was used as the first evaluation score of the individual examinee. The first evaluation process was videotaped and was used for the second and the third evaluation.

The second evaluation was done by five raters (R4, R5, R6, R7, R8) watching the recorded videotape in a quiet room. The third evaluation was done after the raters checked the second evaluation results in order to reduce the examinees' score discrepancies between the raters. The raters conducted the third evaluation job in separate rooms. Also, the third evaluation was made three weeks after the second evaluation lest the raters should remember the examinees' scores.

3.5. Data Analysis: To answer the research question concerning the relationships between the raters' scores, a paired-sample t-test was calculated. Pearson-product moment correlations were applied to find out the correlations between the raters' scores of the second and the third tests. All correlations were significant at $<.05$, unless otherwise indicated. Questionnaires and in-person interviews were used for the English native raters' scoring tendencies.

4. Results

4.1. Correlations between the rater's scores

For the correlations between the raters' scores of the second and the third evaluations, paired-sample t-tests were conducted. Among the means (R4: -19.19, R5: 7.28, R6: -2.75, R7: 17.78, R8: -23.88), standard deviations (R4: 10.86, R5: 6.56, R6: 17.16, R7: 16.47, R8: 11.61), and standard error means (R4: 1.92, R5: 1.16, R6: 3.03, R7: 2.91, R8: 2.05), only R5's and R7's had significant positive correlations with the examinees' scores while the others' were negatively correlated each other.

A comparison of the raters' scores between the second and the third evaluations produced t-values of R4: -9.999, R5: 6.278, R6: -.907, R7:

6.106, R8: -11.634, which were significantly beyond the .05 level. All of those relationships except R6 (p=.372>.05) were significant at the .05 level as shown on the Table 2, by which it can be concluded that the discussion to reduce the score discrepancies after the second evaluation was effective.

4.2. Correlations between the inter-raters' scores

Table 1. Means and Standard Deviations by the Raters (n-32)

	1st.	2nd Evaluation					3rd Evaluation				
Rater	R2/3	R4	R5	R6	R7	R8	R44	R55	R66	R77	R88
Mean	68.59	59.06	82.28	73.28	80.28	63.28	78.25	75.00	76.03	62.50	87.16
Std. Dev.	9.37	13.76	6.42	12.84	10.94	12.55	5.10	8.18	14.81	20.67	5.31

The figures of the first evaluation above are based on the scores combining R2's (pronunciation, audibility, fluency/response time) and R3's (detail/content, comprehension of content). For easier distinction, the figures for the second evaluation are denoted by the raters such as R4, R5, R6, R7, R8; third evaluation ones like R44, R55, R66, R77, R88. The means and standard deviations were vary greatly with the raters as shown on the Table 1 above. The researcher assumed that the various figures with the mean and standard deviation resulted from the raters' inconsistency when they measured the examinees in accordance with the evaluation criteria. However the evaluation criteria are classified in detail, it is by using his or her own subjectivity that a rater judges the examinees.

Therefore, it is natural that the scores vary from rater to rater. Though the mean and the standard deviation of the scores were varied, the correlations of the score between all the raters were very significant except R6's. If the statistical values are significant, we usually take those figures for granted as reliable. However, the correlations of the individual examinee's rank orders between the raters weren't very significant statistically, by which it can be assumed that the examinees' rank orders have no consistency. It is natural to have ranging rank

orders but the range should be acceptable or understandable. The rank orders of the examineees in this study were far beyond the commonsense level and totally unbelievable. The average discrepancies of the rank orders between the top and the bottom was ±22 out of 32.

Table 2. Comparative Rank Orders by Raters (n-32)

	1st.	2nd Evaluation					3rd Evaluation					Avg.	Low-est	High-est	Rank Discr-epancy
Rater	R2/3	R4	R5	R6	R7	R8	R 44	R 55	R 66	R 77	R 88	Rank	Rank (a)	Rank (b)	(a-b)
examine e1	3	5	3	8	12	19	1	11	4	17	19	**9**	**19**	**1**	±18
T2	23	6	31	31	23	28	24	32	23	13	20	**23**	**32**	**6**	±26
T3	2	10	14	14	1	6	17	16	21	4	1	**10**	**21**	**1**	±20
T4	22	30	4	11	2	29	31	25	15	9	28	**19**	**31**	**2**	±29
T5	5	11	6	10	10	14	22	12	18	14	4	**11**	**22**	**4**	±18
T6	29	22	15	22	11	25	18	26	19	6	12	**19**	**29**	**6**	±23
T7	12	15	2	12	13	2	25	10	20	18	5	**12**	**25**	**2**	±23
T8	24	28	24	19	20	15	28	27	29	10	13	**22**	**29**	**10**	±19
T9	31	23	20	20	24	31	30	28	22	29	27	**26**	**31**	**20**	±11
T10	25	16	16	28	5	20	19	31	32	7	29	**21**	**32**	**5**	±27
T11	13	7	18	29	6	7	3	9	1	31	6	**12**	**31**	**1**	±30
T12	30	31	30	1	25	32	32	29	31	30	7	**25**	**32**	**1**	±31
T13	32	32	32	30	30	30	12	30	26	32	21	**28**	**32**	**12**	±20
T14	20	24	22	32	14	26	20	19	24	8	32	**22**	**32**	**8**	±24
T15	18	17	11	23	7	21	23	20	30	19	8	**18**	**30**	**7**	±23
T16	15	18	26	24	3	22	13	13	11	12	22	**16**	**26**	**3**	±23
T17	4	2	7	13	8	8	6	2	5	11	2	**6**	**13**	**2**	±11
T18	11	25	8	6	21	23	26	6	27	24	23	**18**	**27**	**6**	±21
T19	16	3	9	17	26	24	8	5	8	15	14	**13**	**26**	**3**	±23
T20	14	19	19	18	15	9	21	17	28	20	9	**17**	**28**	**9**	±19
T21	8	4	5	26	9	16	7	3	9	5	24	**11**	**26**	**3**	±23
T22	6	8	21	2	16	27	4	4	6	2	15	**10**	**27**	**2**	±25
T23	9	20	12	4	27	3	16	22	16	23	16	**15**	**27**	**3**	±24
T24	26	12	23	27	22	10	9	21	10	21	17	**18**	**27**	**9**	±18
T25	21	29	25	9	28	11	27	18	3	26	10	**19**	**29**	**3**	±26
T26	27	26	29	3	17	17	29	14	14	25	18	**20**	**29**	**3**	±26

T27	28	27	28	15	29	18	14	15	7	27	30	**22**	**30**	**7**	±23
T28	17	21	17	16	31	12	15	8	12	28	11	**17**	**31**	**8**	±23
T29	19	13	13	25	32	13	10	24	17	22	25	**19**	**32**	**10**	±22
T30	10	14	27	21	18	4	11	23	25	3	31	**17**	**31**	**3**	±28
T31	7	9	10	7	19	5	2	7	13	16	26	**11**	**26**	**2**	±24
T32	1	1	1	5	4	1	5	1	2	1	3	**2**	**5**	**1**	±4

The comparative Table 2 shows the individual examinees' rank orders by the raters. The narrowest case of the rank order discrepancy was examinee 32 (T32), which is ±4. Ten examinees (T1, T3, T5, T8, T9, T13, T17, T18, T20, & T24) show relatively reasonable discrepancies, but the rank order discrepancies of the other twenty-one examinees were far beyond the average(±22), which shows that the raters were inconsistent in their rankings. The worst case was found with T12. R6 rated T12 as the top rank, but R4 and R88 rated the same examinee as the lowest one, rank 32, which is totally opposite. Therefore, it is important for novice raters to recognize that an examinee's rank order will vary widely due to rating inconsistency.

4.3. Tendencies of Untrained English Native Raters

The raters generally seemed not to trust their scores for the examinees, which was proved in a questionnaire after the evaluation was over. The raters' distrusting their scores means that they saw their oversights in the process of evaluation. A rater's scoring differences are caused by strictness or generosity, central tendency, restriction of range, and the halo effect (Englehard, 1994; Saal, Downey, & Lahey, 1980).

Raters' errors caused by strictness or generosity occur when raters measure examinees indiscriminately high or low. R4's and R8's average scores of the second evaluations (59.06 / 78.25) were higher than those of the third evaluations (63.28 / 87.87). This seems to be a result of the raters' generosity. The differences of R4 and R8's mean averages between the second and the third evaluations were -19.19 and -23.87,

standard deviations were 8.9 and 8.0, and standard errors were 1.58 and 1.42 respectively. T-values were -12.177 and -16.867, which were very significant at the probability level of p=.000<.05 (two-tailed). This result was enough to draw the conclusion that the raters' discussion after the second evaluation was of use in narrowing their score discrepancies between the second and the third evaluations.

Most of the examinees' rank orders had inconsistencies, which seemed to be caused by the raters' lack of consistency in applying subjectivities. The worst sample cases are shown in Table 3. R4, for example, gave T2 a 75-point score (the 6th in rank) in the second evaluation, but T2 got 74-points (the 24th) in the third evaluation.

Table 3. Comparative Scores and Ranks Differences by the Raters

Rater	Examinee	Score		Rank-order	
		2nd. eval.	3rd. eval.	2nd. eval.	3rd. eval.
R4	2	75	74	6	24
R5	4	89	68	4	25
R6	12	96	50	1	31
R7	11	90	30	6	31
R8	30	75	80	4	31

As mentioned above, an examinee's ability can be differently scored and ranked even though it is evaluated by watching a videotape in accordance with the required same evaluation criteria. The raters in this study have not specially trained for verbal evaluation. Consequently, it is not surprising that they committed mistakes in evaluation. The result of this study confirms that it is a mistake to classify a person as a professional rater only because he or she learned English abroad and is an English teacher (Kim, 1999).

All the raters in this study showed tendencies of avoiding the maximum or minimum scores. The raters were also influenced by the halo effect. A couple of raters commented that they were more relaxed in the third evaluation, which they did separately. R6 and R7 admitted that evaluating five criteria simultaneously was not easy. If a rater failed to evaluate an examinee meaningfully in accordance with the

evaluation criteria, it was due to his failure in terms of evaluating the examinee's potential communicative language skills.

The researcher had asked the raters to evaluate the examinees using a discrete-point evaluation method in accordance with the evaluation criteria, the raters, however, unwittingly used the integrative evaluation method. This was the result of the raters' being more influenced by the impression an examinee was making than by the criteria they should have been continually aware of, as Cooper (1983) asserts.

5. Conclusion and Recommendations

In regard to the first research question of the raters' score correlations, the research showed positive correlations between the scores of R5 and R7 in the second and the third evaluations, while those of R4, R6 and R8 were negatively correlated. The t-test results were very significant statistically except R6's, which can be interpreted as the raters having tried to keep consistent subjectivities for narrowing the examinees' score discrepancies.

With regard to the second question, the correlations between the inter-raters varied with the raters in their means and standard deviations between the second and the third evaluations, which was assumed to be caused by the inconsistent subjectivities. The correlations between the raters' scores, however, were very significant statistically. Notwithstanding the statistical significances above, the examinees' rank orders were not significant, and the rank order showed big differences between the raters, and the average discrepancies were ±22 out of 32. The worst case was found with T12, who was ranked as the top by R6 while R4 and R88 graded him as the poorest one, the result obviously of the raters' incoherent subjectivities in the process of the evaluation.

The third question was about the raters' scoring tendencies. The raters negatively answered the question, "Do you think your evaluation result would be credible and unbiased if you were to get the test score for yourself as an examinee?" If the raters themselves did not trust the

test scores they gave, it means that they knew they were committing mistakes in the process of evaluation. The raters' scoring differences mostly resulted from their strictness to limit the examinees' scores or generosity to increase the examinees' ones.

To summarize the results of this study, the researcher draws the conclusion that an examinee's rank order can significantly vary from unprofessional rater to unprofessional rater unless they are professional raters even though the statistical value is very significant. This result verifies Shin's assertion that we can't rely on the traditional reliability assessment method entirely for English speaking test results (Shin, 2001b).

A rater's subjectivity is changeable in accordance with the time and circumstances, and keeping a consistent subjectivity is pretty difficult. Accordingly, rater training is essential for securing reliability with subjective tests (Lee, 1995), and it is suggested that a rater be trained so as to be able to judge examinees using coherent subjectivity. Applying rank orders in subjective tests can cause dangerous results. It is, therefore, desirable to avoid rank order tests for evaluating communicative language skills. Instead, oral tests should be used only for measuring the levels of the examinees' speaking ability with classifications such as 'excellent,' 'good,' 'average,' 'poor,' etc.

More than two raters are recommended for oral tests, and if any big discrepancy is observed, an average score or an adjusted score should be used for the examinee's final score. If no other rater is available, the oral test process should be videotaped or audio-taped for re-evaluation, which can improve the inter-liabilities of the examinee and the rater. The questions should be presented with an audio or video taped voice in oral tests to lessen the examinee's psychological burden. Also, various types of questions like ones required for simple answers should be included in oral tests in order to improve the test validity with more discrimination.

Last but not least, subjective tests should not be used for examinees' rank orders even though the results are statistically very significant. A rater should expect examinees' potential requisitions for releasing proven

relevant scoring data to confirm their test scores. A rater who is ready to stand by his scoring data is a professional rater. If any unhappy things happen as a result of an oral test, the rater should take the blame. Therefore, a good English teacher must be competent in teaching methods, but no less important, he must be well trained so as to keep consistent subjectivity when evaluating.

References

Koog, M. N. (1998). A development of rating scales for language (English). *Primary English Education, 4(1)*, 189-221.

Kim, M. S. (1999). A study of marking behaviors in performance assessment of English writing. *Journal of Educational Evaluation, 12(2)*, 25-54.

Kim, Y. S. (1999). A study on English oral performance tests methods and measurement reliability. *Journal of the Applied Linguistics Association of Korea, 15(1)*, 171-198.

Shin, D. I. (2001a). Exploring rating patterns with Rasch measurement techniques: implications for training. *Foreign Language Education, 8(1)*, 249-272.

Shin, D. I. (2001b). Validation process of an EFL speaking test on G-theory and other analytical techniques. *Journal of the Applied Linguistics Association of Korea, 17(1)*, 199-221.

Shin, D. I. & Jang, S. Y. (2002). Understanding rating error sources on halo effect. *Foreign Language Education, 9(4), 215-232.*

Shin, D. I. (2003). *English Language Testing in Korea.* Seoul: Hankookmunhwasa.

Lee, Y. S. (1998). An investigation into Korean markers' reliability for English writing assessment, *English Teaching, 53(1)*, 179-200.

Lee, Y. K. (1995). Assessing Korean university students' spoken English proficiency, *English Teaching, 50(1)*, 37-63.

Cho, M. W. & Lee, H. S. (1991). *A Dictionary of English Language Teaching.* Seoul: Hanshinmunhwasa.

Buck, G. (1990). *Testing second language listening comprehension.* Unpublished doctoral dissertation, University of Lancaster.

Carroll, J. B. & Freedle, R. O. (1972). *Language Comprehension and the Acquisition of Knowledge.* Washington, D. C.: V.H. Winston & Sons.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research on the Teaching of English*, 18(1), 65-81.

Chastin, K. (1971). *The Development of Modern Language Skills; from Theory to Practice.* Philadelphia: The Center for Curriculum Development.

Cooper, W. H. (1983). Internal homogeneity, descriptiveness, and halo: Resurrecting some answers and questions about the structure of job performance rating categories. *Personnel Psychology, 36*, 489-502.

Curran, Charles A. C. (1976). *Counseling-Learning in Second Language.* Apple River, Ill.: Apple River.

Engelhard, G. J. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31(2)*, 93-112.

Flowerdew, J. (Ed.) (1994). *Academic listening research perspective.* Cambridge: Cambridge University Press.

Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening* (pp.241-268). Cambridge: Cambridge University Press.

Harris, D. (1969). *Testing English as a Second Language*, New York: McGraw-Hill Book.

Hughes, A. (1989). *Testing for Language Teachers.* Cambridge: Cambridge Universit Press.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach.* Rowley, Mass: Newbury House.

Johnson, D. M. (1992). *Perceiving Writing quality through gender lenses.* Paper presented at the 26th Annual TESOL Convention,

Vancouver, BC.

Millanovic, M., Saville, N., & Shuhong, S. (1993). A study of the decision-making behavior of composition markers. Paper presented at the 15the Language Testing Research Colloquium, Cambridge UK. August 2-4.

Morrow, K. (1977). *Techniques of evaluation for a notional syllabus*. Cambridge: The Royal Society of Arts.

Morrow, K. (1981). Principles of communicative methodology, In K. Johnson & K. Morrow.(Eds.) *Communication in the classroom*. Harlow: Longman.

Oller, J. W. (1976). Evidence of a general language proficiency factor: an expectancy grammar. *Die Neuren Sprachen 76*:165-174.

Pino, B. G. (1989). Prochievement testing of speaking. *Foreign Language Annals, 22(5)*. 488

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88(2)*, 413-428.

Spolsky, B. (1968). 'Language testing: the problem of validation.' *TESOL Quarterly 2*, 2:88-94.

Stansfield, C. W., & Kenyon, D. (1992). Comparing scales of speaking tasks by language teachers and by the ACTFL guidelines. In A. Cumming & R. Berwick, (EDS.), *Validation in language testing* (pp.124-153). Clevedon, Avon: Multilingual Matters.

Stubbs, J. & Tucker, G. R. (1974). The cloze test as a measure of English proficiency. *Modern Language Journal 58*, 239-248.

Terrell, T.D. (1977). A Natural Approach to Second Language Acquisition and Learning. *Modern Language Journal, 61(7)*, 325-337.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge University Press.

Weir, C. J. (1990). *Communicative Language testing*. Hemel Hemstead: Prentice Hall

Wiggins, G. (1990). The Case for Authentic Assessment. *ERIC Digest*, ED328611.

(Appendix 1)

Criteria for conversation test(interview)

Pronunciation (20 points)
1(0): No answer given / 2(5): Poor / 3(10):Average / 4(15):Good / 5(20):Excellent
Audibility (20 points)
1(0): No answer given / 2(5): Difficult to hear
3(10):Could hear speaker with occasional difficulties in hearing answers
4(15):Easy to hear speaker / 5(20):Very audible
Fluency/Response Time (20 points)
1(0): No answer given
2(5): Not confident with speaking English. / Level of vocabulary and grammar poor / Poor interviewee waits long period of time before attempting to answer questions
3(10):Confident answers with errors in vocabulary and grammar Average Pauses before answering questions
4(15):Confident answers with good use of grammar and vocabulary, with few errors. / Good Minimal pause before answering questions
5(20):Very confident and fluent speaker. Excellent use of grammar and vocabulary / Excellent Answers given directly after questions asked
Detail/Content (20 points)
1(0): No answer given
2(5): Poor Incomplete sentences used, one word answers, little to no detail
3(10):Average Answered with little detail but used complete sentences, does not give one word answers
4(15):Good Uses some details to give answer, and uses complete sentences
5(20):Excellent / Excellent answer, uses many details to answer questions, using only complete sentences
Comprehension of Context (20 points)
1(0): No answer given
2(5): Poor Difficult to comprehend interviewee's response
3(10):Average Comprehendible answers and questions with some errors
4(15):Good Comprehendible answers and questions with few errors
5(20):Excellent Answers and questions easily understood

Kim, Ilhong
Language Education Center
Seojeong College
681-1 Yongam-ri, Eunhyeon-myeon,
Yangju-si, Gyeonggi-do 482-777, Korea
Phone: 82-31-860-5034
Email: kih@seojeong.ac.kr

Received: 12 Jun, 04
Revised: 29 Jul, 04
Accepted: 23 Aug, 04