The application of G-theory and IRT in the analysis of data from speaking tests administered in a classroom context

Tomoyasu Akiyama
The University of Melbourne

Abstract

This study[1] explores the potential of applying two measurement approaches (Generalizability theory and Item Response Theory) in the analysis of data from speaking tests for Japanese junior high school students. To date, few studies have been carried out on speaking tests at the school level, particularly for junior high schools in Japan. The present study focuses on the school context and an analysis of the characteristics of items (tasks) and raters with a view to establishing the optimum number of tasks and raters in the particular context. The data used in this study were gathered from a test administered to 109 junior high school students in Tokyo. The test consisted of three tasks and 11 items, and was conducted by five interlocutors. Four raters independently rated the performance of the 109 students. The study shows the potential of using the two measurement approaches as useful tools to examine teachers' self-made tests.

1. Introduction and review of literature

Given the great demand for assessing performance tests in the last two decades, Bachman (2000) states that two measurement approaches have been applied to research of language testing: Generalizability theory and Item Response Theory. Using the two

[1] This paper is a revised and expanded version of Akiyama (2001) shown in ELEC Bulletin.

Melbourne Papers in Language Testing 2001 Volume 10.1 pp. 1-21. The Language Testing Research Centre, The University of Melbourne.

measurement approaches, researchers have investigated facets of the testing context, including the optimum number of raters and tasks (McNamara and Lynch, 1997; Lynch and McNamara, 1998), item analysis (McNamara, 1996), and rater behavior (Wigglesworth, 1993; Lumley and McNamara, 1995; Weigle, 1998). In the following sections, studies to which two measurement approaches are applied are presented briefly.

Generalizability Theory (G-theory)

Generalizability theory is an extension of the framework of classical test theory. Classical test theory treats error as undifferentiated and random, so that it cannot identify sources of error (Bachman, 1990). In contrast, G-theory yields estimates of the relative effects of each source of error. In G-theory, the object of measurement is a person's ability and all other components or facets, such as raters, items and tasks and all possible combinations of them, are considered as sources of error. Bachman (2000) points out that G-theory analysis enables one to distinguish between sources of measurement error and the persons, by estimating the variance component (G-study). If much of the variance components in test scores is attributed to persons, test scores can be interpreted as a reflection of a person's ability. Thus G-theory enables researchers to identify the influence of facets other than the object of measurement.

A Decision study

Based on the variance component of each facet estimated by G-study, the Decision study (D-study) shows the potential reliability (dependability) according to scenarios: how many items and raters are needed to ensure targeted dependability. D-study provides two coefficients based on potential combinations of all facets: G-coefficient and phi coefficient. The former refers to traditional reliability, which concerns relative ranking order. The latter refers to criterion-referenced dependability, which concerns absolute decisions such as mastery/non-mastery. The two coefficients provide useful information in a test situation. For example, test developers, who design speaking tests involving raters and various tasks, need to decide on the test conditions: how many raters and tasks will be needed. In order to reach targeted dependability in terms of cost and time effectiveness in speaking tests, test developers and administrators need to how many items and raters are needed.

McNamara and Lynch (1997) conducted a series of D-study in performance tests, called the *access*. The purpose of *access* is to screen 'immigrants' who come from non-speaking English counties. One part of this test requires candidates to demonstrate proficient oral communication skills. As the assessment will determine whether the prospective immigrant can come into the country, this context requires an absolute decision (pass or fail), which is associated with phi coefficient. Based on *phi* coefficient indexes, the D-study shows potential dependability on the number of items and raters. McNamara and Lynch (1997) suggested that two raters would be required to reach acceptable reliability, however, a third rater would be needed if there is a major discrepancy between the first two raters. This provides practical and useful information not only to the researchers but also to test developers and administrators.

Item Response theory (IRT)

Item Response Theory is a mathematical measurement theory. This allows us to identify differences between actual data patterns in response to items and patterns estimated by IRT. If these differences are outside the acceptable range based on all interactions between item and test-takers' response, IRT shows that there are problems with items and test-takers. Henning (1987) points out that one of the strengths of IRT is that IRT allows both persons' ability and item difficulty to be estimated using a common scale 'logit'. This unit enables all facets to be compared to each other on the same scale. Rasch measurement is a more sophisticated measurement technique than classical measurement theory, in that IRT considers each item difficulty and individual ability.

A family of IRT models forms the basis for what is known as Rasch measurement. The basic Rasch model investigates the relationship between an item difficulty and a person's ability. Over the last two decades, Rasch models have been further extended to include the relationship between test-takers' ability, item difficulty, and rater severity, and other facets of the assessment setting, called many-facet Rasch measurement (McNamara, 1996).

Investigation of construct validity

One of the greatest contributions to language testing is the investigation of construct of validity in using IRT, including Rasch

measurement. Investigation of construct validity, based on item analyses, is concerned with to what extent items correctly measure what items purport to measure. According to McNamara (1990b: 109), 'Rasch analysis does identify a number of items as not contributing to the definition of a single measurement dimension'. McNamara (1990a) investigated the construct of the speaking test of the Occupational English Test for health professionals, using 6 items (*Overall communicative effectiveness, Intelligibility, Fluency, Comprehension, Appropriateness, Resources of grammar*). The item analysis provided a test developer with information regarding the 'performance of items' as to whether or not items make a contribution to measuring the targeted construct. The identification of 'misfitting items' with Rasch measurement shows which items do not perform well in measuring the targeted construct. McNamara (1990a) found that there were no misftting items, whereas the two items were 'overfitting', indicating that the items did not make 'independent contribution' to measuring the construct. This signals that an overfitting item overlaps other items. His research showed that the majority of items were helpful in making a contribution to the construct being assessed. This study also showed the possibility of utilizing the IRT as a tool to explore construct validity. Thus, IRT can be applied to assist teachers in examining their self-made test as to whether or not the items are correctly measuring the targeted ability.

Investigation of rater behavior

Lumley and McNamara (1995) explored rater severity and rater consistency using a Rasch model. Speaking tests often requires subjective judgements by raters. In investigating rater consistency, the advantages of using IRT, according to McNamara (1996), is that Rasch measurement can indicate who is a consistent or inconsistent rater. As misfitting item above, inconsistent raters are identified as 'misfitting raters'. The raters identified then should undergo for the rater training or excluded from conducting tests.

Rasch measurement also shows exactly how severe one rater is compared to another rater. Lumley and McNamara (1995) showed that rater training improved intra-rater reliability overall. However, in terms of the relationship between individual raters and each item, they found that there was variability between them. In similar vein, Weigle's (1998) study showed that raters were not equally harsh or lenient, and that rater variability can be narrowed, regardless of

rating experience. Regardless of rater variability, another advantage of using IRT that it can adjust the differences of harshness between raters when raters are within the acceptable range of consistency. Thus, the important thing is how consistent raters are rather than how much agreement there is between them.

Recent research using these two measurement approaches in language testing has primarily investigated test validity and reliability for adults, including immigrants in a second language context. Few studies have been carried out o speaking tests at the school level, particularly for junior high schools in Japan. Japanese junior high school students (equivalent to Years 7 to 9 students in Australia) undertake entrance examinations for senior high school (Years 10 to 12) when they are in Year 9. Admission decisions for senior high schools are made based on both teacher implemented assessment (school-based assessment) and entrance examinations (external standardized tests). The subject 'English' require both their assessment. Thus, teacher-generated (school-based assessment) grades for Year 9 students have a significant impact on students and their parents because school-based assessment represents 50 % of the basis for admission decisions for senior high school students. Nevertheless, little research into school-based assessment, in particular, of speaking skills has been conducted in Japanese junior high schools. Teacher-implemented assessment involves high-stakes decisions. It is important to investigate whether speaking tests developed by English teachers can deliver interpretable scores. Therefore, this study explores the potential of applying two measurement approaches (Generalizability theory and Item Response Theory) to the analysis of data from speaking tests for Japanese junior high school students where more than 35 students are in the classroom.

This study addresses mainly three questions:

1) What is the optimum number of raters and tasks (items) needed to reach relatively high dependability in the current classroom condition?
2) To what extent do items developed by Japanese English teachers assess students' speaking skills?
3) To what extent are teachers as raters consistent?

2. Research methods

Participants

109 Japanese students, who had studied English as a foreign language for two or three years, mainly in the classroom setting, participated in this study. The students were 14 to 15 years old and were from three different public schools in Tokyo. The students usually have only three or four English classes (50 minutes per class) per week and as such their English ability can be assumed to be limited. Table 1 shows the school, interlocutors' identification, and number of the candidates, including their gender and grades (8^{th} and 9^{th}). The interlocutors were (A, C, D, and E) four Japanese English teachers who taught the participants at their schools and the researcher (B).

Table 1: A summary of research participants

School name	Interlocutor ID	Number of students	Year of the students
1	A and B	34 (*M19.** F15)	9^{th}
2	C and D	40 (M29, F11)	8^{th} (20) and 9^{th}(20)
3	E and B	35 (M15, F19)	8^{th}

*M = Male **F = Female.

Test structure

The length of the test was approximately 10 minutes per student. The speaking test consisted of three assessed sections following section 1, which was an unassessed warm up. The test consisted of four sections as follows: (See Appendix 1: A summary of the test task).

Section 1 (30 seconds). Warm up: The purpose of this section was to get the student to relax and to understand the test procedure. This section was unassessed.

Section 2 (3 minutes). This section was divided into two parts. In the first part, the student was required to answer two questions about illustration. The second part required the student to describe illustration in English within a given time limit.

Example: (Look at the illustration 1). This is Akiko's family in the living room. Could you describe the illustration in English? You have 15 seconds to think about it. After that, you have 1-minute description time.

Section 3 (3 minutes). Situational responses: In this section, students were required to respond orally in English to oral Japanese prompts. This task required the student to produce four language functions such as asking for information, asking for permission, and making excuses, in accordance with different prompts.

Example: (Japanese prompt from the interlocutor). What do you say to your teacher in English when you want to be absent from class? (translated into English)

Section 4 (3 minutes). Role play: In this section, an interlocutor played the role of a cashier at a fast food restaurant and the student played a role of a customer who wanted to buy a hamburger and orange juice.

Raters and scoring criteria

Four raters participated in this study. The raters had had 10 years teaching experience. Raters were two Japanese English teachers (1 and 2) and the other two (3 and 4) were assistant language teachers (ALTs), native speakers of English, who assist Japanese English teachers and students to improve communicative skills. All four raters independently rated 109 students.

The scoring criteria needed to reflect the language skills developed in class. English classes of the three participant schools generally focus on mastering basic language knowledge and use, so grammar and vocabulary were chosen. A nonverbal (eye contact, facial expressions and gestures) and intelligibility criteria were chosen as global criteria. Performance on each item was rated on a scale from 0 to 4. Different levels of performance were carefully described for each item. In order to measure oral performance, each item was defined in accordance with the degree of students' performance. For example, a '0' score on the criterion of 'Vocabulary' indicated 'no response or irrelevant response' and a score of '4' meant 'uses vocabulary precisely and appropriately' (See Appendix 2: scoring criteria). The global criteria such as intelligibility and nonverbal skills were marked on the degree to which the student performed not in each section, but on the whole

test. All students were videotaped in order that intelligibility and nonverbal skills could be rated later.

Analysis methods

In order to answer question 1, the analyses were carried out using GENOVA (Crick and Brennan, 1984) software, which is the application of G-theory. The Decision study (D-study) gives two indicators regarding dependability based on potential combinations of all facets: G-coefficient and *phi* coefficient. ConQuest (Wu, Adams and Wilson, 1998) program, which is the application of a generalised IRT, was used in order to answer research questions 2 and 3. ConQuest provides specific information of items based on the degree of fit to the IRT model and information on rater consistency and rater harshness.

3. Results

Results with G-study

The G-study design used in this study is a 'random effects' model with two facets: 4 raters and 11 items. The term 'random effects' refers to the assumption that 4 raters and 11 items interacted interchangeably. The focus here is dependability of test scores using full facets (4 raters and 11 items). First, this study examines the relative effects of the variance components of persons (students), raters, items and the combination of them.

Table 2 shows the variance component of each facet. Using the variance components, 87% of the total variance is attributable to persons, which is acceptably high variance. In other words, person ability accounts for 87% in this context. However, rater-related variance amounts to approximately 11%, indicating that raters vary more or less according to students' scores. The variance component of items is only approximately 2%, which did not influence variability in this assessment.

Table 2: D-study variance components (4 raters x 11 items)

Effect	Variance component	Standard errors	% of total variance
Persons (P)	0.718	0.101	87.0
Raters (R)	0.062	0.040	7.6
Items (I)	0.002	0.001	0.3
PR	0.026	0.002	3.2
PI	0.010	0.000	1.2
RI	0.001	0.000	0.2
PRI	0.005	0.000	0.6
Total	0.819		

Table 3 presents the G-coefficient and phi-coefficient in the cases of different raters and items. It is widely accepted that the G-coefficient is larger than the phi-coefficient. This suggests that test candidates are ranked very similarly among raters, whereas there is some disagreement to test scores (Lynch and McNamara, 1998).

Table 3: Dependability estimates for different numbers of raters and items

Number of raters	Number of items	G-coefficient	Phi-coefficient
1	11	0.84	0.65
1	9	0.84	0.64
1	6	0.82	0.63
1	3	0.77	0.59
2	11	0.91	0.78
2	9	0.90	0.78
2	6	0.89	0.76
2	3	0.85	0.73
3	11	0.93	0.83
3	9	0.93	0.83
3	6	0.92	0.82
3	3	0.88	0.79

Most interestingly, there is little difference between use of 11 items and 9 items regardless of the number of raters in terms of the two coefficients. This indicates that a plausible exclusion of the two items

does not decrease the quality of dependability. According to the two coefficients, discrepancies arise with a number of raters scoring items. A large discrepancy occurs when one rater only used than when two or more raters are used. Table 3 shows that exclusion of two items makes little difference in terms of two coefficients.

D-study for different scenarios with two coefficients

The focus of this study was time efficiency: how many raters and tasks could be reduced. The scenarios would be that two teachers (a JET and an ALT) administer the test for 35 students and rate them independently. Each of three test tasks was approximately three minutes. In this context, therefore, reduction of one task, from three to two tasks could save approximately 105 minutes (35 students x 3 minutes). The table 4 includes six different scenarios, estimating two coefficients. For example, scenario 1 (two tasks and one rater) is 0.78 (G-coefficient) and 0.64(phi-coefficient). As can be seen, at least two tasks and two raters would be necessary to achieve relatively high coefficients (0.85 G-coefficients and 0.75 phi-coefficient). Table 4 shows that two raters would be necessary to obtain more than 0.75 regardless of number of tasks.

Table 4: D-study for different scenarios with two coefficients

No. of Scenario	No. of tasks	No. of raters	G-coefficient	Phi-coefficient	Time required per student / minutes
1	2	1	0.78	0.64	6 *(105 minutes)
2	2	2	0.85	0.75	6 (105 minutes)
3	2	3	0.87	0.79	6 (105 minutes)
4	3	1	0.81	0.66	9 (158 minutes)
5	3	2	0.88	0.77	9 (158 minutes)
6	3	3	0.90	0.82	9 (158 minutes)

*Parenthesis is time required for 35 students for one of two teachers to administer the test.

Item analysis

The focus is on how each item contributes to constructing speaking ability (Wright and Masters, 1982:93). In other words, the question is whether or not the majority of items correctly measure speaking ability. Table 5 shows the analysis of all items used in the speaking test. The first and the second columns show item numbers and the name of 11 items, including the relevant test section. The third column indicates the difficulty of each item. The disparity between the easiest item 1 (-0.46 logit) and the most difficult item 6 (0.02) was roughly 0.5 logits. The range was very small which suggests that most items were equally difficult. The fourth column shows the error which indicates accuracy of estimation in terms of each item difficulty. The error is usually expected to be less than 0.2. The fifth (T) column is more popularly called, *Fit* (T) value (Wright and Master, 1982: 99). Basically, T indices show the extent to which expected scores obtained from the Item Response model differ from the actual data. In terms of Fit (T) index, the acceptable range is $-2 < T < 2$. The index less than –2 is called *overfit* and the index larger than + 2 is called *misfit*. The overfit item is based on the evidence that actual data shows less variability than expected scores. It indicates that the overfit item shows 'unexpected interdependence' (Wright and Master, 1982: 96). Thus, an *overfit* item does not make an independent contribution to constructing separated ability to be measured. On the contrary, a misfit item occurs when the difference between both scores is so large that the Rasch measurement cannot confirm the responses due to inconsistency. The misfit item signals that an item does not fit the model due to lack of consistency. In practice, the misfit item does not discriminate between low and high ability students as consistently as other items do.

Table 5: Item analysis with Rasch measurement

Item	Difficulty[2]	Error	T
1 Fluency *S2	-0.46	0.03	***2.1
2 Grammar S2	-0.24	0.03	-1.0
3 Vocabulary S2	-0.41	0.03	1.4
4 Appropriacy S3	-0.11	0.03	1.4
5 Fluency S3	-0.08	0.03	-0.5
6 Grammar S3	0.02	0.03	-1.3
7 Appropriacy S4	-0.31	0.03	-0.2
8 Fluency S4	-0.27	0.03	-0.2
9 Grammar S4	-0.15	0.03	-1.3
10 Intelligibility (**G)	-0.22	0.03	****-4.1
11 Nonverbal (G)	-0.07	0.04	-1.6

Separation Reliability =0.939.
*S= section, **G= Global criteria, *** misfit item, **** overfit item

Item 1 was slightly larger than 2 (misfit item), which was on the borderline of an acceptable range. This means that the item assessed student performances inconsistently compared with other items. On the other hand, item 10 (intelligibility) was an overfit item. This indicated that that this item was dependent on other items. It could be often interpreted that this item was a redundant item which did not make an independent contribution to the construct. All items but items 1 and 10 were within the acceptable range. These results suggest that most of the items made an independent contribution to the construct of speaking ability in this context.

Rater severity and consistency

Table 6 provides a summary of rater severity and consistency obtained from ConQuest. Raters are identified in the first column, and the second column indicates the severity of raters. For instance, Rater 3 with the highest positive sign (0.446) was the most severe. On the contrary, the most lenient rater was Rater 4 with the highest largest negative figure (-0.265). The discrepancy between Raters 3 and 4 was

[2] Note that the item difficulty in Table 5 sets the mean of the latent ability distribution at zero in order to obtain the Fit (T) value of the last item. (Wu et al, 1998:21), although it is usual to set the mean of the item difficulty of parameters to zero in Rasch measurement.

approximately 0.7 logits. In the last column, T statistics indicates how consistently each rater marked the students' performances. Like item analysis, the figure with less than –2 is called overfit and larger than +2 is called misfit.

Table 6: Rater severity and rater consistency

Rater	Severity	Error	T
1 (the researcher)	-0.064	0.02	0.1
2 (Japanese English teacher)	-0.065	0.02	1.8
3 (native speaker of English)	0.446	0.02	-0.8
4 (native speaker of English)	-0.265	0.02	4.6

Separation Reliability = 0.98

In terms of raters, the latter indicates that 'misfit rater' marked inconsistently. As can be seen, all raters except Raters 4 were within acceptable ranges. The value of Rater 4 is far larger than + 2. This indicates that Rater 4 rated with unexpected inconsistency. Finally, separation reliability in IRT, which refers to 'the proportion of the observed sample variance which is not due to the measurement error', was 0.98. (Wright and Masters, 1982:106). This indicates that raters differed significantly in severity overall. It may be possible that rater inconsistency may arise from differing complexities of the topics covered. Content areas might also need to be further examined as a fix facet (the extent to which ratings differ across content areas, whether raters tend to be more reliable and accurate in certain content areas than in others).

Relationship between raters and each item

Raters marked in very various ways in terms of each item. Table 7 shows the estimations of all items estimated for each rater, including means and standard deviations of raters and items. The first column indicates the item name, including section and item number. Each item has item difficulty estimated by each rater obtained from ConQuest.

Table 7: Item difficulty estimated by each rater (logit score)

Section Item (n)	Rater 1	Rater 2	Rater 3	Rater 4	Mean	S.D.
*2F(1)	-0.86	-0.84	-1.6	-0.14	-0.86	0.6
2G(2)	-0.11	0.0	-0.55	-0.2	-0.12	0.24
2V(3)	-0.64	-1.16	-0.77	-1.5	-1.02	0.36
3A(4)	0.21	0.68	**-1.03**	0.11	-0.01	0.73
3F(5)	0.37	0.71	-0.27	**1.17**	0.5	0.61
3G(6)	0.87	0.89	0.64	0.96	0.84	0.14
4A(7)	-0.46	-0.37	-1.52	-0.53	-0.72	0.54
4F(8)	-0.5	0.26	-0.55	0.11	-0.17	0.42
4G(9)	0.15	0.29	0.45	0.41	0.33	0.14
I (10)	**0.95**	-0.03	-0.41	-0.15	0.09	0.56
N (11)	0.01	-0.43	**5.36**	-0.22	1.18	2.79
Mean	0.00	0.00	0.00	0.01	0.00	0.00
S.D.	0.59	0.65	1.92	0.72		

A= Appropriacy, F= fluency, G= Grammar, Vocabulary, I= Intelligibility, N = Nonverbal.

The Rasch measurement sets zero for the mean of each rater's item difficulty. Bold numbers in Table 4 above are worthy of comment. In particular, Item 11 (Nonverbal: 5.36 by Rater 3) is an extremely high value, which indicates that the rater marked this item particularly severely. Similarly, both Intelligibility (10) by Rater 1 and Fluency (5) by Rater 4 appear to be severe. In contrast, Appropriacy (3) by Rater 3 is relatively lenient item. However, these were not serious problems compared with the case (nonverbal) of Rater 3. In terms of each item difficulty, the standard deviation of the three 'Grammar' items account for the smallest in each section, which shows that these item are more stable than other items among raters.

4. Discussion

The aim of this study was to establish the optimum number of tasks and raters in the particular context and to analyse characteristics of items (tasks) and raters using G-theory and IRT. Results based on G-theory showed that an acceptable amount of variance component was attributed to student ability and that at least two raters and two tasks were needed to reach relatively high dependability of test scores. Using IRT, it was found that all items except one contribute to

measuring students' speaking ability and that three out of the four raters rated consistently overall.

The two measurement approaches adopt both global and specific views and provide useful approach to resolving issues inherent in speaking tests. G-theory provides 'total balance' of all facets of a group such as items, raters and persons in the test. Results informed by G-theory provide useful information for designing tests: what is the optimum number of raters and tasks in the specific context? On the other hand, IRT provided specific individual information such as specific items, tasks and raters. For example, IRT provides who needs to undergo further training and which items need to be revised. Bachman *et al.* (1995: 256) point out the strength of the two when applied to language testing:

> *G-theory identifies the relative effects of facets such as raters or tasks, as well as the relative effects of combination of these facets (interactions). Many-facet Rasch measurements allow us to identify specific raters, specific tasks and specific combination of raters, tasks and persons that are affecting the dependability of our judgments.*

G-theory can estimate to what extent the variance components are attributed to each facet of the tests. Following the results of G-study, D-study can assist test developers in designing the targeted reliability, showing a particular set of conditions for each facet. Then, IRT provides test developers with more specific information, such as individual item difficulty and raters. In attempting to highlight the usefulness of the two theories, Lynch and McNamara (1998: 176) use the analogy of a microscope.

> *Using the microscope as an analogy, FACETS turns up the magnification quite high and reveals every potential blemish on the measurement surface. GENOVA, on the other hand, sets the magnification lower and tends to show us only the net effect of the blemishes at the aggregated level. **There is not to say that 'tuning up the magnification' is the same as increasing accuracy. It merely suggests that there is a different level of focus (individuals versus groups).*** [emphasis added]

It is important to note that the statistical and theoretical limitations of both G-theory and IRT models. For example, a G-study model is relatively simple to analyse and interpret, but most likely does not

exhaustively partition the errors in ratings into their respective sources to gain a full understanding of sources of variance in ratings. A more complex multifaceted model might include raters' confidence in ratings as measurement facets, so that the proportion of variance in ratings attributable to such sources could be estimated. In terms of raters, limitations in human information processing and attentional capacity may cause fatigue and perhaps boredom, thus introducing a high degree of error into the rating obtained from the raters.

Although it is argued that such statistical technique would be useful for the design and analysis of speaking tests, the question is whether it is possible or necessary for junior high school teachers to master such statistical skills. English teachers at the junior high schools in Japan develop, administer speaking tests and rate their students. As stated before, teacher-generated grades represent half of admission decision for senior high schools. This context requires teachers to deliver reliable scores. In addition, the main goal of the guidelines issued by the Japanese Ministry of Education (1999) emphasizes the development of speaking skills. In order to allow teachers' assessment of speaking skills to produce reliable sources, it is clear that much investment is needed to assessment skills. Brindley also (2000: 153) clearly expresses serious concern over teacher-implemented assessment:

> *If they are to be expected to design and conduct assessments which can provide valid and dependable information, teachers need the opportunity to develop the skills necessary to do so. While formal degree courses and professional development activities undoubtedly play an important role here...*

One practical solution, as the study results indicate, is that at least two raters are needed to reach relatively high dependability. This suggests that in administering the speaking test under the current teaching context, an assistant language teacher's (ALT) cooperation could reduce time and alleviate class size problems. However, procedural variations in how rating procedure is conducted need to be considered as caution, since various methods and contexts can influence the psychometric properties of the obtained data. Therefore, it is not clear whether the optimal number of raters indicated by the result of this study will generalise to other procedural variants. It is clear that further research needs to be conducted.

5. Conclusions and implications

This paper argues that the two theories of measurement can provide useful information to analyse the test data from both global and specific point of views. G-theory provided 'total balance' of all facets as a group, such as items, raters and tasks and IRT provided specific individual information of specific items and raters. These measurement approaches can complement each other and employ the strengths of each other, depending upon information needed in specific contexts.

If applying the two measurement approaches using statistical techniques are impractical (or impossible) for English teachers to master, then expert development and analysis of speaking tests would be necessary. Another potential solution to the problem of accurate assessment of speaking skills for Japanese junior high school students in a high-stakes context is to introduce speaking tests in the entrance examination for senior high schools. Such high-stakes tests need to be developed and administered with experts in cooperation with teachers. Thus, it is clear that further research is needed on how the introduction of speaking tests would have impact on teachers and students. If the introduction of speaking tests might have a great impact on teaching methods as well as on the assessment of English, such tests would gradually approach the goal of the guidelines issued by the Japanese Ministry of Education.

References

Akiyama, T. 2001. An analysis of spoken tests for junior high school students using IRT and G-theory. *ELEC (English Language Educational Council) Bulletin*. No 108. 56-62.

Bachman, L. F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L.F. 2000. Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing*, 17(1), 1-42.

Bachman, L.F., Lynch, B. K. and Mason, M. 1995. Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12(2), 238-57.

Crick, J. E. and Brennan, R. L. 1984. *GENOVA: a general purpose analysis of variance system. Version 2.2.* Iowa City, IO: The American college testing program.

Brindley, G. 2000 Task difficulty and task generalisability in competency-based writing assessment. In Brindley, G (ed), *Studies in immigrant English language assessment. Research series 11,* 125-157. Sydney: National Centre for English for Language Teaching and Research, Macquarie University.

Henning, G. 1987. *A guide to language testing: development, evaluation and research.* Heinle & Heinle Publishers.

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

Lynch. B.K. and McNamara, T. F. 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(2), 158-180.

McNamara, T. F. 1990a. Item response theory and the validation of an ESP test for professionals. *Language Testing,* 7(1), 52-76.

McNamara, T. F. 1990b. *Assessing the second language proficiency of health professionals.* Unpublished Ph.D. Thesis. University of Melbourne, Australia.

McNamara, T. F. 1996. *Measuring second language performance.* London: Longman.

McNamara, T. F. and Lynch, B. K. 1997. A generalizability theory study of rating and test design in the writing and speaking modules of the access test. In G. Brindley and G. Wigglesworth (eds). *Access: Iissues in language test design and delivery,* 197-214. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Ministry of Education. 1999. CHUGAKU *GAKUSHUUSHIDO YOURYO*: [The guidelines of teaching foreign language]. Tokyo Shoseki.

Wigglesworth, G. 1993. Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3), 305-335.

Weigle, S. C. 1998. Using FACETS to model rater training effects. *Language Testing*, 15(2), 263-287.

Wright, B. J. and Masters, G. N. 1982. *Rating scale analysis*. Chicago. IL: MESA Press.

Wu, M. L., Adams, R. J. and Wilson, M. R. 1998. *ACER ConQuest. Generalized Item Response Modeling software*. The Australian Council for Educational Research.

Appendices

Appendix 1: A summary of the speaking test

Test structure	Task type	Language function	Criteria
Section 1 (30 seconds) Including explanation of test procedures	interview 1-2 set questions	Greeting Stating personal questions	Unassessed
Section 2 (3 minutes) Including explanation and 15 seconds planning time	2 set of questions about the illustration 2 Description	Question and answer Detailed description	Fluency Grammar Vocabulary
Section 3 (3 minutes) Including explanation of test procedures	4 situational questions (response to each Japanese stimulus and answer in English)	Questioning Excusing Refusing	Appropriacy Fluency Grammar
Section 4 (3 minutes) Including explanation of test procedures	Role play (situation dialogue)	Greeting and asking for information Questioning Answering	Appropriacy Fluency Grammar

(Global scoring criteria): Intelligibility (0 to 4 Likert scale). Nonverbal (0 to 4 Likertscale)

Appendix 2: Scoring criteria

A. *Fluency (smoothness, smooth flow of utterances)*
4: speaks fluently with only occasional hesitation
3: speaks with some hesitations without impeding communication
2: a marked degree of hesitation impedes communication
1: speech is fragmented due to unacceptably frequent long hesitation, and pauses.
0: response or irrelevant task

B. *Grammar (control of complex and simple construction and grammatical basic rules)*
4: no major or minor errors in structure
3: no major errors but only a few errors
2: some errors impede communication
1: somewhat frequent minor errors and major errors
0: no response or irrelevant response

C. *Vocabulary (breadth and knowledge of vocabulary)*
4: uses vocabulary precisely and appropriately
3: vocabulary is adequate to express some ideas
2: limited vocabulary restricts expression to simple ideas only
1: very limited vocabulary and only some words and phrases
0: no response or irrelevant response

D. *Appropriacy (the degree of politeness and suitability of timing to prompt)*
4: almost no errors in the socio-cultural conventions of language
3: signs of developing attempt at response to role, and setting. But misunderstandings may occasionally arise through inappropriateness
2: able to operate only in a very limited capacity: responses characterized by socio-cultural inappropriateness
1: unable to function in the spoken language
0: no signs of appropriateness

Global criteria

E. *Intelligibility (naturalness, stress, intonation, rhythm and tone)*

4: no conspicuous mispronunciation but would not be taken for a native speaker

3 : marked 'foreign accent' and occasional mispronunciations which do not interfere with understanding.

2: frequent gross errors and a very heavy accent make understanding difficult, require frequent repetition

1: speech frequently unintelligible

0: no response or unintelligible

F. *Nonverbal (eye contact and gestures and facial expressions)*

4: can communicate effectively with eye contact, gestures, and facial expressions all the time.

3: communicate adequately with eye contact, gestures and facial expressions most time.

2: can sometimes communicate with eye contact, gestures and facial expressions.

1: can rarely communicate with eye contact, gestures and facial expressions.

0: never or lack of communicative delivery