Diachronic and Synchronic Rasch Measurement Analysis of Entrance Examination Listening Tests

David Aline

The beginning of this century has seen a sudden growth in the amount of research conducted on university entrance exams in Japan. While this new research looks at many aspects of the exams, some researchers have started to focus on analysis of the exams for the purpose of improving their quality. This paper is a continuation of an action research cycle through which feedback on the performance of the exams is provided for future exam construction. Using the statistical tool of Rasch measurement, which provides information on item difficulty and examinee ability on a single scale, this paper explores the 2004 and 2005 English entrance exam subtests. The results show that while the exams are adequately measuring the examinees for the purpose of entrance selection, some adjustments could be made to the questions to improve the overall quality and make the exams even better.

Key words: Japanese university entrance exams, Rasch measurement, language testing, action research, reliability

Introduction

Despite the high stakes decisions that are made based on the results of university English language entrance examinations in Japan, little information is made public on the quality of the exams and little research is conducted for the purposes of improving the reliability and validity of the tests. While there is no lack of voices in the controversy

that swirls around the entrance exams concerning their washback effect or lack of effect on teaching (Brown, 1997 ; Cheng, Watanabe, Curtis, 2004 ; Mulvey, 1999 ; Stout, 2003 ; Watanabe, 1996a, 1996b), there is less solid research conducted on the exams themselves. In a call for greater empirical research, Watanabe (Newfields, 2005) criticized policy makers in noting that change in exam policy seems to be based more on general opinion than on empirical evidence. Some research has begun to appear on entrance exams in Japan in the form of both surveys of the level of difficulty of reading passages and type of items employed (Brown & Yamashita, 1995 ; Kikuchi, 2006), studies on the validity of the vocabulary used in the exams (Hasegawa, Chuujyou, & Nishigaki, 2006), theoretical proposals for utilizing exam scores as diagnostic tools and for curriculum design (Weaver, 2005), analyses of optimal number of distractors, (Shizuka, Takeuchi, Yashima, & Yoshizawa, 2006), use of exams in assessing oral communicative competence (Weaver & Romanko, 2005), analysis of exams using Rasch measurement (Weaver & Sato, 2005), and rating scale performance (Weaver, 2006). Though some research in this area is just now beginning to appear on the horizon, in light of the fact that these exams have major ramifications for the future of each examinee, research is woefully lacking in all areas of entrance exam design, production, administration, and analysis.

In a progress report on the state of English language entrance exams in Japan, Brown (2002) outlines a plethora of research questions begging for attention. Two of these questions, falling under the heading of "Roles of assessment," are : (a) How sound are the entrance examinations in terms of reliability and validity? and (b) How can we develop better entrance exams? These questions are central to traditional test development but take on an altogether different character when viewed in light of current entrance exam practices.

A traditional approach to test writing begins with (a) item writing, which includes editing and discussion of the questions among item writers, (b) piloting, which involves administering the test to a sample

population to assess the quality of the items, (c) statistical analysis of the items, (d) revision of the items based on the statistical analysis so that distractors function to reduce achievement through guessing, (e) administration of the test to the target population, (f) reanalysis of the items, (g) banking of items that perform well for the test construct, (h) writing of more items to include in the test, (i) and a continuation of this process to ensure the quality of the exam and exam security. For entrance exams in Japan this process is not possible as the examinees are allowed to take their test booklets home with them after the test administration and the exams are later published.

It is supposed that this practice is in part to ensure fairness in the selection process as the public can judge for itself the quality of the document upon which entrance decisions are based. For example, many examinees take their exam booklets back to their exam preparation schools, where the tests are analyzed to determine the correct answers, and the examinees immediately given their rough test scores. Unfortunately, this practice precludes reuse of the test items on future tests in the university. Consequently, it is impossible to produce a bank of quality items from which to draw for future administrations of a test, and therefore each test must be produced ad hoc for each administration. Another possibility is to pilot the items prior to the actual administration. This would allow for some adjustment of distractors. Regrettably, this procedure is also not practicable due to difficulties with test security and time constraints. Working within these constraints, test writers in Japanese universities have little besides their own intuition to rely on when it comes to preparing a new exam. This has led some researchers to question how, within these constraints, better entrance exams can be developed.

Action Research Approach

One avenue open for improvement of entrance exams is an action research approach. Action research is a process whereby individuals as part of a community of practice pursue change through research on

issues pertinent to their practice in a progressive and cyclic process. In the action research approach to test development presented here and in the other papers of this series, the test writers themselves analyze the tests they have written so as to foster a better understanding of the level of the examinees they are testing in terms of the relation of examinee level with test questions, and to understand the quality of the item types that are being employed. Such an approach has been taken by one research team in a post-administration analysis of previous exams and application of the analysis to the next annual exam writing cycle (Aline & Churchill, 2004, 2006 ; Churchill & Aline, 2004, 2005). This approach is cyclical in nature in that the steps are to be repeated each year for each test. The steps taken, as displayed in the graph in Figure 1, are to first write the test, second administer the test, next analyze the test, then share the findings with fellow item writers, and finally apply these findings to writing the exam for the next year.

The analyses contained herein sustain and augment this line of research by analyzing the English entrance exam listening tests from 2004 and 2005, and comparing the results with the analysis of the 2003 English entrance exam listening test (Aline & Churchill, 2004).

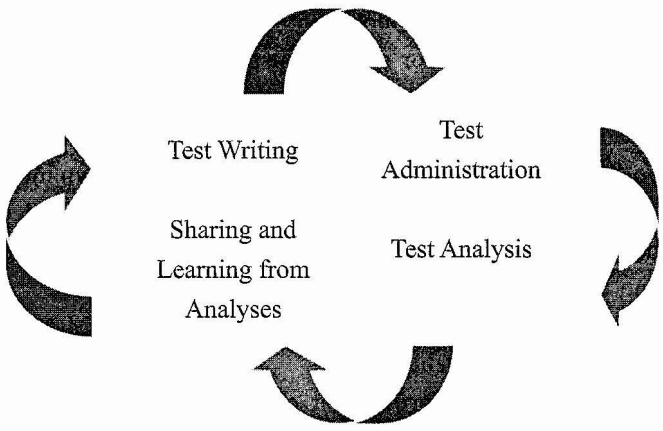

Figure 1. Action Research Approach to Exam Development (Aline & Churchill, 2004 ; Churchill & Aline, 2005)

Purpose of the Research

As a continuation of the action research cycle, this paper advances the analyses of the entrance exams by conducting a comprehensive analysis of the 2004 and 2005 English entrance exam listening tests in terms of the reliability, item functioning, and change over time, and reviews the results of the analysis of the 2003 listening tests in relation to the 2004 and 2005 listening tests. It is anticipated that through analysis of these tests a deeper understanding of the results of the examination can lead to an application of the findings to future test development. For this purpose, the following research questions were outlined:

(1) What is the person sample reliability for the 2004 English entrance exam listening test and for the 2005 English entrance exam listening test? How do these reliability estimates compare with the reliability for the 2003 English entrance exam listening test?
(2) How does the difficulty of the exams compare with the overall examinee ability level? Is there an overall match or mismatch?
(3) Which item types are performing well in measuring the ability of the examinees? Which item types are not performing well?
(4) How have the exams changed over time?

Method

Participants

As stated earlier, the main exams analyzed for this paper are the 2004 and 2005 English entrance exam listening tests. These exams were administered to 160 and 163 examinees, respectively, applying for the departments of Trade and English. The 2003 listening test, previously analyzed (Aline & Churchill, 2004), is included here for diachronic analyses. The 2003 exam was taken by 224 Trade and English majors. Applicants for admission to the university have a number of options to choose from; the listening test is one of these options. Examinees selecting the listening exam must also take written exams in English

and Japanese. Therefore, the examinees who took the listening tests were self-selected, and probably selected the listening test based on their perceived abilities in listening or essay writing. These perceived abilities could be either from self-rating or from their high school teacher's suggestions about the student's strengths in these areas.

Materials

For the purpose of continuity in comparisons of the structure of the listening tests, the structure of the 2003 listening test is first reviewed, and then the structures of the 2004 and 2005 tests are outlined. The listening tests are 45-minute tests, including initial instructions for each section and a few minutes at the end of the listening for examinees to double check their answer sheets. The question format, the same for all three tests, was multiple-choice with four-options for each question.

The 2003 test was produced with six sections : (1) picture matching (12 items), examinees select the picture that best matches the four-turn conversation ; (2) question/answer (15 items), examinees choose the best response to each question ; (3) synonymous statements (12 items), examinees select the sentence closest in meaning to the sentence heard ; (4) conversation meaning (10 items), examinees listen to a two-turn conversation, hear a question about the conversation, and select the best answer to the question ; (5) extended discourse (10 items), examinees listen to an extended conversation of about eight-turns, after which they hear one question for each of three items for each of the three conversations ; and (6) short lectures (6 items), examinees listen to two short lectures and then hear three questions after each lecture. The total number of items on this test was 65. The number of questions for each section and the question numbers assigned to each question for the previously presented analysis (Aline & Churchill, 2004) are given in Table 1.

Table 1 2003 Listening Exam Structure

		Number of questions	Question numbers
Part 1	Picture matching	12	1.1-1.12
Part 2	Question/answer	15	2.1-2.15
Part 3	Synonymous statements	12	3.1-3.12
Part 4	Conversation meaning	10	4.1-4.10
Part 5	Extended discourse	10	5.1-5.10
Part 6	Short lectures	6	6.1-6.6

The 2004 test, with a total of 70 items, was prepared with four sections: (1) question/answer (20 items), examinees choose the best response for each question; (2) statement/response (20 items), examinees select the best response to each statement; (3) conversation meaning (10 items), examinees listen to a two-turn conversation, hear a question about the conversation, and select the best answer out of four from their test booklet for that question; (4) short lectures (20 items), examinees listen to two short talks, played twice, and then choose the best completion to a sentence stem. The number of questions for each section and the question numbers assigned to each question for this analysis are presented in Table 2.

Table 2 2004 Listening Exam Structure

		Number of questions	Question numbers
Part 1	Question/answer	20	1.1-1.20
Part 2	Statement/response	20	2.1-2.20
Part 3	Conversation meaning	10	3.1-3.10
Part 4	Short lectures	20	4.1-4.20

The 2005 test was constructed with five sections with a total of 69 items: (1) question/answer (20 items), examinees choose the best response for each question; (2) statement or question/response (15 items), examinees select the best response to each statement or question; (3) conversation meaning (12 items), examinees listen to a two

-turn conversation, hear a question about the conversation, and select the best answer to the question; (4) extended discourse (9 items), examinees listen to three extended conversations of about nine-turns each, followed by three spoken questions for each conversation; (5) short lectures (13 items), examinees listen to three short talks, hear four or five questions after each lecture, and select the appropriate answer for each question. The number of questions for each section and the question numbers assigned to each question for this analysis are presented in Table 3.

Table 3 2005 Listening Exam Structure

		Number of questions	Question numbers
Part 1	Question/answer	20	1.1-1.20
Part 2	Statement or question/response	15	2.1-2.15
Part 3	Conversation meaning	12	3.1-3.12
Part 4	Extended discourse	9	4.1-4.9
Part 5	Short lectures	13	5.1-5.13

Results

Analysis

The statistical analyses for this study were conducted through WINSTEPS (Linacre, 2003), a computer software program which utilizes Rasch measurement for analyses. A dichotomous Rasch analysis looks at the likelihood of an examinee (termed "person" in Winsteps) answering a question on the exam correctly, and, conversely, at the difficulty of an item to be answered correctly. Rasch provides an estimate of an examinee's ability level through an analysis of the probability of that examinee marking any given answer as correct or incorrect. The examinee is then placed on an ability level scale with the other examinees ranked in order of their abilities on the test, from having a high probability of answering correctly to having a low probability. Furthermore, Rasch estimates the difficulty levels for the

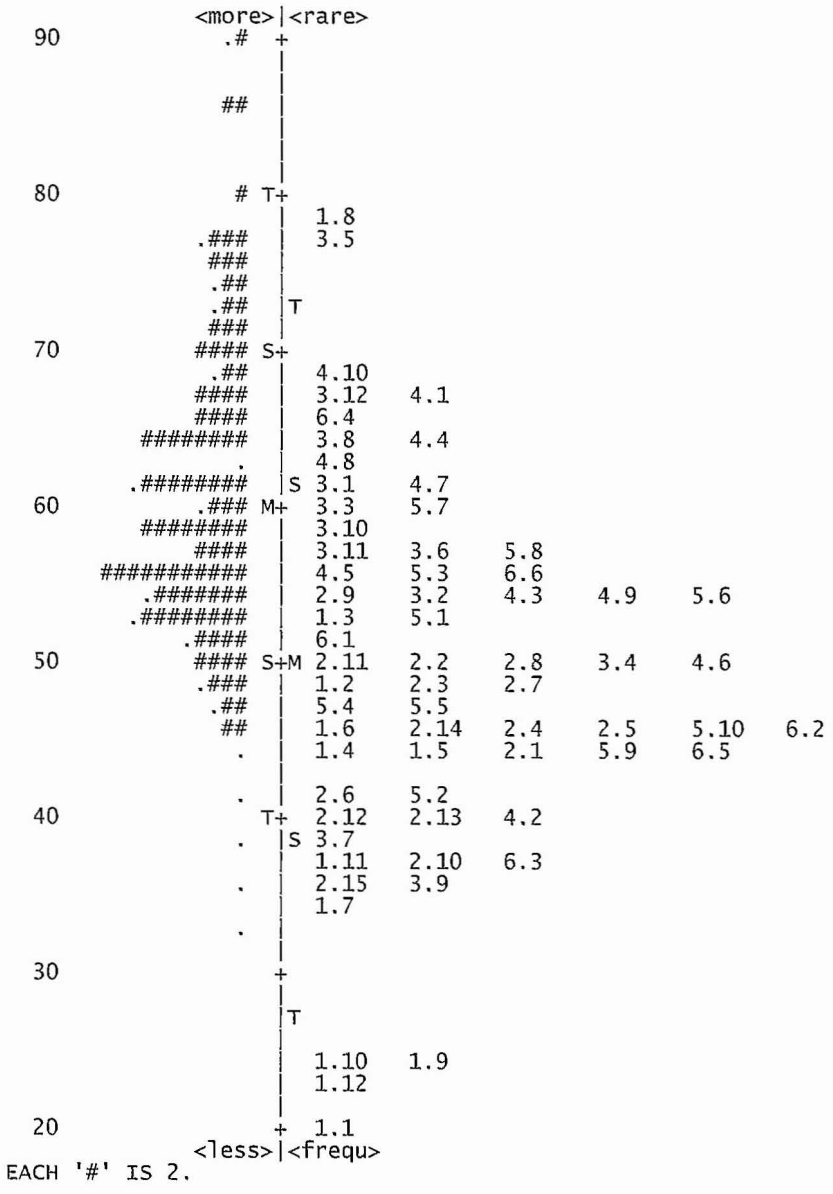

Figure 2. Person-Item Map for 2003 English Listening Exam

test items and ranks them according to their difficulty, very difficult to correctly answer to very easy to correctly answer, on the same interval scale on which the examinees' are ranked according to ability. This can best be understood by looking at one of the output tables produced by Winsteps. Figure 2 presents the person-item map (examinee-question graph) for the 2003 listening test.

Each hash mark, "#", on the left side of the center line, the interval scale, represents, for this analysis, two examinees. On the far left of the graph is the probability scale, represented in units ranging from 1 to 100, with 10 units on the scale equaling one logit, the unit of measurement used in Rasch. The ability scale runs from less ability at the bottom to greater ability at the top. On the right side of the scale are the exam item numbers ranked from top to bottom according to difficulty estimates, with the most difficult items toward the top and the items that are easier to answer correctly nearer the bottom. For example, an examinee ranked on the ability scale at a probability of 50 has a 50 percent chance of correctly answering an item ranked at the same level, such as items 2.11, 2.2, 2.8, 3.4, and 4.6 in this analysis. This same examinee has a much higher likelihood of correctly marking item 3.7, ranked at 39 on the scale, and has a much lower likelihood of getting a correct answer for item 3.5, ranked at 78 on the scale. In the same way, an examinee ranked at 68 on the scale has a 50 percent chance of correctly answering items 3.12 and 4.1, and a progressively increasing likelihood of correctly answering all the items ranked below that point on the scale. Therefore, the scale provides an easy overview of the examinee abilities and the item difficulties on one scale, allowing for a quick understanding of exam difficulty and examinee ability.

Reliability

Reliability, a ratio of true variance and observed variance, is a statistic expressing the reproducibility of a test, but not the quality or accuracy. If a test were administered a second time to the same examinees, to what degree would it provide the same score for each

examinee? This is a question of consistency rather than accuracy in measuring a construct. Nevertheless, reliability does provide some measure of the quality of an exam in that any measure of language would be of little value if it failed to measure without a significant degree of consistency. Winsteps reports reliability as "person sample" reliability, which is equal to the test reliability of classical test theory. As shown in Table 4, the reliability estimates for the listening tests from 2003 to 2005 are .88, .87, and .86, respectively.

Table 4 Reliability for Three Listening Exams

Exam year	Number of examinees	Reliability
2003	224	.88
2004	160	.87
2005	163	.86

As a rule-of-thumb, reliability estimates below .8 require some adjustments be made to the exam. While these estimates are acceptable, there is still some room for improvement. If these exams were to be repeatedly administered, the reliability would need to be improved. It is imperative to remember that the listening tests are subtests of a battery of exams upon which the entrance decisions are based. With the increase in the number of items in aggregate, the reliability will be higher.

Two ways to improve reliability are (a) by testing a wider ability range, and (b) increasing the person measurement precision so as to decrease the average person standard error by increasing the number of items on the test. The feasibility of testing a wider ability range is precluded by the fact that the examinees select themselves, and thus examinees in the wider ability range have opted not to take this test because of their lower ability. Moreover, the examinees taking this test are in a narrow band of ability as they have been through an educational system of tests and teacher recommendations that have directed high school students with similar abilities to apply for this level of

university and this type of test. The second approach is to increase the number of items on the test. This solution is discussed below.

Standard Error

One measure of test quality is standard error (S.E.). Standard error provides information about the precision of the exam. To rectify problems with high S.E.s for the examinees, the test length should be increased, and for high S.E.s for the items, the number of examinees should be increased. These conditions are similar to those of reliability estimates because lowering the S.E.s serves to improve reliability. The S.E.s for persons and items on the 2003, 2004, and 2005 listening tests administrations are presented in Table 5.

Table 5 Standard Error for Three Listening Exams

Exam year	Person standard error	Item standard error
2003	3.37	1.81
2004	3.28	2.19
2005	2.94	2.02

The S.E.s for all of these exams are relatively high. These measures indicate that the listening test in general is in need of some adjustments in order to improve the capacity of the exam to accurately measure the examinees' competence in understanding spoken English. The obvious solution seems to be to increase the number of items on the exams so as to lower the standard error and increase the reliability. However, it is imperative first to look at the quality of the questions and how they are functioning in relation to the examinees as there may be other means by which to lower the standard error.

Average Measure

The average person measure is the mean of the examinee ability ranking. The average item measure is the mean of the item difficulty and is set at a standard of 50 out of 100 for the Winsteps' person-item

map. A simple comparison of these two measures on an exam evinces any disparity between the level of the examinees and the level of the test. Table 6 reports the person average measures and item average measures for the three exams under review. The person average measure, showing the ability level of the examinees, is noticeably higher on all three tests than the item average measure, the difficulty level of the items. These numbers clearly demonstrate that the test questions in aggregate were far too easy on the 2003 and 2004 exams, and a bit too easy on the 2005 exam. Looking back at the 2003 listening exam person-item map in Figure 2, the mismatch between the person ability and item difficulty is unmistakable. The person average measure is indicated with an "M" on the left side of the middle line, and the item average measure is indicated with an "M" on the right side. Since the difference is ten units apart on the scale, and the examinees are distributed in a curve that sits much higher on the scale than the distribution of the item difficulty on the right side, it is clear that the difficulty of this exam could be significantly increased. With a difference of nearly 12 units, the 2004 listening test, graphically displayed in Figure 3, is obviously not well matched to the ability level of the examinees as a whole. Whereas the 2005 test, with a difference of only 4.5 units, exhibits a closer match between test and examinees. This closer match, however, is not reflected in the standard error or reliability of the 2005 test as this test is as low on those measures as the other tests.

Table 6 Average Measure for Three Listening Exams

Exam year	Person average measure	Item average measure
2003	60.64	50.00
2004	61.57	50.00
2005	54.50	50.00

Person-Item Map : 2003 Exam

The 2003 English listening subtest, shown in Figure 2, was previously examined (Aline & Churchill, 2004) and the analyses discussed in terms of item difficulty, discrimination, and INFIT and OUTFIT. The recommendations for future test writing drawn from these analyses were : (a) remove Section 1 (picture matching) as it was simply too easy, (b) perform an in depth analysis of the other sections to capture a more vivid picture of which items could benefit from a clearer focus, and (c) increase the difficulty of the items in Section 6 (short lectures). Section 1 was subsequently removed, although not as a direct result of these analyses, but rather due to the difficulty of finding satisfactory and unambiguous pictures and to the intuition of the test writers that the section was too simple. The most salient deficits of this test, from the panoramic perspective of the person-item map in Figure 2, are (a) the number of items of extremely low difficulty, mostly from Section 1, that contribute little to and actually detract from the accurate measurement of the examinees, and (b) the paucity of items in the higher difficulty range.

Person-Item Map : 2004 Exam

Figure 3 presents the person-item map produced by Winsteps. While the test items on the whole are too easy for these examinees, taking a closer look at each section tells a different story. The items from Section 1, question/answer, numbered 1.1 to 1.20, range from a low difficulty of about 25 to a high of 65. As the test is presumably designed to be progressively more difficult, the range of the first section measuring the lower ability levels is acceptable. Section 2, statement/response, numbered 2.1 to 2.20, ranges from a low of 42 to a high of 80, which provides good coverage of the 40 to 92 ability range of the examinees. Section 3, conversation meaning, item numbers 3.1 to 3.10, covers a range from 35 to 78, although with some items falling far below the ability levels of these examinees and a large gap in the higher measurement range. The difficulty of this section could be slightly increased by

increasing the turn length of the conversations, which consist mostly of simple sentences in this version of the test.

The section with the greatest mismatch with the examinee ability is Section 4, short lectures. The items, numbered 4.1 to 4.20, range from 25 to 66, but with more than half falling below the average difficulty of the items. This section should be measuring the higher ability ranges of the examinees. The very high mismatch between the difficulty of Section 4 and the average examinee ability could explain much of the high standard error and low reliability. In comparison with the 2003 and 2005 tests, the lectures for the 2004 test are relatively short. The 2003 test has two lectures of about 19 lines each ; the 2004 test has four lectures of about 5 lines each, and the 2005 test has three lectures, two of 6 lines each and one of 9 lines. Increasing the length of the short lectures would be one alternative to attempting to increase the difficulty of the items. However, there are other differences that could be affecting the difficulty. Both the 2003 and 2005 tests use spoken questions about the lectures with only the four multiple-choice answers written in the test, while the 2003 test uses a sentence stem written in the test booklet with the four multiple-choice answers. There is the possibility that this question type is fundamentally easier, but this assumption requires further research. Another possibility is that the answers themselves are easy. In reviewing the answers to the items in Section 4 with very low difficulty, it can be seen that the correct answers often utilize the same vocabulary or phrases as occurred in the short lectures, such as "by car/by car," "Kanto and Kansai/Kanto and Kansai," "at least 38 magnificent plays/at least 38 plays," "American pioneer/American pioneer," "simple tools/simple tools," etcetera. The examinees may have simply selected the answers based on the similarity of the words heard. Answers that used synonymous words or phrases were more difficult, and those questions that required some inference or used similar words in both the correct answer and the distractors were the most difficult. To increase the difficulty of the short lecture section, more questions requiring the examinees to infer the answers should be

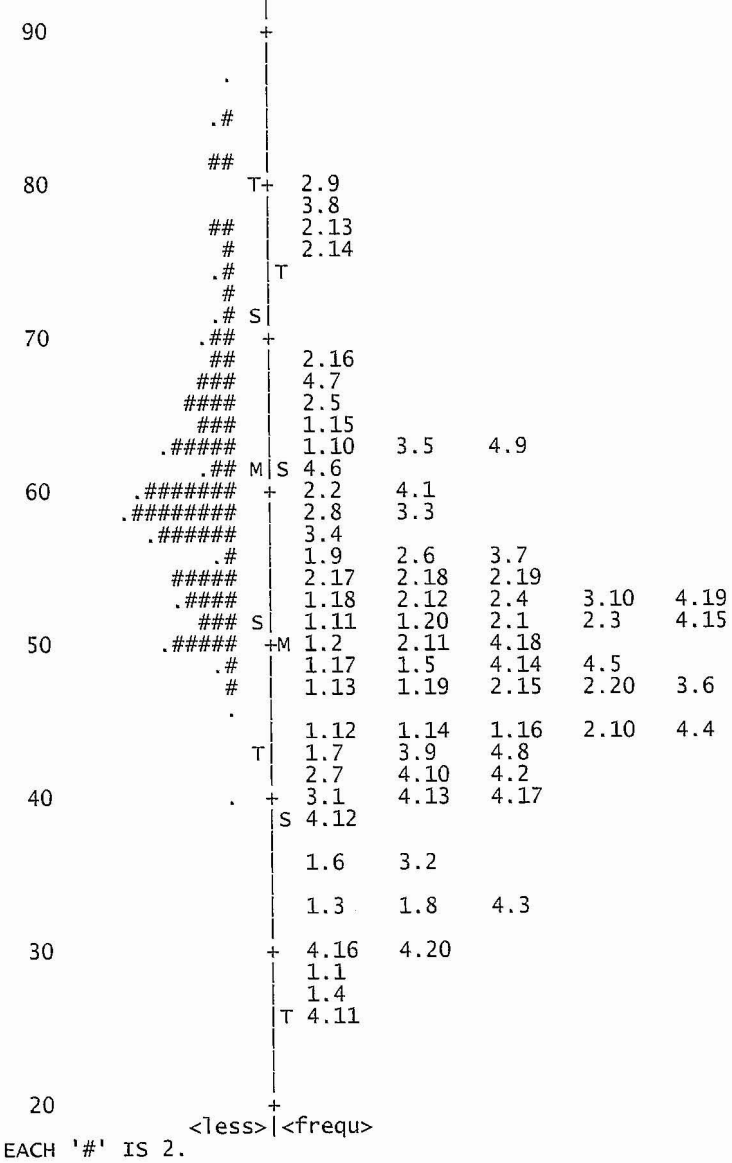

Figure 3. Person-Item Map for 2004 English Listening Exam

used and repetition of the same words or phrases from the lectures in the answers should be avoided.

Person-Item Map : 2005 Exam

The results of the 2005 English listening test are graphically portrayed in the person-item map in Figure 4. The difficulty of the exam overall corresponds well with the ability range of these examinees. Moreover, each of the individual sections complements the examinees' range. The ability of most of the examinees ranges from about 40 to just under 70.This range is covered satisfactorily by each of the five sections with a good distribution of items. However, if the test were designed to incrementally increase in difficulty with each section, this map would provide evidence that there is no increase in difficulty by sections and that some redesign of the exam would be warranted. On the other hand, in its current form the exam discriminates well between the examinees as there are many items measuring each of the levels, and each level is being measured by the different types of items from each section.

A number of items on this exam are far below the ability levels of the examinees on this test, and are decreasing the quality of the test by markedly increasing the standard error. For example, the four items with the lowest difficulty estimates, 2.3, 5.9, 4.4, and 1.11, have standard errors of 7.15, 3.91, 3.32, and 3.06, respectively. Proscribing such items from future exams will lower the standard error of the test, increase accuracy of the measurement, and improve reliability. One problem with some of these items is similar to that found with the 2004 exam, namely that the examinees can guess by matching vocabulary in the text with the words in the answer. Items 4.1 and 4.4 are relatively easy on this exam. The correct answer for 4.1 is "Basketball" which is repeated four times in the text, and the correct answer for 4.4 is "camera," repeated three times in the text. Similarly, in Section 3, item 3.6 uses "tastes" and "taste" in the text and "tasted" in the correct answer, and item 3.2 uses the word "test" at the end of the short dialog

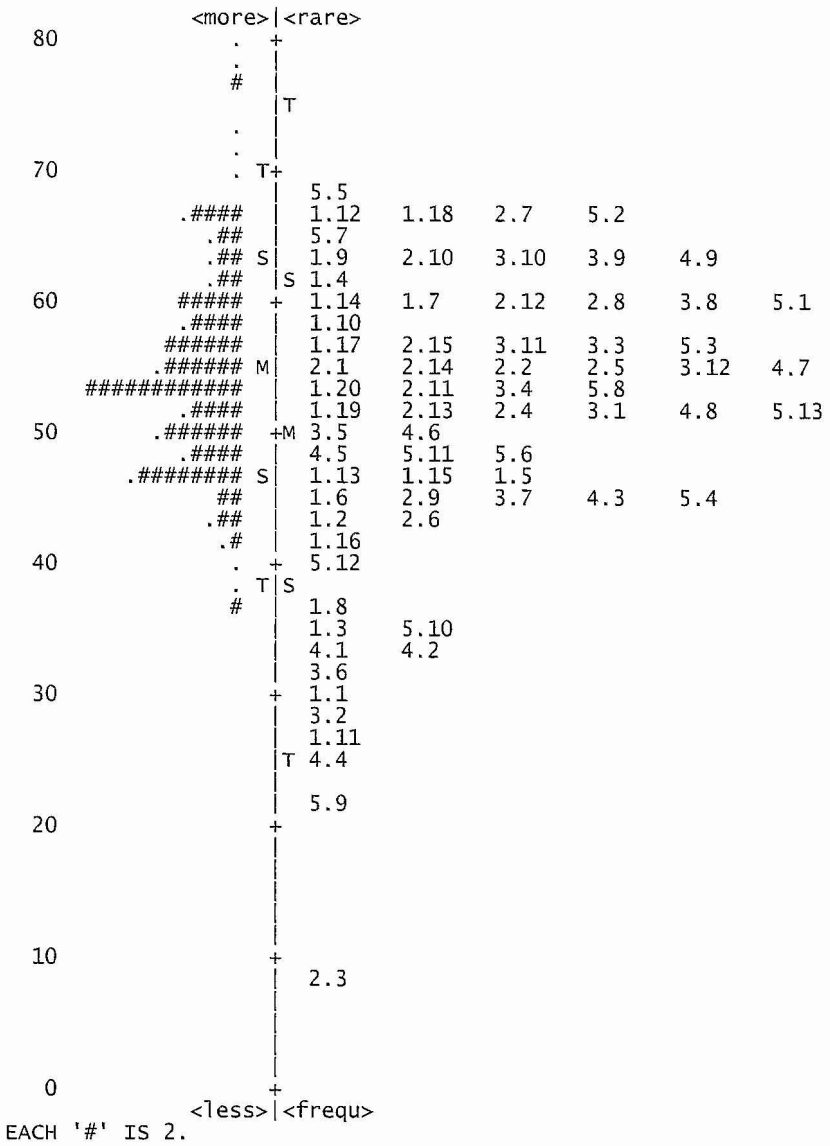

Figure 4. Person-Item Map for 2005 English Listening Exam

and the same word in the answer. In item 3.2 the word "movie" also appears, and the distractor using this word was the most selected distractor on this item, thus providing some indication of the presence of guessing. If an examinee were to employ guessing strategies based on word frequency or similar words, they could select the correct answers to these items. Although this does not explain the lack of difficulty on these items, it is a phenomenon to be avoided on future tests, or possibly to be used as a distractor in order to diminish the effect of employing this guessing strategy.

INFIT / OUTFIT Statistics

Although there are other statistics produced by Rasch that provide a variety of information about how each test and the specific test items were functioning, such as INFIT and OUTIFT statistics, there were no salient patterns exposing any problems with the test. On the 2004 and 2005 listening tests in this analysis, the INFIT and OUTFIT statistics exhibit a very acceptable pattern. When INFIT and OUTFIT mean -square statistics fall outside of the range from 0.5 to 2.0, then the exam is not measuring accurately and further examination of the specific items is warranted. All of the mean-squares are well within this range, indicating that the test is measuring accurately.

Conclusion

This paper examined the 2004 and 2005 English language entrance exam listening sub-tests and reviewed the findings from the 2003 listening test analysis (Aline & Churchill, 2004) as part of a continuation of the action research cycle implemented in previous research (Aline & Churchill, 2004, 2006 ; Churchill & Aline, 2004, 2005). The results of the analyses employing Rasch measurement affirm that the listening tests are functioning adequately to measure the examinees as the reliability estimates are at acceptable levels, the items are distributed across the range of examinee ability, there is a sufficient number of items measuring the examinees at the various ability levels serving as entrance

selection cutoff points on each test, and the quality of the tests as measured by standard error and INFIT and OUTFIT statistics is acceptable.

The research questions framed for this study outlined a basic analysis of the quality of the listening exams, a quality that has been found to be acceptable for the purposes of the exams.

The first research question, pertaining to the reliability levels of the exams, has elucidated reliability estimates that are all but the same across the three years of the exams analyzed here, demonstrating continuity of production of reliable exams, and show the exams to be measuring with consistent accuracy both diachronically and synchronically.

The difficulty of the exams has been found to be incompatibly lower than the examinees' ability level as measured on these exams. The large number of items falling far below the examinees' ability level, while not detracting to any great degree from the quality of the exams, could be removed on future exams through closer analysis of the items so as to improve the accuracy of the exams.

While previous research has demonstrated how some item types on the exams could be removed or adjusted (Aline & Churchill, 2004, 2006 ; Churchill & Aline, 2004, 2005), the results presented here drew the locus of inspection to problems within specific questions rather than to any inadequacies of the major types. While it was found that some items types, such as the short lectures, could be adjusted to measure higher levels of ability, there were no egregious deficiencies in the broadly defined overall types.

As noted above, the exams have changed little over time in terms of reliability. The average measure indicates some movement toward an exam offering a preferable match between item difficulty and person ability. More meaningful comparisons of the exams over time are somewhat impeded by the changing structure of the exams. As the use of listening comprehension materials increases in high schools throughout Japan, the listening tests should be developed to reflect the tasks

employed in teaching language so as to augment the validity of the exams.

The analyses of the English entrance exam listening tests presented here represent but one stage in the cycle of action research described above. It remains for item writers to apply the conclusions they draw from these analyses to the next step in the process, preparing the next listening test. One principle of success for action research is the working together of individuals in a community of practice to collaboratively solve problems they mutually face in endeavors for organizational change. Refining the quality of entrance exams necessitates participation of all members of the community of practice in all stages of the action research cycle so that results of analyses by the community becomes bottom-up feedback upon which the community can proceed with the next action, and individual analysis becomes individual action through personal analysis of one's own production in the exam preparation cycle.

References

Aline, D., & Churchill, E. (2004). *Item analysis of a university entrance examination listening test*. Paper presented at the annual meeting of the Temple University Applied Linguistics Colloquium 2004, Tokyo, Japan.

Aline, D., & Churchill, E. (2006). Analyzing entrance exam item types with Rasch. *Kanagawa University Studies in Language*, 28, 125-142.

Brown, J.D. (1997). Do tests washback on the language classroom? *TESOLANZ Journal*, 5, 63-80.

Brown, J.D. (2002). English language entrance examinations: A progress report. In A.S. Mackenzie & T. Newfield (Eds.), *Curriculum Innovation, Testing and Evaluation : Proceedings of the JALT Pan-SIG Conference*, Kyoto, Japan (pp.95-105). Tokyo: Japan Association for Language Learning.

Brown, J.D., & Yamashita, S.O. (1995). English language entrance examination at Japanese universities: What do we know about them? *JALT Journal, 17*, 7-30.

Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing : Research contexts and methods*. Mahwah, NJ: Erlbaum.

Churchill, E., & Aline, D. (2004). *Applying Rasch measurement to entrance exams*. Paper presented at the annual meeting of the Japan Language Testing Association, Hikarigaoka, Chiba, Japan.

Churchill, E., & Aline, D. (2005). Applying Rasch measurement to the analysis of entrance exam item types. *Kanagawa University Studies in Language, 28*, 125-142.

Hasegawa, S., Chuujyou K., & Nishigaki, C. (2006). A chronological study of the level of difficulty and the usability of the English vocabulary used in university entrance examinations. *JALT Journal, 28*, 115-134.

Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities after a decade. *JALT Journal, 28*, 77-96.

Linacre, J.M. (2003). WINSTEPS Rasch measurement computer program. Chicago : Winsteps. com.

Mulvey, B. (1999). A myth of influence : Japanese university entrance exams and their effect on junior and senior high school reading pedagogy. *JALT Journal, 21*, 125-142.

Newfields, T. (2005). Voices in the field : An interview with Yoshinori Watanabe. *JALT Testing & Evaluation SIG Newsletter, 9*(1), 5-7.

Shizuka, T., Osamu, T., Yashima, T., & Yoshizawa, K. (2006). A comparison of three-and four-option English tests for university entrance selection purposes in Japan. *Language Testing, 23*, 35-57.

Stout, M. (2003). Do the university entrance exams in Japan affect what is taught? *The ETJ Journal, 4*, 1-7.

Watanabe, Y. (1996a). Investigating washback in Japanese EFL classrooms : Problems of methodology. In G. Wigglesworth & C. Elder (Eds.), *The language testing cycle : From inception to washback* (Series S, Number 13) (pp.208-239). Canberra, Australia : Applied Linguistics Association of Australia.

Watanabe, Y. (1996b). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13*, 318-333.

Weaver, C. (2005). How entrance examination scores can inform more than just admission decisions. In P. Ross, T. Newfields, Y. Ishida, M. Chapman, & M. Kawate-Mierzejewska (Eds.), *Lifelong Learning : Proceedings of the JALT Pan-SIG Conference*, Tokyo, Japan, (pp.114-121). Tokyo : Japan Association for Language Learning.

Weaver, C. (2006). Evaluating the use of rating scales in a high-stakes Japanese university entrance examination. *Spaan Fellow working papers in second or foreign language assessment*, 4, 57-79.

Weaver, C., & Romanko, R. (2005). Assessing oral communicative competence in a university entrance examination. *The Language Teacher, 29*(1), 3J-9.

Weaver, C., & Sato, Y. (2005). Kobetsu-gakuryoku-kensa (eigo) no Rasch bunseki [A Rasch analysis of the English section of a university entrance examination]. *Daigaku Nyuushi Kenkyuu Journal, 15*, 147-153.