Analyzing Entrance Exam Item Types with Rasch*

David Aline
Eton Churchill

大学入試は日本人英語学習者にとって最も大切な節目の一つである。しかし，入試問題の分析はめったに行われていなく，その分析の発表の数は更に少ないと言えるだろう。本研究で，我々はアクション・リサーチの視点に基づき改正されたテスト開発サイクルを提案する。そしてこのサイクルの応用を単一の試験（2003 年 2 月 B 神奈川大学英語試験）のラッシュ分析を通し説明する。試験全体の信頼性の説明の後，どのような種類の設問が有効でないかを見分ける為特定の項目を分析する。本研究ではラッシュがいかに問題のある項目を算出するか示し，項目能力に影響を及ぼす要素は何であるかを検討し，そして将来の試験作成者への提案を提供する。我々はこの分析を用いる事によって今後の試験の開発の為により良い選択が出来る事を示す。

Key words: Japanese education, entrance exams, Item Response Theory, Rasch Analysis, language testing

Introduction

Entrance exam test writers in Japan are caught between a rock and a hard place when it comes to test development. Published analyses of past exams reveal that there is ample room for improvement with regards to test reliability (e.g., Brown & Yamashita, 1995), and yet test developers' hands are tied in terms of their ability to refine exams. For security reasons, exam writers are discouraged from piloting test items, and due to the publication of tests following their administration, item

banking becomes problematic. Under these constraints, it is difficult to follow the traditional test development cycle of item development, item piloting, test development and analysis. Rather, in the test development phase, the prevailing assumption appears to be that test writing committees can do little more than rely on their collective intuitions about how items and distractors may perform. As a result, most entrance exams in Japan are administered with no prior knowledge of how well items are measuring the ability of examinees and no indication of the overall reliability of the exams. Given the high-stakes nature of these tests in the Japanese educational system, the question of what can be done to improve this state of affairs is begging to be answered.

Purpose of the Research

Given the apparent need to improve the reliability of entrance exams in Japan and the systematic constraints on test development, it is the purpose of this paper to advocate an alternative approach to the creation of entrance exams in Japan. We first propose a revised test development cycle drawing from an action research perspective. We then illustrate three stages in this cycle in a case study of one exam. We begin with an analysis of a recent entrance exam drawing on item response theory. Following a presentation of the overall reliability of the exam, we analyze specific items in the interest of identifying question types that are not performing well. We argue that through this analysis, we can learn to make better informed decisions in the construction of future exams.

An Action Research Approach to Test Development

Ideally, by the time an examinee sits for an exam, the items making up the exam have been reviewed and edited, distractors have been revised and the items have been piloted on a similar population of

examinees. Items performing well are then banked for future use. It is also standard practice to ensure that the reliability of the exam is sufficiently high, so that all stakeholders in the exam can be confident about what it is purporting to measure. In order to carry out the piloting and banking of items and in calculating the reliability of tests, test developers depend on test security to ensure that particular examinees do not have an unfair advantage in the actual administration of the test. Unfortunately, in the Japanese context, in which universities regularly release exam booklets on the day of the exam and later publish the exams, it is not possible to bank items that perform well so that they can be used on future exams. Moreover, the test development and editing cycle, which usually runs from April through the fall, does not allow much time to pilot items and this practice is actively discouraged for security reasons at most institutions. As a result, in contrast to many professionally developed tests (e.g., the TOEFL, TOEIC), most entrance exams administered in Japan rarely make it past the stage of test writing, revision and editing because of the various systematic constraints outlined above.

While acknowledging these constraints, we believe that more can be done to improve the quality of entrance exams in Japan. In particular, we would like to suggest that exam writers can make better informed choices in the test writing and revision stage if they can benefit from a post-hoc analysis of previous exams. Based on this perspective, we are currently taking an action research approach to our test development. In our approach (See Figure 1), beginning with the administration of a set of entrance exams, we conduct an analysis through which we attempt to better educate ourselves about how the tests are performing, how the structure of the test is affecting the results and how specific test sections and items are performing. We are sharing our findings with fellow exam writers currently involved in the process of test construction through informal discussions, in-house presentations (Aline & Churchill, 2004) and our university's bulletin (Churchill & Aline, 2005). It is our belief that this process allows us all to make

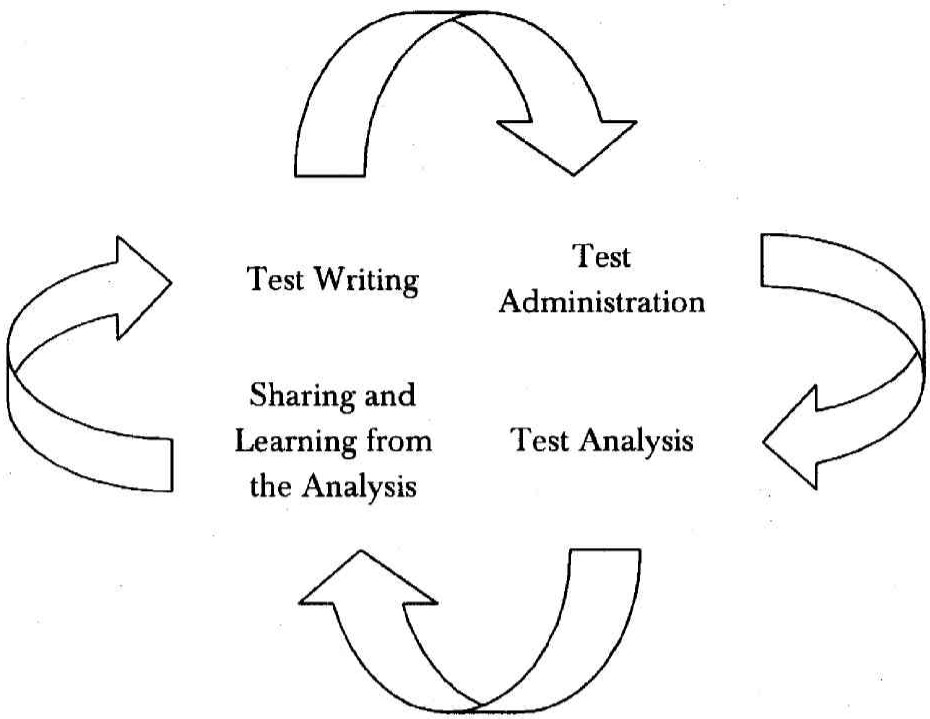

Figure 1. An Action Research Approach to Test Development

more informed decisions in the creation of future exams. To illustrate how this approach works, we have outlined our analysis of our 2003 February B English entrance exam below. Guiding us in our analysis were the following research questions:

1) What is the reliability, as measured in a Rasch Analysis, of the 2003 February B English Exam?
2) Which question types are not performing well in terms of item fit and discrimination? In other words, which item types should we avoid or modify in the writing of future exams?
3) Which question types are performing well? In other words, which item types should we continue working with on future exams?

Method

Participants

The exam that we analyze here is the February B exam administered in 2003. Sitting for this exam, there were 3,968 examinees applying to a variety of departments of which there were 239 prospective English and Spanish majors. Only this subset of the examinees was required to take all five sections of the test. The remaining 3,729 examinees were evaluated based on the first three sections of the test (a total of 45 questions).

The 2003 February B Exam

There were five sections in the exam comprising a total of 82 questions as outlined in Figure 2. Item numbers are identified here and

Section 1 (Reading text: Becoming a Flutist) 12 items (1.1–1.12)
- 4 vocabulary questions (1.1–1.4)
- 4 synonymous phrase questions (1.5-1.8)
- 4 true-false questions (1.9–1.12)

Section 2 (Reading text: Cotton) 14 items (2.1–2.13)
- 6 vocabulary questions (2.1–2.6)
- 3 synonymous phrase questions (2.7–2.9)
- 4 true-false questions (2.10–2.13)

Section 3 ("Conversations" and Grammar) 20 items (3.1–3.20)
- 10 conversation completion questions
- 10 grammar questions

Section 4 (Reading text: Junko Tabei and Mt. Everest) 19 items (4.1–19)
- 6 vocabulary questions (4.1–4.6)
- 4 synonymous phrase questions (4.7–4.10)
- 4 true-false questions (4.11–4.14)
- 5 identifying the accent (4.15–4.19)

Section 5 (Grammar) 18 items (5.1–5.18)
- 8 questions (one paragraph) on the article system (5.1–5.8)
- 5 complete the sentence questions (Grammar and Vocabulary) (5.9–13)
- 5 questions on conjunctions (5.14–19)

All items are multiple choice except for the true-false questions

Figure 2. Test Structure of the 2003 February B Exam

throughout this paper by decimal numbers such that, for example, 5.9 indicates item 9 in section 5. Briefly going over the structure of the test, there were 3 sections based on long reading passages (500–700 words)—Sections 1, 2 and 4. In these sections, questions focused on vocabulary and comprehension of the texts based on true-false statements. Section 4, taken only by the English and Spanish majors, differed from Sections 1 and 2 only in that an additional 5 items on identifying the accent in words were added to the section. Sections 3 and 5 focused primarily on grammar with the exception being that Section 3 also included 10 questions aimed at testing coherence in short conversations. While Section 3 tested more general features of grammar, Section 5 focused in on specific elements (e.g., the article system and conjunctions).

WINSTEPS and Rasch measurement

The Rasch analysis performed on the participant responses was conducted with a WINSTEPS version 3.47 statistical package. Rasch analysis distinguishes itself from approaches used in classical testing theory in that it looks at the interaction between test items and examinees through the lens of probability (Bond & Fox, 2001). The output in a Rasch analysis allows one to examine and compare the relative difficulty of items and the ability of examinees. Moreover, a hierarchical map of participant ability is statistically linked to and juxtaposed with a corresponding map of item difficulty on a single interval scale. As illustrated in Figure 3, the person-map for the 2003 February B Exam, this allows for easy comparison of overall test difficulty and the ability of the examinees.

As we will illustrate later in our analysis, a Rasch analysis also allows one to compare the average ability of examinees who have chosen different distractors for a particular item. This information is helpful in determining which items and distractors are performing as expected and which ones may be candidates for revisions.

Analysis

As outlined in the discussion of the participants and the exam, two groups of examinees took different versions of the test—all 82 questions for the Spanish and English majors and only the first 45 questions for the other majors. As a result, the analysis of the last two sections of the test was based strictly on the performance of the Spanish and English majors on these items.

Results

Effects of Test Design and Administration on Reliability

To appreciate how well the test reliably measured the two distinct groups of examinees, analysis was performed on Sections 1–5 of the test for the Spanish and English majors and then on Sections 1–3 of the test for the remaining examinees. As we can see in Table 1, the five sections (82 items) taken by the 239 Spanish/English examinees led to a respectable person reliability of .83. Meanwhile, for the majority of the examinees, who were only required to take the first three sections of the test (45 items), the person reliability was considerably lower at .70. Thus, one of the first points observed about the design and administration of the February B 2003 English Exam (and others that we have analyzed with a similar structure—e.g., Churchill & Aline, 2005)—that is a different number of examinees taking different sections—is that they lead to different person reliability measures for the different sub-groups of examinees taking different sections (Sections 1–3 vs. Sections 1–5) of the exam. The smaller group taking more items is likely to exhibit a higher reliability while the larger group of

Table 1. Effects of Test Design and Administration on Reliability

	Examinees	Mean Measure	# of Items	Person Reliability
Spanish/English Majors	239	55.44	82	.83
Other Majors	3729	51.08	45	.70

examinees taking fewer items is likely to have a lower person reliability.

The Person Map of Items

In an initial analysis of all examinees and all test items, Rasch produced the following person map of items (Figure 3). In a person map of items, the examinees are located on the left hand side of the map with their relative ability in ascending order. At the same time, items are displayed on the right hand side of the map—again, in ascending order in terms of difficulty. The scale on the far left ranging from 20 to 80 is not a raw score, but rather the measurement unit common to both person ability and item difficulty, adjusted here to center on 50, the default setting for the average item difficulty in our analysis. This is a scale of equal intervals such that the difference between 20 and 30 and 60 and 70 is the same. However, for formatting reasons, this map has been truncated a bit on both ends.

In this particular map, each pound sign represents 40 examinees and each period (.) represents a smaller number. As mentioned earlier, to identify items in our analysis, we are using decimal numbers such that 3.2 represents the second item in Section 3. This item has a low difficulty measure of 24. Looking at one further example, item 5.9 has a difficulty measure of 56. Examinees with this same ability measure represented by the four pound marks and one period amount to 152+ examinees. Since these examinees have the same ability measure as the difficulty of item 5.9, they would have about a 50% chance of getting this item correct. On the other hand, they would have a much stronger chance of getting item 3.2 correct and a lower chance of getting a difficult item such as 1.4 correct. In this way, we can easily identify items from different sections on the map and compare their difficulty with that of other items and with the ability of examinees.

On this scale, the mean difficulty of items (+M) is arbitrarily set at 50 and all other measures are calculated in relation to this figure. On the right side of the map, +S and +T signify respectively the first and second standard deviations for the items. Looking closer at this map,

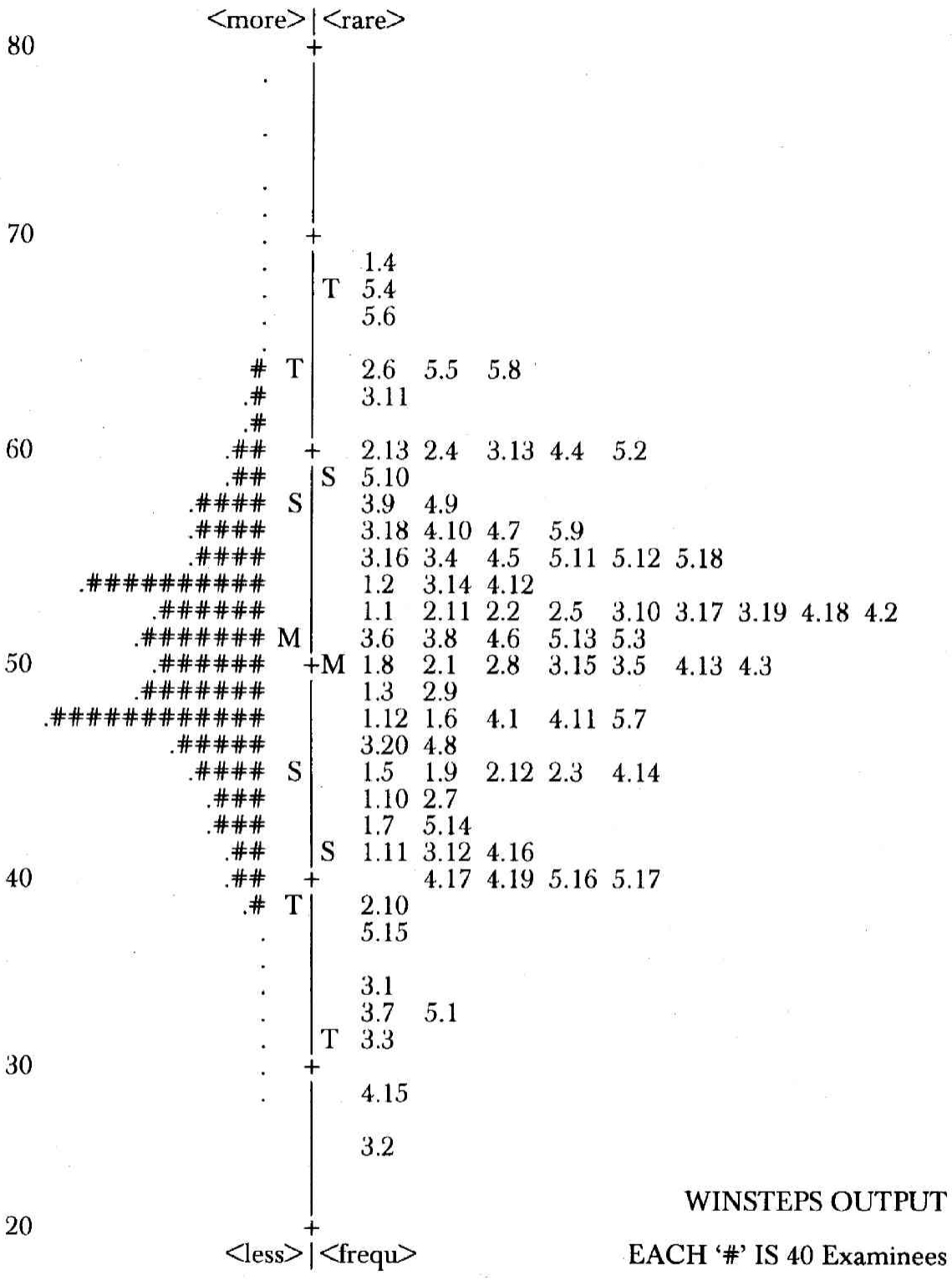

Figure 3. Person Map of Items for all Examinees for the 2003 February D Exam

the first thing to note is that there is a very good match between the difficulty of the test and the ability of the examinees as the means for both the items and the examinees are just about at 50. In cases where a test is too difficult for the examinees, the mean ability will be lower, while in cases where examinees find a test less difficult, the mean will be higher. In fact, as noted earlier, this was the case for the Spanish/English majors taking this exam, as their mean ability was 55 (See Table 1). On the left side of the map, S and T refer to the first and second standard deviations of the examinee ability.

A second observation is that in the distribution of items, there are places where there are only a few items. Ideally, it would be nice to see more items in these places to help increase the reliability with which the items are separating the examinees. If we were reusing this test and had a bank of items to draw from, we would add items to the test at these difficulty levels. Thirdly, the person item map allows one to identify potential outliers, such as 3.2 which is at an extremely low level of difficulty compared with all other items.

Finally, looking very broadly at how items in specific sections performed, items at the higher end of the scale (50 and above) tend to be questions related to grammar and vocabulary, while the "conversation" items (3.1–3.10) and questions (e.g., items 5.14, 5.15, 5.16, 5.17) related to simple conjunctions (because, and, but, if) are at the lower end (50 and below). Reading comprehension questions are spread across the middle. This relative ranking in difficulty, with vocabulary and grammar questions more difficult than conversation items and reading comprehension questions comprising the middle range of difficulty is the same pattern that we observed in our analysis of our March 2003 exam (Churchill & Aline, 2005).

Looking specifically at Sections 4 and 5, several items related to the article system and vocabulary are among the more difficult items in the test and also seem to be matching the ability level (55) of or challenging the English and Spanish majors taking these sections. On the other hand, there are items in these later sections which are far below the

ability level of the examinees taking these sections. Exam writers may need to increase the difficulty of some items in the later two sections of the test so that they more closely match the abilities of the language majors.

While the person item map can provide a lot of information on the overall match of the test to the examinees, allows one to reflect on the design of the test, and provides a window to begin looking into the relative difficulty of items, it does not allow one to closely examine how specific items are performing. However, Rasch does give the researcher a bag of tools to look at specific test items and it is to this analysis that we now turn.

Correct Data Entry

One application of Rasch is to ensure that data has been entered correctly. This is an important first step when viewing the person item map. In our first run of the data, we found that item 5.2 was extremely high in difficulty level, around 98 on the scale on the person item map. Therefore, that was the first item that we looked at closely. We found that the correct answer was miss-entered on our answer key, although it was entered correctly on the answer key in the testing center. We then changed the answer and reran the data.

WH Questions

Another outlier, item 3.2—the easiest item according to the person map of items, also drew our attention. This item type requires the examinees to select the missing phrase from a short, four-turn conversation. As we learned from the analysis of our March test for this question type (Churchill & Aline, 2005), if selection of the answer only requires noticing a simple WH question adjacency pair, then the item will fall on the easy end of the scale. The examinees do not even need to read the first two turns in this conversation in order to select the correct answer. Moreover, the distractors fail, as they are too simple for this range of examinees. Furthermore, although they are possible

Item 3.2
A: Have you made any plans for your vacation?
B: Not yet. I'm still thinking about where to go.
A: (2)
B: No, but I've heard it's beautiful.

	# of Examinees	Average Measure
a. Did you enjoy Okinawa?	86 (2%)	44.20
d. Okinawa is a great place.	50 (1%)	44.58
c. Going to Okinawa might be good.	171 (4%)	47.63
b. Have you ever been to Okinawa?*	3,658 (92%)	51.17

Figure 4. "Conversation Items" That Were Too Easy

responses in some contexts, the distractors C and D are not questions (See Figure 4).

Conjuncts

One focus of our research is to hopefully find item types that are performing poorly as measures of our examinees so that we can exclude them from use on future tests. Items 5.14 through 5.18 are testing the language majors' knowledge of conjuncts (See Figure 5). As indicated on the person map of items (Figure 3), almost all of these items are rather easy, except for 5.18 with the correct answer of "hence." Keeping in mind that the person map of items includes all of the examinees and that only the language majors answered items in Section 5, these 5 items are even easier for the language majors than they appear on the map. Indeed, all of the correct answers for these items had a very high correct selection frequency of around 80%. Therefore, we could conclude that this type of item should not be used in future tests. But there are a couple of things to consider before excluding it: first, these sentences are short and do not require much thought on the part of the examinees. Moreover, the conjuncts are being tested as a set of five items with five answers so that if an examinee gets four answers correct, the last answer is automatically known. For a comparison we can look at item 3.18. This item's correct

(5.14) 1. It's past midnight, (1) the pubs are still open.

(5.15) 2. Peter got wet (2) it was raining.

(5.16) 3. Well, (3) you ask me, she's improved a lot.

(5.17) 4. I was 10 years old (4) the first man landed on the moon.

(5.18) 5. I had little time during the day, (5) I studied at night.
a. because b. but c. if d. hence e. when

(3.18) 8. Physical fitness exercises can cause injuries (8) the participants are not careful.

Data Code	# of Examinees	Average Measure
a. with	1,071 (27%)	48.81
d. is	689 (17%)	49.00
c. to	739 (19%)	49.14
b. if*	1.439 (37%)	53.92

Figure 5. Items Involving Conjuncts

answer is the conjunct "if," and it is a good item as demonstrated by its position on the person item map, its average measure, and its discrimination estimate of 1.23. Two differences between this item and the other conjunct items are that this item is not included in a set of similar conjunct items, and it does not have other conjuncts as distractors. Therefore, it would be premature to eliminate all items testing conjuncts, but we should consider eliminating items testing the same grammar form in sets, and have more distractors than questions in order to eliminate guessing to some extent.

Average Measure

The Rasch Model includes other statistics besides the person item map for analyzing test items. Because the Rasch Model places the examinees on an interval scale according to the probability of an examinee selecting an item correctly in relation to the other items an examinee has answered correctly, that is to say the ranking of the examinee abilities, we can find distractors in items that are selected by examinees who are at a higher ability level than the difficulty level of

the item. In other words, these examinees should be answering a particular item correctly as the item is easier than their ability level, but they are answering it incorrectly. Item 5.9 appears to be a good item on the person item map in that it seems to be performing well as its difficulty level is at a relatively good match with the ability level of the examinees. However, when we look at the average measure of examinees selecting different distractors, we can see that some examinees who answered correctly to items that are ranked as more difficult than this item, answered this item incorrectly (See Figure 6). The correct answer here is A, "sociable."

In Figure 6, we have the data code (multiple choice answers), count (number of examinees who selected that answer), percent of those selections, and the average measure, which is, according to the WINSTEPS manual, "the observed, sample-dependent, average measure of persons in this analysis who responded in this category" (Linacre, 2003). The manual goes on to note that the average measure is "a quality control statistic for this analysis." If the correct item receives an asterisk, that indicates that examinees with a higher average ability have answered this item incorrectly. Examinees with an average ability of 55.96 selected the correct answer, while examinees with a higher average ability of 56.88 selected distractor D, "sophisticated." In other words, higher ability examinees are choosing the wrong answer. The other distractors are clearly incorrect for examinees at this ability

Item 5.9

The Smiths are very (a. sociable* b. society c. solitary d. sophisticated) people, and they love to give parties every now and then.

Data Code	# of Examinees	Average Measure
B	50 (20%)	51.58
C	42 (17%)	54.41
D	41 (17%)	56.88
A*	115 (46%)	55.96

Estimated discrimination: 0.48

Figure 6. Example of a Vocabulary Item That is not Performing Well

level as "solitary" is not semantically related to "parties," and "society" is grammatically incorrect. Examinees at a higher ability level might be selecting "sophisticated" as they see "sociable" as being an answer that is too simple because it is very close in meaning to "parties," and they may be viewing "sophisticated" as including the meaning of "sociable" as both words can sometimes be semantically related through such words as "urbane" or "worldly." However, this explanation as to why examinees at a higher ability level than this item are selecting "sophisticated" is only conjecture. We would need to actually ask the examinees if we wanted to more completely understand their selection of D as the answer.

Another way we could look at item 5.9 is from the estimated discrimination. The Rasch model expects a discrimination of 1.0, so that values greater than 1.0 are over-discriminating and values less than 1.0 are under-discriminating. Item 5.9 has a discrimination value of .48, which indicates extreme under-discrimination, that is, it is not differentiating well between higher ability and lower ability examinees. It is quite clear why the discrimination is so low as examinees at a higher ability level are missing the correct answer on this lower difficulty item, while examinees answering it correctly have a lower average ability.

Misfit Order

Another statistic produced by Rasch that we can use to more closely examine items is the misfit order, which tells us about the mean-square infit and outfit of items. These measures show us to what extent the items are not fitting the Rasch model. According to the WINSTEPS manual, infit is "more sensitive to unexpected behavior affecting responses to items near the person's measure level," while outfit is "more sensitive to unexpected behavior by persons on items far from the person's measure level" (Linacre, 2003, p. 174). The Rasch manual further notes that items with values greater than 1.5 indicate "noise," which doesn't provide us with useful information about the

Item 1.4 Outfit = 1.30

Playing the flute gives me a chance to take the notes on the page—to (4)take what somebody else gave me—and make it something personal.

Data Code	# of Examinees	Average Measure
a. remove	180 (12%)	49.00
d. memorize	1,471 (37%)	50.96
c. make up	1,384 (35%)	51.01
b. use*	619 (16%)	51.24

Figure 7. Example of a Misfitting Vocabulary Item

test, and values less than 0.5 are overly predictable and trick us into believing that our test is measuring better than it actually is, (also called the Attenuation Paradox). All of the items on this exam fell within this acceptable range of 0.5 to 1.5, but we must think of this range as a guide rather than an exact rule.

On this exam, Item 1.4 had the highest outfit value at 1.30, while the next highest outfit value is 1.13. Item 1.4 is further indicated as an outfit item on the items fit graph. Continuing with the outfit data, Rasch produces a table of the most unexpected responses of the examinees. In this table is a list of the examinees who have unexpectedly answered this Item 1.4 correctly even though they had only a .02 or .04 probability of answering correctly. Checking these examinees' on the measure order for persons, we see that they have a lower ability as indicated on the person item map, ranging around 32–37.

The high outfit statistic for this item is a result of examinees with a low ability level, around 35, scoring correctly on an item with an ability level of about 68! As to why these lower ability examinees are answering correctly, we can only speculate. This is a vocabulary item contextualized within a 700 word reading passage. The examinees must select the correct synonymous lexical item to fit the underlined lexical item. It may be that the correct answer "use" is too simplistic for examinees to want to select it, but at the same time it causes this to be a difficult item because it does require the simple answer. The distractor

B could be a possible answer because musicians do memorize the notes, and then interpret them differently, but memorize complicates the meaning in a way that "take" and "use" do not. It could be that the lower ability examinees are selecting the answer "use" because it is the only vocabulary item that they understand. But again, this is as far as we can go with the statistical analyses as any further analysis requires some kind of feedback from the test takers themselves. In the future, we hope to extend our research into this area.

Conclusion

From our analyses of our university's February B form entrance exam, we make the following conclusions. First, the overall test is neither too easy nor too difficult and there is a very nice match between the test and the examinees. Second, the grammar and vocabulary questions tend to be more difficult, while the conversation and conjuncts are easier. In the future, we should ensure that the conversation questions do not require selection of a simple question/answer pair, and that conjuncts and other grammar forms are tested separately, not as a set, and that they have a greater number of answers than there are questions so as to decrease answering correctly by guessing. Third, distractors in vocabulary items should not be too closely related semantically, such as "sociable" and "sophisticated." Fourth, we should initially check to make sure that our data has been entered correctly and that our answer key is completely correct. Finally, our examination of the 2003 February B entrance exam is part of the action research cycle that we outlined at the beginning of this paper. Similarly, this report works towards sharing our findings with other exam writers and learning from our analysis. We hope that this analysis is informative to future test writers and that it contributes to improving the quality of our entrance exams within the constraints that they are being created.

* An earlier version of this paper was presented at the annual meeting of the Japan Testing Association in Hikarigaoka, Chiba, Japan in September of 2004. This research was partially funded by a research grant from the Center for Language Studies at Kanagawa University.

References

Aline, D., & Churchill, E. (2004, July). *Item analysis of entrance exam item types: An application of Rasch measurement.* Paper presented at a meeting of the Kanagawa University Comparative Linguistics Research Group, Kanagawa University, Yokohama, Japan

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* London: Lawrence Erlbaum Associates.

Brown, J. D., & Yamashita, S. O. (1995). English entrance exams at Japanese universities. What do we know about them? *JALT Journal,* 18(1), 7–30.

Churchill, E., & Aline, D. (2004, September). *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 types. *Kanagawa University Studies in Language,* 27, 85–105.

Linacre, J. M. (2003). *A User's Guide to WINSTEPS Ministep Rasch-Model Computer Programs.* Chicago: WINSTEPS.