Improving English Test Questions Through Rasch Analysis

OBERMEIER, Andrew

（京都教育大学）

Rasch 分析による英語リスニング問題の改善

オーバマイヤー・アンドリュー

（京都教育大学）

Received November 2008

Abstract: This research investigated whether the questions in the listening section of final examination of Kyoto University of Education's general English requirement course, *Grammar for Communication,* are at an appropriate difficulty level. Through a Rasch analysis, the difficulty of the listening questions with regard to the students' ability level and the difficulty other sections of the test was evaluated. The research concluded that the questions are of an appropriate difficulty level, and fit in well with the other questions on the test.

抄録：本研究では京都教育大学の共通科目「コミュニケーションのための英文法」の期末試験のリスニング問題の難易度を調べた。Rasch 分析によるとリスニング問題の難易度は相応しいとの結果が証明された。

キーワード：Rasch Analysis, Testing English as a Foreign Language,

Ⅰ．INTRODUCTION

Beginning with the 2006 academic year, as part of efforts to improve the required English curriculum, the English Department revised the contents of *Grammar for Communication* by unifying the syllabus and final examination for all eight sections of the course. The course's main purpose is to teach the verb tenses and how they are used in communication, with a text (Murphy, 2004) that emphasizes explanations and examples of how English is used in daily situations. Significantly, a listening section was added to the final examination, intending to urge teachers and students to focus on spoken English. This was quite an innovation, since grammar has traditionally been taught through explanations in Japanese, and tested through translation and multiple choice type tests. As the audio medium is a new and different way of testing grammar, it was worthwhile to determine whether the listening section fit in with the rest of the test. In this research, I investigated whether listening items were at an appropriate difficulty level, and whether they tested grammatical ability coherently with the rest of the test.

Ⅱ. THE LISTENING SECTION

The ten items that make up the Listening Section are part of the one-hour, sixty-item final examination for *Grammar for Communication.* Each section of the test, including the Listening Section, covers the verb forms, as explained in Murphy (2008). There are five sections: A) Multiple Choice; B) Fill in the Blanks; C) True-False Multiple Choice; D) Error Correction; and E) Listening. On the listening section, a fifteen-minute audio CD is played once. Each of the ten items has a question prompt and four corresponding answer choices, all of which are repeated twice. Students hear a question and then hear four choices to respond to it. Three of the choices are grammatically incorrect distracters, and one choice is the correct answer. An example of a listening item follows:

Figure 1:

Example Listening Item

(Students hear the following, twice.)

1. "Would you like to go to a movie tonight?"

a. "Yes, I do."

b. "Yes, I am."

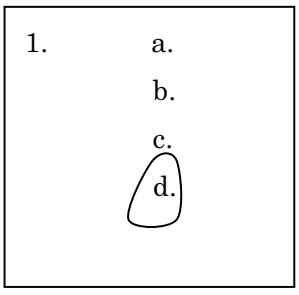
c. "Yes, I will."

d. *"Yes, I would."*

In the above item, "d" is the correct answer because it is the only one that is grammatically correct. One could argue that all of the answers above are communicatively correct because they all adequately convey that the speaker is accepting the invitation to go to a movie. But grammatically, only choice "d" is correct. The students' answer sheet has no text, clearly showing that the item can only be solved through listening:

Figure 2:

Example Answer



The implications of this item design are important. Since students have to hear and understand both the question prompt and the response answers, it is purely a listening item in the sense that there is no reading involved. Our hope is that this will urge teachers to teach grammar orally through drill work and active

oral questioning, and avoid explaining grammar in Japanese.

However, this format does have some problematic aspects. First, when responding to spoken language in real life, one must respond to the contents of what is said, only rarely needing to comment on grammatical correctness. Therefore, it could be said that the task is inauthentic. More concerning, the format makes superfluous demands on memory in the process of testing grammar. Students must remember all four distracters until they choose their answer. Remembering each distracter and simultaneously considering whether they are correct or not requires mental exertion peripheral to the abilities we want to test. Such mental capacity is beyond the scope of what we are trying to teach in the class, and requires extra work from the students. On a traditional written multiple-choice test, the student can carefully compare between two or three possible answers by rereading them, so such a format may more genuinely test grammar knowledge. To ameliorate this extra burden on memory, we make distracters that differ in only one grammatical point and by repeat each question twice. Also, many students take notes while listening to all of the questions, and then go back to choose their answer from their notes.

Since this new item type is an innovative change, and due our uncertainty about how students and teachers would respond to a listening section, the English Department introduced it with caution. Only twenty percent of the total test grade comes from this section. With the above issues in mind, I conducted this present study to find out how the listening items compare to the items on the rest of the test. My research questions were as follows:

1. Are the listening questions more difficult than those on the rest of the test?
2. Do the listening questions fit in with the construct of grammatical competence as measured by the test as a whole?

I investigated the above questions through Rasch Analysis, a statistical technique used in education and psychology to measure abstract constructs.

Ⅲ. RASCH ANALYSIS

Rasch measurement aims to provide social scientists with the means to produce genuine interval measures (Bond & Fox, 2007). By converting item scores to logarithms, and then calculating the odds of each student to answer each item correctly, the scores become meaningful calibrations of difficulty and ability. On this grammar test, for example, we can understand how the difficulty of each item compares with the others. We can also understand how the students' grammatical abilities, as measured by the items, compare with each other. The fundamental questions that drive Rasch Analysis apply to each item and each student:

1. How difficult or how easy is this item?
2. How high or low is this student's ability?

Rasch Analysis is used by psychological and educational testers to turn abstract constructs like item difficulty and student ability into concrete measures. Researchers use one of a number of Rasch software

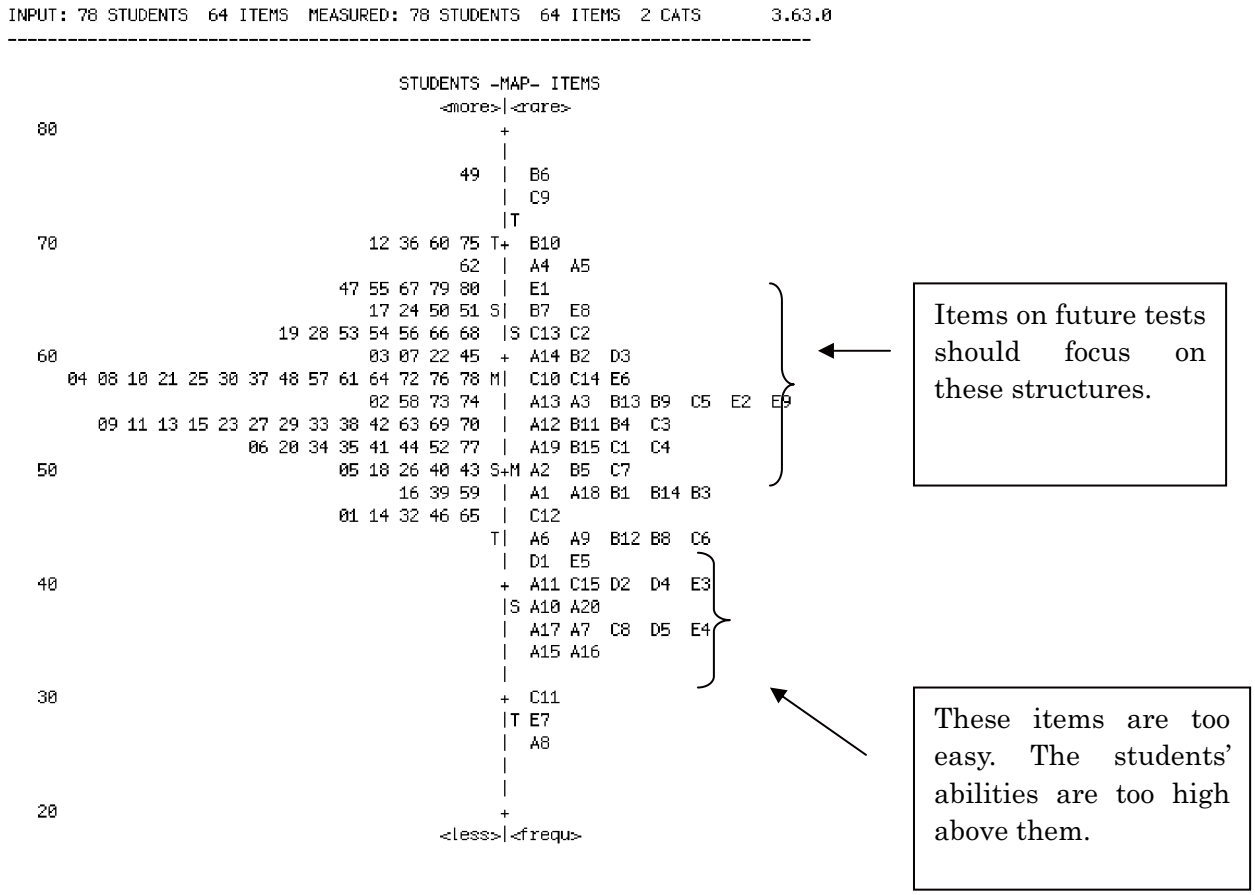
packages, and for this analysis I used *Winsteps* (Linacre, 1995). Proponents of the model strive to make measures that are as consistent and reliable as those in the physical sciences. They often make an analogy with a thermometer to emphasize the assertion that measuring a concept like students' grammatical ability should be as consistent and reliable as measuring temperature. In the same way that temperature is a construct that is measured by the reliable thermometer in incremental units such as Celsius degrees, grammatical ability should be likewise measurable in incremental units. The issue is whether the test is a reliable instrument, which Rasch modeling helps researchers to determine.

A Rasch analysis of a test places all the items onto an incremental scale of difficulty. Student ability is measured in terms of test item difficulty. One very illuminating aspect of the model is that the test items and students can be mapped on a chart according to these ability and difficulty estimates.

The ability and difficulty map of the final test in *Grammar for Communication* is shown below in Figure 3.

Figure 3

Ability/Difficulty Map of Students and Items



The 80 students in this study are mapped on the left side of the map under the heading "Students." The most able student is number 49, and five students, 01, 14, 32, 46, and 65, are tied as the least able. These students' ability level corresponds to the item difficulty level. The high ability of student number 49 matches the difficulty level of item number B6. The low ability of the aforementioned five students matches the relatively easy item number C12.

On the right side of the map, the test items are represented by letter for each section. Letter A represents the multiple-choice section, so A1 is the first item in that section. Similarly, letter E represents the listening section. As it happens, the first problem in the listening section, E1, is also the most difficult. We know this because it has the highest Item Difficulty and is thus charted toward the top of the map. The transcript of Item E1 follows in Figure 4:

Figure 4

The Most Difficult Listening Item

Item E1:

"How was your dentist appointment?"

a. *"Not so bad, it might be worse."*

b. *"Not so bad. It should have been worse."*

c. *"Not so bad. It could have been worse.*

d. *"Not so bad. It ought to be worse."*

The correct answer is "c", but the distracters are all only subtly different. Students have to know the different uses of the words "could," "might," "should," and "ought." Also, being the first item in the section, some students may have missed it because they were getting used to the test format. On future tests, to relieve student anxiety, we should make sure that the first item in this section is easier. Nevertheless, the item is an appropriate difficulty level for at least 11 of the 80 students. Looking at Figure 3, we can see that there are five students (47, 55, 67, 79, 80) whose ability level corresponded to this difficulty level. That means that these five students had a 50% chance of getting item E1 correct. In addition, the students mapped above the level of E1 (12, 36, 49, 60, 62, 75) all probably got E1 correct, because their ability level is higher than the difficulty level of E1. Since E1 was the most difficult question on the test, and was at an appropriate level for many of the students, we can conclude that none of the listening items were too difficult.

Investigating the middle of the range of the map in Figure 3, we find E8, E6, E2, and E9, items that blend in with many items from the other sections of the test. These items are at difficulty levels that correspond with the majority of the students, meaning they are perfectly appropriate, neither too difficult nor two easy. Item E6 lies horizontal to the symbol M, meaning that its difficulty is perfectly in the middle of the range, at the mean ability level. In future revisions of the test, items like these should be included

on the test.

Four of the problems in the section, E5, E3, E4, and E7, were too easy to test the students' ability. We can see this visually on the table because there are no students at ability levels that correspond to the difficulties of these items. Since all the students are charted above these items, we can assume that they all got these items correct. An example of such an easy item follows:

Figure 4

The Easiest Listening Item

Item E7:

"I hear you are planning to go on holiday soon. Where are you going?"

a. "I went to the sea."

b. "I go to the sea."

c. "I will go to the sea."

d. *"I am going to the sea."*

This item is clearly more straightforward than the difficult item in **Figure 3.** Distracter choice "a" responds to the question about the future with an answer about the past. Distracter "b" similarly mistakes the tense, answering about the habitual present. The only difficulty in this problem lies in choices "c" and "d" which are both about the future. Choice "d" is correct because the Present Continuous is used to talk about the planned future. Choice "c" is wrong because the Future tense is only used at the point of making a decision, not in describing planned future events like the question refers to. Therefore, only two of the four distracters on this item caused any difficulty. Further, since the Present Continuous is used in the prompt, many students probably correctly reasoned that the Present Continuous would also be used in the answer.

Although these final four easy items probably did not provide a sufficient challenge for the ability level of the students who took the test, it is worthwhile to include a few similar items on the test, and preferably as one of the first items in the section. Such items serve as an easy warm up, giving students confidence, and ameliorating test anxiety. The abundance of items from the first section of the test in this "too easy" range (A6, A9, A11, A10, A17, A15, A16, A8) are similarly justified because they serve this "warm-up" function to alleviate test taker anxiety and help students to bring their full ability to the test without being distracted by confusing items when they begin the test.

Ⅳ. FIT ANALYSIS

The above discussion assumes that the test as a whole is functioning properly. In this section, I will explain the results of an analysis conducted to determine whether the data from the test fit the Rasch

statistical model. That is to say, the test should measure one construct.

The Infit and Outfit statistics were 1.00 and .96 respectively, showing that there was little measurement error. The items were measuring the same construct, in this case, grammatical ability. The Category Function Measure showed that students who got answers correct were on average 12.48 logits above the items, while those who got items wrong were .98 logits below. This means that the easier items were *very* easy, but the items that were difficult were only *slightly* difficult. This can be confirmed visually on the map in Figure 3, very few of the items are far above the student ability levels, but many of the items are far below the student ability levels. The coherence estimate was .68 for incorrect answers and .77 for correct answers, so the model predicted slightly more accurately for correct answers. The Item Misfit Analysis showed that all of the items productively constructed measurement, since the Outfit Mean squares ranged within the recommended zone, between .82 and 1.21. In all, the fit analysis showed that the data was suitable to be measured by Rasch Analysis.

Ⅴ. CONCLUSION

This study sought to find out whether the listening questions on the final test of *Grammar for Communication* were appropriately difficult, and concluded that if anything they were too easy. The study also concluded positively that the listening items fit in with the rest of the items on the test through a fit analysis. Rasch analysis is a useful tool for analyzing tests, allowing a researcher to understand how the test items function to measure the students' ability.

REFERENCES

Bond, T. & Fox, C. (2007). *Applying the Rasch Model.* Mahwah, New Jersey: Lawrence Earlbaum

Brown, J. D. (1995). *The Elements of Language Curriculum.* Boston: Newbury House.

Brown, J. D. & Hudson, T. (2002). *Criterion-referenced Language Testing.*

Johnson, R. K. (1989). *The Second Language Curriculum.* Cambridge: Cambridge University Press.

Linacre, M. (1995). *Winsteps.* Software currently available online at www.winsteps.com.

Murphy, R. (2004). *English Grammar in Use.* Cambridge: Cambridge University Press.