論　文 ニダバ (*Nidaba*) 41(2012), 127-136

The Order of Grammar Item Difficulty

—A Preliminary Analysis —

Atsuko NISHITANI

Introduction

Communicative language teaching has been popular for approximately three decades, and the role of grammar teaching in the L2 communicative curriculum has been questioned (Purpura, 2004). However, Japanese teachers of English are still teaching grammar in secondary school classrooms to meet students' immediate needs to pass entrance examinations or to get high scores on proficiency tests such as the Test of English for International Communication (TOEIC). Even though high-stakes tests in Japan continue to include tests of learners' grammatical knowledge, the relative difficulty of different grammatical structures has not been investigated or established yet (DeKeyser, 2005; Ellis, 2006). As DeKeyser (2005) pointed out, few researchers have compared the difficulty of a wide range of grammatical structures. Previous researchers have been generally focused on single grammatical categories; thus, the interrelationships among different grammatical structures have been discussed relatively little. For example, although researchers have investigated the acquisition order of certain morphemes, the developmental sequence of negatives, the developmental sequence of interrogatives, and the difficulty order of relative clauses, few researchers have investigated which negative form is acquired before which interrogative form and vice versa. In other words, language testers tend to write grammar items based on their intuition or general and vague perception of difficulty of different grammatical structures, which is obviously unsatisfactory (Ellis, 2001). Thus this study was conducted in an attempt to answer the following research question: What is a hierarchy of grammatical difficulty for Japanese learners of English?

Participants

A total of 854 Japanese university students, 435 male and 419 female students, participated in this study. Of the 854 students, 460 were first-year students, 323 were second-year students, 48 were third-year students, and 23 were fourth-year students. Thus, the

majority of them presumably remembered the English grammar they had studied in junior and senior high school for six years reasonably well. Their English proficiency was mixed, ranging from the false beginners to the advanced proficiency learners (TOEIC 290-870).

Instrumentation

Five sets of tests and a total of 12 test forms were administered in this study. All of the tests used a multiple-choice format with a correct answer and three distractors. The following is a sample item:

This new plan is supported by young mothers as well as by the ( ) opinions of experts.
(a) object (b) objects (c) objectively (d) objective

All items were scored dichotomously. The participants were familiar with this test format, as high-stakes examinations, such as Japanese secondary school and university entrance examinations and the TOEIC test, use the same multiple-choice format.

Test 1

Prior to this study, a pilot study was conducted (Nishitani, 2011). After 278 grammar items from TOEFL and TOEIC practice tests taken by 1,409 university students were analyzed using a Rasch analysis, pairs of items that were testing the same grammar point in a similar way (i.e., having a blank in the same position in the sentence) and had similar difficulty estimates were extracted. Average difficulty estimates of each pair were calculated, and the difficulty order of 21 grammatical structures was tentatively established. Therefore, Test 1 was designed to determine if the order obtained in the pilot study would be replicated with similar items: Two parallel items for each of the 21 structures, a total of 42 items, were adapted from items originally found in TOEFL and TOEIC practice tests.

The items were revised to predominantly include high frequency vocabulary so that they would measure the participants' grammatical knowledge and their ability to identify the correct answer would not be influenced by their lexical knowledge. The vocabulary used was within the first high frequency 3,000 words of English, which Japanese students are supposed to learn before graduating from high school. An English-Japanese dictionary, *Genius* (Konishi & Minamide, 2006), was used for checking the frequency level of the words on the test, as it is frequently used as a reference when making university entrance examinations.

Test 1 had two parallel forms that were made up of items testing the same grammatical structures; however, different lexis was used in the sentences testing the same structure. The assumption was that the two sentences would have the same or similar Rasch difficulty estimates when administered to different participants. Each form consisted of 30 items: the 21

items mentioned above (i.e., the structures selected from the pilot study), and nine items from the other form (i.e., nine items from Form A were included on Form B, and vice versa), which were used as anchor items. In other words, out of 42 items written for Test 1, 18 items were taken by all 265 participants.

Test 2

Out of the 21 pairs of items, only six pairs showed similar difficulty estimates; thus, the items designed to measure 15 grammatical structures were revised to be more similar.

Test 2 had two forms as well and each form was made up of 15 items. Of the 30 items, six items were also on Test 1, so they were used as anchor items when combining the data from Tests 1 and 2.

Test 3

Of the 15 grammatical structures included on Test 2, six pairs did not show similar difficulty estimates despite having been revised to be more similar. Therefore, the six grammatical structures were retested, this time with three items testing each of the six grammatical structures: Of the 18 items, three items were the same items as on Test 2, nine items were revised, and six items were newly written.

Test 4

An additional 17 structures that did not show consistent difficulty estimates in the pilot study were selected for inclusion on Test 4. Three items were written for each structure in the hope that at least two items (if not all three) designed to measure each structure would show similar difficulty estimates. Each of the three forms was made up of 30 items, 13 of which were anchor items from the previous tests.

Test 5

After the data from Tests 1 through 4 were combined and analyzed, 11 grammatical structures from the tests did not display consistent difficulty estimates; thus, they were revised and retested. Two forms were made, each of which was made up of 22 items: 11 revised items and 11 anchor items from the previous tests. Of the 11 anchor items on each form, six were common to both forms but the other five differed from each other. In other words, Test 5 included a total of 16 anchor items. Most (but not all) pairs showed similar difficulty estimates; thus, no more tests were administered for the purpose of measuring the difficulty estimates of different grammatical structures.

Procedures

Test 1 was administered at three universities in Kyoto and Nagoya. A total of 265 participants took either Form A (n = 129) or Form B (n = 124) of the test. Twelve participants, who happened to be in two classes where the test was administered, took both forms. The participants were given 10-15 minutes to complete the test.

The data were combined using 18 anchor items to place the items from the different test forms onto a common scale. Note that the 12 participants who took Form A and Form B answered the anchor items twice. If they answered differently on the two test forms, those answers were excluded from the analysis.

The Rasch analysis was conducted using the WINSTEPS computer software. First, the infit mean-square statistics for each item were examined to determine whether they were in the range of the mean ± twice the standard deviation of the mean square statistic. According to McNamara (1996), "the infit statistics ... are the ones usually considered the most informative, as they focus on the degree of fit in the most typical observations in the matrix" (p. 172), and "for n sizes of 30 or more, the [acceptable] range is the mean ± twice the standard deviation of the mean square statistic" (p. 181).

Next, the difficulty estimates for each pair of items designed to measure the same grammar point were examined to determine whether they had similar estimates. According to Linacre (2010), "when we want to say 'Item A is definitely more difficult than Item B'... their measures need to be more than 3 S.E.s [standard errors] different." In this study, however, a 2 S.E. difference was used as the criterion to ensure that the two items had similar difficulty estimates. The pairs that did not meet the 2 S.E. criterion (i.e., 15 pairs) were revised and tested once again in Test 2.

Test 2 was administered at one university in Osaka. The participants (n = 55) took both forms of the test with a two-week interval separating the administration of the first and second test form. To avoid the possibility of an order effect, one class was given Form A first, and the other class was given Form B first. Because some students were absent either the first week or the second week, the number of the students who took both forms, only Form A, and only Form B was 48, 2, and 5 respectively. The test takers were given 5 to 8 minutes to finish the test. The test results were analyzed with the dichotmous Rasch model and the Rasch difficulty estimates of the items within each pair were examined to determine whether the difficulty estimates were within 2 S.E.s of one another. The pairs that had difficulty estimates greater than the 2 S.E. criterion (i.e., six pairs) were revised and retested in Test 3.

Test 3 (6 items x 3 forms) was administered at one university in Kyoto. The participants (n = 51) took all three test forms at one-week intervals and were given 2-3 minutes to finish the

test. To avoid an order effect, the order of the forms (Forms A, B, and C) given was randomized. Thus 16 students took the tests in the order of ABC, 9 students in the order of BCA, 12 students in the order of CAB, 8 students in the order of CBA, and 6 students in the order of BAC. Out of the 51 participants, 48 had taken Test 1 also; thus, the data gathered from those students were used as anchors when combining the data from Tests 1, 2, and 3. After obtaining the results from Test 3, the combined data of Tests 1, 2, and 3 were analyzed using the dichotomous Rasch model. The difficulty estimates of the items designed to measure the same grammatical structure were examined to determine whether they were sufficiently similar. They all were found to have similar difficulty estimates; thus, further revisions were unnecessary.

Test 4 (30 items x 3 forms) was administered at three universities in Osaka, Kobe, and Kyoto. The participants (n = 242) were given 10 to 15 minutes to finish the test. The number of participants who took Forms A, B, and C was 82, 79, and 81, respectively. The data from the three forms were combined using 13 anchor items. Again, after obtaining the data from the Test 4 administration, a Rasch analysis was conducted on the combined data of Tests 1, 2, 3, and 4. The item difficulty estimates of the items designed to measure the same grammatical structure were examined to determine whether they were sufficiently similar; those items with difficulty estimates that did not meet the 2 S.E. criterion (i.e., 11 pairs) were revised and retested in Test 5.

Test 5 (22 items x 2 forms) was administered at three universities in Osaka and Kobe. The participants (n = 289) were given 7-11 minutes to complete the test. The number of participants who took Form D and Form E was 131 and 158, respectively. The data were combined using 16 anchor items. The Rasch analysis was conducted not only on the data from Test 5, but also on the combined data of Tests 1, 2, 3, 4, and 5. The item difficulty estimates of the items designed to measure the same grammatical structure were examined to determine whether they showed similar difficulty estimates.

After inspecting at the results of the Rasch analysis for Tests 1 through 5, two items with similar difficulty estimates were selected for each grammatical structure. The mean estimates of each pair were calculated and compared to examine the difficulty order of 38 grammatical structures.

Results and Discussion

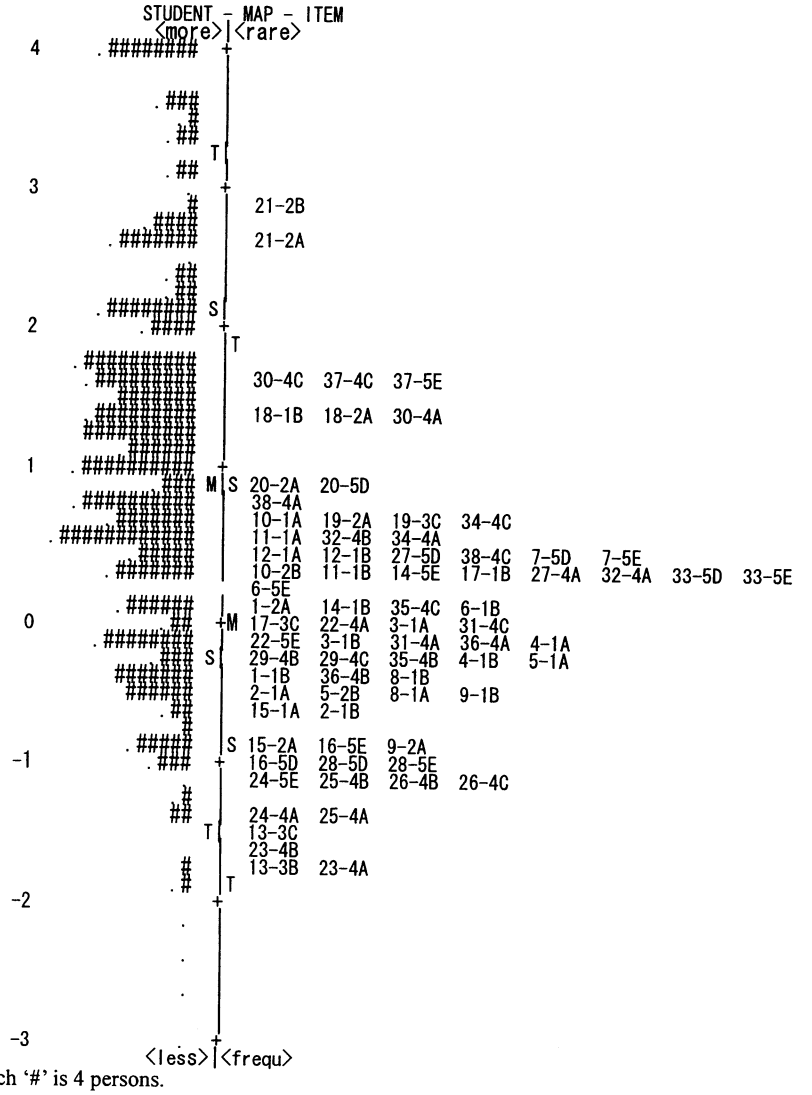
A Rasch measurement analysis was conducted using the WINSTEPS computer program. Tables 1 and 2 show the estimates and fit statistics of the students and the items.

Table 1. *Estimates and fit statistics of the students*

	score	measure	error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
M	10.1	1.05	.72	1.00	.1	.98	.1
SD	3.9	1.30	.24	.24	.8	.44	.8
SD (adjusted)		1.06					
Separation		1.39					
Reliability		.66					

Table 2. *Estimates and fit statistics of the items*

	score	measure	error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
M	113.4	.00	.23	1.00	.1	.99	.0
SD	64.8	.91	.10	.11	1.2	.24	1.3
SD (adjusted)		.88					
Separation		3.49					
Reliability		.92					



Note. Each '#' is 4 persons.

Figure 1. *Distribution of Student Ability Estimates and Item Difficulty Estimates*

Table 3. *Means of Difficulty Estimates of 38 Grammar Items*

Grammar item	Mean
Subjunctive	2.76
Past perfect	1.61
Relative clause "which" (vs. "where")	1.50
Adjective (after "be" + Adverb)	1.38
Present tense	0.90
Adjective (after "be")	0.60
Adjective (after sensory verb)	0.60
Subject-Verb agreement	0.57
Adverb (between Subject and Verb)	0.42
Conjunction (+meaning)	0.38
Conjunction (vs. Preposition)	0.38
Passive	0.37
Future tense	0.36
Noun (after Verb)	0.28
Causative "–ed"	0.26
Adverb (between "be" and Adjective)	0.23
Adverb (between "be" and Past participle)	0.20
Past tense (with another past verb in the sentence)	0.13
Preposition (vs. Conjunction)	-0.01
Subject pronoun	-0.03
"ing" (after Noun)	-0.07
Past participle (after "be")	-0.10
Noun (between Article and Preposition)	-0.13
Possessive pronoun	-0.16
"ed" (after Noun)	-0.27
Relative clause "that"	-0.28
Adjective (between Article and Noun)	-0.38
Base form (after "in order to")	-0.44
"oneself" (after "by")	-0.57
"To + Verb" (after Verb + Noun)	-0.71
Noun (after Verb + Adjective)	-0.73
"ing" (after Preposition)	-0.96
Verb (between Subject and Object)	-1.04
Relative clause "who"	-1.15
Object pronoun (after Verb + "to")	-1.26
Adverb (after Verb + Noun)	-1.27
Past tense (with a past-tense keyword in the sentence)	-1.65
Present perfect (with "since" in the sentence)	-1.71

The separation index and the reliability index of person ability were low – 1.39 and .66 respectively, which means the students can be divided into about 1.4 levels of ability and the items did not distinguish between the students. However, this does not concern me because the focus of this study is not on the student's ability but the item difficulty.

The separation index and the reliability index of item difficulty were much higher – 3.49 and .92 respectively, which means the items can be divided into nearly 3.5 levels of difficulty

and item difficulty would remain quite stable even if the tests were given to other groups of people. It infers that "we have developed a line of inquiry in which some items are more difficult and some items are easier, and that we can place confidence in the consistency of these inferences" (Bond & Fox, 2001, p. 32).

Figure 1 is an item-ability map, which shows the distribution of student ability estimates relative to item difficulty estimates. The mean of item difficulty is set at zero, and the greater the value the higher the ability of the student and difficulty level of an item. This item-ability map visually shows that there was a ceiling effect – there were not enough items to distinguish higher-level students. It also shows that there are gaps along the item difficulty hierarchy, which means that items to distinguish the students at those levels more precisely are needed. But again the focus of this study is to investigate the difficulty of different grammar points and not to develop a test that can distinguish students at all levels. Thus the lack of such items is not considered as a problem here.

Table 3 shows the 38 types of grammar items that are arranged according to the difficulty estimates. The most difficult item was subjunctive (2.76) and the easiest one was present perfect tense with a keyword "since" in the sentence (-1.71). There are items that are testing the same parts of speech but with different sentence positions, and different sentence positions seem to have effects on item difficulty. Placing an adjective after a be-verb and an adverb was the most difficult (1.38), followed by immediately after a be-verb (0.60), after a sensory verb (0.60), and the easiest was placing one between an article and a noun (-0.38). Placing an adverb between a subject and a verb was the most difficult (0.42), followed by between a be-verb and an adjective (0.23), between a be-verb and a past participle (0.20), and the easiest was after verb and a noun (-1.27). Placing a noun after a verb was more difficult (0.28) than between an article and a preposition (-0.13), the easiest was after a verb and an adjective (-0.73). Pronoun items seem rather easy. Among them, choosing a subject pronoun was more difficult (-0.03) than a possessive pronoun (-0.16), a reflexive pronoun after "by" (-0.57), and the easiest was an object pronoun after "to" (-1.26). As for tense items, choosing past perfect was the most difficult (1.61), followed by present tense (0.90), future tense (0.36), and past tense with another past-tense verb in the sentence (0.13), while past tense with a keyword such as "yesterday" (-1.65) and present perfect with "since" in the sentence (-1.71) were much easier. As for relative pronoun items, choosing "which" over "where" as in "She will live in Kyoto ( ) is one of the oldest cities in Japan." was much more difficult (1.50) than placing "that" in front of a verb (-0.28) and "who" in front of a verb (-1.15).

Conclusion

In this study 38 pairs of items were designed to test the same grammar point in a similar way (i.e., the same sentence positions) and have similar difficulty estimates. They were placed on a linear scale of difficulty using the Rasch model, and their difficulty estimates were compared. The most difficult item was subjunctive (2.76) and the easiest one was present perfect tense with a keyword "since" in the sentence (-1.71). As with the pilot study, this study showed that the sentence position seems to have an effect on item difficulty. For example, placing an adjective after a be-verb and an adverb was far more difficult (1.38) than placing one between an article and a noun (-0.38).

For future research, adding more items of different grammar points to investigate their difficulty estimates would be beneficial. Knowledge of the order of grammatical difficulty can be used to select grammar items for tests, identify a learner's current stage of development, and plan more efficient curriculums. I hope that this study provides some implications for decision making in educational institutions.

Acknowledgement

This work was partially supported by KAKENHI (Grant-in-Aid for Scientific Research (C) 22520596).

References

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Langauge Learning, 55,* 1-25.

Ellis, R. (2001). Some thoughts on testing grammar: an SLA perspective. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 251-263). Cambridge: Cambridge University Press.

Ellis, R. (2006). Modelling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge. *Applied Linguistics, 27,* 431-463.

Konishi, T., & Minamide, K. (Eds.). (2006). *Genius English-Japanese dictionary.* Tokyo: Taishukan.

Linacre, J. M. (2010). Winsteps Help for Rasch Analysis. Retrieved from http://www.winsteps.com/winman

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

Nishitani, A. (2011). A Rasch measurement analysis of grammar item difficulty. In D. Meyer & D. Beglar (Eds.), *Proceedings of the 2010 Applied Linguistics Colloquium* (pp.30–36). Tokyo: Temple University Japan.

Purpura, J. E. (2004). *Assessing grammar.* Cambridge: Cambridge Uniresity Press.