Maintaining Validity and Reliability in Test Item-Reduction

ROBSON, Graham G.

I. Introduction

Literature Review

Placements tests in language learning seek to group "together students of similar ability levels" (Brown, 1996, p. 11) in a specific program or institute. Placement tests would try to determine where students have similar skills, say, listening, grammar or reading comprehension, and then students of similar abilities would be grouped together for maximum teaching efficiency. Without good placement tests to stream individuals into ability groups, there is a danger of having too wide a range of abilities in one class. Not only can such classes be extremely difficult for teachers to facilitate, but also some of the higher ability students may be in danger of being taught a version of skills that equates to an average level of proficiencies within one class. Therefore, making good placement tests that allow correct placement decisions is important to channel the skills of the students and allow the teacher to maximize materials for a more homogeneous level group.

In order for placement tests to be called good they should pay attention to how they can separate students' abilities, and should have high levels of validity and reliability. Furthermore, the level of test-takers themselves should be taken into consideration. In order to channel skills of the students, the range of questions on a test should be such that the competent students can answer more questions correctly, and the less competent students answer less of the questions correctly. The shape of scores should resemble a normal distribution, spreading the students out on a continuum of abilities, with fewer students at the outer ranges of the scores, and the majority of scores placed around the mean. Assuming that there is a sufficiently differing range of students' proficiencies, when students display their true proficiencies of a given construct in this way, teachers' ability to make placement decisions becomes easier. Next, validity deals with the need to justify how we measure the particular area of construct that we are interested in, so our validity score should reflect the "extent to which we can interpret a given test score as an indicator of ability...we want to measure" (Bachman & Palmer, 1996 p. 21). That means that if we issue a test of grammar that the items on the test should be grammar questions, and not measuring another construct. Just as important as validity is reliability, which we can describe hypothetically as the ability of a test to show consistency among scores if that test was issued to the same group at different times. In reality, for a placement test, reliability should reflect the consistent ability to order students from low to high levels of proficiency in a test. Finally, the test takers themselves should be considered. By this I mean that the test instrument should match the abilities of the students themselves. For example, if we administered a very difficult grammar test to low-proficiency students, we would not expect anybody to score high. Conversely, if we gave a test that was too easy to a group, we would get ceiling effects, where true abilities could not be measured because of the ease of a test, i.e. all the test takers score maximum points. Tests should endeavor to be of sufficient difficulty so that no-one scores either zero or the maximum score, and that everyone lies on a continuum of scores between the two, with most scores concentrated around the average. In summary, a placement test should seek to spread out the abilities of the test-takers, while maintaining high levels of validity and reliability, and keeping a close match between the test-takers ability and the difficulty of the test.

Traditionally in Japan, it is believed that a longer test will display all the necessary qualities of a good test mentioned above. Tests with a large number of items have a certain face-validity. In other words people can be more confident that a longer test will measure a construct better than a shorter test. Meaning that the larger the maximum point score is, the more believable the scores become as we compare them to the maximum score. This is a product of reduction in the amount of error that a larger number of items has over a smaller number, but it stands to reason that when a longer test is devoid of items that create error and redundancy, it becomes a shorter, but, potentially, better test. Therefore, instead of merely producing a long test to measure a construct, test-makers need to find some kind of cut-off point for the number of items, making tests not longer, but more efficient. Above this cut-off point more items in the test would produce redundancy, rather than a better test. Related to the perceived increased face-validity of longer tests, the answer to where to find this cut-off point has not been researched much in second language testing. One article by Bell and Lumsden (1980), concluded that all tests could be reduced by more than 60% without appreciable decreases in validity, but the tests used in that study were 100-item tests, much shorter than the tests used for this study. On the other hand Burton (2005), looked briefly at the sparse number of articles related to the length of tests. He concluded that "60 items are too few for a adequate sampling of a large knowledge domain". His argument is based upon shorter tests having problems with sampling effects and students who possibly guess. Beyond these two articles little has been mentioned on the topic in ESL/EFL academia.

Purpose

This paper analyzes a part of a placement test concentrating on grammar items that was used on first years at a Japanese university, and attempts to make the test more efficient by reducing the number of items on the test by use of various statistical methods. In other words, part of the test would be shortened to measure varying abilities of grammatical-based knowledge to help make placement decisions. A revised test based on the first placement test was then given to another group of first-year students at the same university as a measure of proficiency. This paper seeks to answer the following question:

1) Can the grammar part of the placement test be shortened and made more efficient, while maintaining acceptable levels of score dispersion, validity and reliability?

II. Method

Participants

The participants in this study were males and females from the total first year population of 231, who joined a tourism department at university in Japan for the new semester that started in April of 2004. These students included a small number of overseas students from China, but mostly consisted of students who had recently graduated from high schools in Japan. No other information about any of the students was available as background information for this study. The second population who used a revised version of the first placement test were first years that joined in the academic year starting 2005. The sample population used in this group was 105 students from four in tact groups (from a total of nine groups separated by a placement test. The students in the study were in groups two to five) of the English program, assumed to be of similar proficiencies to the previous year's group, in order to pilot the revised version of the placement test to see if it can be used for future placement decisions.

Procedure

The placement test used in this study was part of a TOEIC test. Rather than being a full version of the test that would lead to a full score of 990 points, the test was a compilation of items taken from a TOEIC preparation book (Lougheed, 1999). When the test was compiled, sometime shortly before April 2004, the test compiler chose three sections to the test as follows: The first section was reading comprehension, and involved a passage of about 260 words with ten comprehension items and three possible answers to choose from; the second section was listening broken into two parts. In the first part was 15 items that consisted of questions that the students listened to, and chose a suitable response from three choices that, again, they listened to. In the second part of the listening the students were required to read five questions on the test sheet and listen to a short dialogue. Next, they had to choose the answer that best fitted the answer to the original question. Finally, the third part was a grammar section consisting of 40 items that were a combination of single sentences and two-turn dialogues, with students choosing the best answer out of a choice of four to go into a missing space in a sentence. After analysis of the initial grammar section, a revised version of the original test was used to establish proficiency of a group of first years that joined the university one year later. This paper has been divided into two sections. In the first section I will look at how the first placement test was analyzed and made shorter, and then in the second section I will show a comparison of the first and revised tests and discover whether the dispersal of scores, the reliability and validity of the test was maintained for both tests.

III. Results and Discussion

Revision of First Test

Two types of statistical analysis were used to shorten the first placement test. The first type was incorporated in Winsteps software that used the Rasch model, a model of fit of test-taker and test difficulty, in order to show which items were not working very well at producing a wide range of individual ability in the grammar construct test. Once the redundant items were identified, statistics through excel were used to identify yet more redundancy of items and to determine how well the distractor items, or answers other than the correct answer were able to attract the test-takers into choosing the correct or incorrect answer.

Table 2 in Appendix A is based on the Rasch model that used the dichotomous results of the test, whether an answer was scored correctly, or not. This data was used to construct a developmental pathway based upon the relationship between the items and the students that took the test. The further the item is at the top the more difficult the item was. In this case Item 14 was the most difficult question overall on the test, and conversely, Item 6, at the bottom of the chart, was the easiest.

The figures down the side indicate a calculation of how much more difficult one item is compared to the other items on the test. Our pathway chart showed that the items have spaced out and indeed if you look to the left at the sharp marks, which represent the students in the pathway, you can see an indication of a normal distribution. The students IDs were too long for the information to be shown it its entirety, and that information was not necessary for this paper, but it does show a bell curve shape. However, even though we have a normal distribution the student separation value of 1.34 from Winsteps was quite low perhaps telling us that the students might have been in a range of similar abilities, so that no matter what test of grammar they were given, a narrow range of scores would have ensued.

In order for the pathway to be perfect there would have needed to be an item at every point along the vertical line, but the construction of an item to fit in, say, in between Item 15 and Item 37 would be both difficult and subjective. Working with the data we have, we can now begin to shorten the test and remove extraneous items. This was done by selecting items that were found on the same developmental difficulty line as others. One example could be Item 1, Item 21 or Items 25, all of which would not be necessary in the new version of a test. Two of the items that produce the same level of difficulty could be removed, leaving just one item, while maintaining the reliability of the test. The choice of which items to remove on the same line should be a matter of keeping a balance of different types of grammar questions.

After removing by far the easiest questions, Items 17 and 6, I chose 27 items that should remain in the new test that would form a line of developmental abilities in the students who took the test. These have been indicated by bold on the chart. Once the item fit was established, it was time to use excel statistics to again show which items were not separating well, as well as, which distractors were not functioning properly. These changes can be seen in Table 2 in Appendix B, and include removal of an item altogether if the Item Discrimination (ID) was too low, meaning that the item was not separating the top (77 highest-scoring students) from the bottom (77 lowest-scoring students) ability test-takers. I used Ebel's (1979, p. 276) guideline for rejection of items. Those guidelines stated that items with an ID of less than .19 are candidates for complete revision or rejection. In Table 2 the four items surrounded by boxes were rejected because of their unacceptable IDs.

Next, by dividing the test-takers into three groups of 77 each, I created a high, medium and low scoring group for each item. It was then possible to view the ID and how each distractor was working for each question with a certain ability group. The distractor could be either made more of an attractive option to test-takers, which is indicated by a arrow pointing up on the right of the distractor, or a distractor could be made less of an attractive option, thereby making the correct answer more obvious, indicated by an arrow to the right of a distractor pointing down. In other cases the three distractors and the correct answer were attracting the right amount of people in the correct proportions (above 0.7 is a standard figure commonly agreed on) and didn't need changing. It is worth noting that the number of lower IDs also mirrored the results of the student separation score in Winsteps.

That is because the students possessed a similar range of grammatical knowledge.

In summary, despite similarities between the test-takers as indicated by the low separation and IDs, it was possible to create a normal distribution necessary to separate students into groups of similar abilities for placement into the English program. Using statistics some redundant items that measured at a similar level of difficulty were removed, shortening the original test from 40 items to 27 items. Further statistics were able to pinpoint which items were not separating the lower from the middle and upper scoring students in the test. Also, within the items, the distractors were analyzed to see which were attracting the right level students and which were not. Both items and distractors were either removed or revised. The first revision of 27 items was brought down to 23 items in the final version. After the revisions were put into place a new version of the grammar test part of the placement test was finalized for future placement decisions.

Comparison of First Test and Revised Test

After revising the first placement test from 40 to 23 items, the revised form was given to a different group of first years at the same university who started a year later, in 2005. This group totaled 106 students out of the total 234 first year population, who had already been placed in English program in eight different groups. The students were chosen from four whole classes at different levels within the total population. At this time, the test was not used to make placement decisions, but was piloted to see how placement decisions could have been made. The data that was collected from both tests (the original 40-item placement test and the revised 23-item pilot test) were analyzed and then compared by looking at the general descriptive statistics, separation, reliability and validity.

Table 1 shows the general statistics of the two tests. We were looking for normal distribution in the scores of both tests. One way to check that the data meets normality is by checking the skewness or kurtosis that are an indication of how wide or high the normal distribution was. In the case of the two tests, using the criterion of ±1.96, we could assume normality of kurtosis and skewness for both tests. The means and standard deviations seemed to suggest that the second test is about half that of the first test, both showing some ability to separate the students. This separation ability of the test (not the test-takers) was confirmed in both normality of distribution and in the separation of scores. The scores of 5.84 and 4.08 indicate that the tests were divided into respectively six and four distinct developmental areas of difficulty that the test produced in the test-takers. These figures suggest that group who took the test first may have had a wider range of abilities than the group that sat the revised version of the test. This difference may also be explained by a smaller N size for the second group compared to the first group.

Table 1

Descriptive Statistics For The Grammar Part Of The Two Placement Tests

		First test N = 231 40 items	Revised test N = 105 23 items
Mean		20.26	11.26
95% confidence interval	Lower Bound	19.62	10.74
for mean	Upper Bound	20.90	11.79
Standard Deviation		4.90	2.730
Skewness		.028	−.226
SES		.160	.235
Kurtosis		−.519	−.251
SEK		.319	.465
Cronbach Alpha		.97	.94
Separation		5.84	4.08

Along with good separation, was a high reliability score for both tests, (a score of one would be perfect reliability). The Rasch item reliability index from Winsteps was a summary of the replicability of the Rasch linear measures of item difficulties, and it decreased from 0.97 on the 40-item form to 0.94 on the 23-item. This indicates that if the items were administered to a different group of examinees, the replicability of item placements would be substantially stable on the reduced-item test.

Validity was measured in terms of fit statistics, which showed items that do not conform to the expected difficulty pattern determined by the mathematical model. The fit statistics for the two tests can be seen on Tables four and five on Appendices C and D. The two sets of figures to pay attention to are the infit and outfit mean squared values. The infit measurement was a probability value of how well the item fitted a specific ability group in the test-taking population and the outfit was a probability measure of how the item matched all of the students who took the test. The acceptable limits of both these sets of figures could be as low as .78 and as high as 1.30 (Bond & Fox, p. 177), with a perfect fit at 1.00. Even though more of the fit scores are slightly nearer the acceptable limits in the second test, compared to the first (the bold items on Table 4 are those retained in the revised version), all the items on both tests fitted the realm of the predicted model. In other words both test items and test-takers were performing in an expected fashion, confirming that the validity of the two tests was acceptable.

In summary despite some homogeneity in the ability of the test-takers in the first group, it was possible to produce a highly reliable first test that resembled a normal distribution necessary to separate students into groups of similar abilities for placement into the English program. Using statistics some redundant items that measured at a similar level of difficulty were removed, shortening the original test from 40 items to 27 items. Further statistics were able to pinpoint which items were separating the lower from the middle and upper scoring students in the test. Also, within the items, the distractors were analyzed to see which were attracting the right level students and which were not. Both items and distractors were either removed or revised. The first revision of 27 items was brought down to 23 items in the final version. This final revised version of the test was given to another group of first year students at another time, assuming similar proficiencies. The results from the second test showed slightly less levels of separation, reliability and validity compared to the first test, but both tests are well within reasonable limits for all facets of both tests. The new revised test will serve as a placement test of first years that join the university's English program in the coming years, assuming that future years' students will be similar in abilities to the ones that took the tests.

IV. Conclusion

Tests of all kinds have tended to be long in nature because it is perceived that the longer the test is the more reliable and valid it is. Little research in second language testing has been given over to showing how tests can be made shorter and function just as well as longer tests. Using primarily the Rasch model, I have shown how it was possible to answer in the affirmative for the question in the purpose section of the introduction that asked if one grammar part of a placement test used at a Japanese university could be reduced and still maintain acceptable levels of reliability, validity, and separation of scores. Of course, this paper does not presume that the grammar test used is in any way a perfect description of the large and incredibly complex construct of grammar knowledge, but merely shows that tests don't necessarily need to be long to function properly. Assuming that the abilities of the students who sat the tests are not too different from students who will sit the placement test in the future, the revised version of the grammar and test could be used as a reliable placement tool. The advantages of improved test efficiency are most practical at the individual school level, where less time needed for assessment and scoring means an increase in the time available for teaching. It is also likely that shorter more efficient tests would lead to lower levels of stress in students. A further need, indicated by Brown (1996 p. 192), calls for a need to base placement decisions on reliable tests, citing time, money and effort to be invested into studying as the main reason for the necessity of consistency of such tests.

References

Bachman, L.F. & Palmer, A.S. (1996). *Language Testing In Practice.* Oxford University Press, NY.

Bell, R. & Lumsden, J. (1980). Test Length And Validity. *Applied Psychological Measurement. Spr [4-2], 165-170.*

Bond, T.G.. & Fox, C.M. (2001). *Applying The Rasch Model: Fundamental Measurement in The Human Sciences.* Lawrence Erlbaum Associates: Mahwah NJ.

Brown, J.D. (1996). *Testing In Language Programs*. Prentice-Hall: Upper Saddle River, NJ.

Burton, R.F, *(2005).* Multiple-choice and True/false tests: myths and misapprehensions. *Assessment & Evaluation In Higher Education. Feb [30], 65-72.*

Ebel, R.L. (1979). *Essentials of Educational Measurement.* 3rd Ed. Prentice-Hall: Englewood Cliffs, NJ.

Lougheed, L. (1999). *Barron's How to Prepare For the Toeic Test.* 2nd Ed. Barron's Educational Series, Inc. Hauppuage, NY.

Appendix A

Table 2 - Item Map for First Grammar Test As Part of Placement test

```
2                 +
                  |  Item 14
                  |T
                  |
                  |  Item 38
                  |  Item 12
                # |  Item 39
                  |
                #T|
              ### |  Item 26
                  |  Item 7
1            .### +  Item 31
                  |S Item 15
              ### |
             #### |
                 S|  Item 37
            ##### |  Item 18   Item 23
        ######### |  Item 28
                  |  Item 34
          ####### |  Item 35
       ########## |
                  |  Item 33   Item 5
            ##### |  Item 9
0     .##########M+M Item 24
                  |  Item 10
           ###### |  Item 1    Item 21   Item 25
         ######## |  Item 40   Item 8
       ########## |  Item 30
                  |  Item 20   Item 3    Item 36
            ##### |  Item 22   Item 32
          #######S|  Item 11
                  |  Item 2    Item 4
              ### |
              ### |  Item 13   Item 16   Item 29
                  |S
-1            ### +
                  |  Item 19
               ##T|  Item 27
                  |
                  |
                  |
                  |
                  |  Item 17
                  |T
-2                +
                  |  Item 6
                  |
                  |
                  |
                  |
                  |
-3                +
```

Note : Items in **bold** are to be kept for the new version of the test.

Appendix B

Table 3 - Item Distractor Analysis And Item Discrimination For First Grammar Test

Item Number	ID	Group	1.	2.	3.	4.	n/r
2	.069	High	.688*	.312	.000	.000	.000
		Middle	.675*	.299	.013	.013	.000
		Low	.610*	.338	.026	.026	.000
3	.397	High	.078	.766*	.117	.039	.000
		Middle	.078	.636*	.169	.117	.000
		Low	.065	.403*	.234	.299	.000
5	.190	High	**.026** ↑	.597*	.013	.364	.000
		Middle	**.052**	.403*	.039	.506	.000
		Low	**.065**	.377*	.169	.377	.013
7	.276	High	.494	**.026** ↑	.416*	.065	.000
		Middle	.688	**.039**	.182*	.091	.000
		Low	.494	**.078**	.208*	.208	.013
9	.603	High	.766*	.026	.039	.169	.000
		Middle	.506*	.117	.169	.208	.000
		Low	.195*	.312	.195	.299	.000
10	.224	High	**.026** ↑	.688*	.078	.195	.013
		Middle	**.052**	.429*	.130	.390	.000
		Low	**.104**	.442*	.195	.247	.013
11	.310	High	**.000** ↑	.104	.805*	.091	.000
		Middle	**.039**	.182	.584*	.195	.000
		Low	**.039**	.221	.545*	.195	.000
12	.000	High	.169*	.610	.013	.208	.000
		Middle	.182*	.468	.065	.260	.026
		Low	.208*	.429	.078	.286	.000
14	.207	High	.156	.571	.052	.221*	.000
		Middle	.208	.623	.026	.143*	.000
		Low	.156	.701	.065	.065*	.013
15	.241	High	.169	.403	.416*	**.013** ↑	.000
		Middle	.208	.416	.312*	**.065**	.000
		Low	.429	.325	.195*	**.052**	.000
16	.397	High	.052	.844*	.078	**.026** ↑	.000
		Middle	.026	.740*	.221	**.026**	.000
		Low	.091	.494*	.364	**.052**	.000

* Denotes correct answer

Item Number	ID	Group	1.	2.	3.	4.	n/r
18	.362	High	.571*	**.013** ↑	.078	.338	.000
		Middle	.299*	**.052**	.143	.506	.000
		Low	.234*	**.065**	.208	.481	.013
19	.276	High	.117	**.013** ↑	.052	.818*	.000
		Middle	.104	**.013**	.104	.779*	.000
		Low	.169	**.169**	.052	.610*	.000
22	.534	High	.091	.039	.052	.818*	.000
		Middle	.091	.104	.130	.675*	.000
		Low	.195	.117	.312	.338*	.039
24	.310	High	.662*	**.000** ↑	.260	.078	.000
		Middle	.455*	**.039**	.416	.091	.000
		Low	.377*	**.156**	.351	.091	.026
25	.517	High	**.000** ↑	.766*	.039	.195	.000
		Middle	**.000**	.545*	.013	.442	.000
		Low	**.065**	.338*	.065	.494	.039
26	.155	High	**.636**	.065	.299*	.000	.000
		Middle	**.649**	.065	.273*	.013	.000
		Low	**.481** ↓	.182	.221*	.078	.039
27	.534	High	.948*	.052	**.000** ↑	**.000** ↑	.000
		Middle	.779*	.169	**.013**	**.039**	.000
		Low	.519*	.312	**.039**	**.078**	.052
28	.172	High	.182	.104	.260	.455*	.000
		Middle	.182	.117	.286	.416*	.000
		Low	.169	.143	.299	.312*	.078
30	.466	High	.779*	.091	.013	.117	.000
		Middle	.610*	.156	.065	.169	.000
		Low	.377*	.234	.065	.247	.078
31	.259	High	.530*	.224	.333	.271	.000
		Middle	.288*	.388	.364	.339	.000
		Low	.182*	.388	.303	.390	1.000
34	.414	High	.244	.297	.442*	.294	.000
		Middle	.356	.297	.358*	.412	.100
		Low	.400	.406	.200*	.294	.900

* Denotes correct answer

Item							
Number	ID	Group	1.	2.	3.	4.	n/r
35	.276	High	.200	.167	.343	.440*	.000
		Middle	.400	.500	.343	.300*	.214
		Low	.400	.333	.314	.260*	.786
37	.259	High	.432*	.436	.155	.381	.063
		Middle	.284*	.400	.379	.381	.125
		Low	.284*	.164	.466	.238	.813
38	.241	High	.356	.296	.561*	.167	.158
		Middle	.333	.361	.317*	.444	.105
		Low	.311	.343	.122*	.389	.737
39	.034	High	.300*	.255	.489	.364	.118
		Middle	.380*	.392	.298	.333	.118
		Low	.320*	.353	.213	.303	.765
40	.414	High	.100	.432*	.291	.000	.111
		Middle	.200	.341*	.418	.333	.167
		Low	.700	.227*	.291	.667	.722

* Denotes correct answer

Appendix C

Table 4 - Infit And Outfit Statistics For First Grammar Test - 2004

		INFIT		OUTFIT	
Item Number	Error	MNSQ	ZSTD	MNSQ	ZSTD
Item 1	.14	1.00	.1	.99	−.2
Item 2	.14	1.12	2.1	1.21	2.9
Item 3	**.14**	**.98**	**−.5**	**.98**	**−.4**
Item 4	.14	1.04	.8	1.08	1.2
Item 5	**.14**	**1.06**	**1.6**	**1.07**	**1.7**
Item 6	.21	.99	.0	.96	−.1
Item 7	**.15**	**1.02**	**.3**	**1.13**	**1.4**
Item 8	.14	1.04	1.2	1.05	1.1
Item 9	**.14**	**.88**	**−3.5**	**.87**	**−3.2**
Item 10	**.14**	**1.05**	**1.4**	**1.07**	**1.6**
Item 11	**.14**	**1.01**	**.2**	**.99**	**−.1**
Item 12	.17	1.11	1.0	1.25	1.8
Item 13	.15	.96	−.6	.92	−1.0
Item 14	**.19**	**.96**	**−.3**	**.89**	**−.6**
Item 15	**.15**	**1.06**	**.9**	**1.10**	**1.3**
Item 16	**.15**	**.96**	**−.7**	**.96**	**−.5**
Item 17	.19	1.01	.1	1.07	.5
Item 18	**.14**	**.96**	**−.7**	**.97**	**−.4**
Item 19	**.15**	**1.00**	**.1**	**1.00**	**.0**
Item 20	.14	.92	−1.9	.90	−1.8
Item 21	.14	.90	−2.8	.88	−2.7
Item 22	**.14**	**.91**	**−2.1**	**.90**	**−1.8**
Item 23	.14	1.19	3.7	1.22	3.5
Item 24	**.14**	**1.03**	**.8**	**1.03**	**.8**
Item 25	**.14**	**.90**	**−2.6**	**.89**	**−2.6**
Item 26	**.15**	**1.08**	**1.0**	**1.16**	**1.7**
Item 27	**.16**	**.88**	**−1.6**	**.80**	**−2.2**
Item 28	.14	1.09	2.1	1.09	1.7
Item 29	.15	.94	−.9	.92	−1.0
Item 30	**.14**	**.95**	**−1.1**	**.95**	**−1.0**
Item 31	**.15**	**.97**	**−.4**	**.98**	**−.2**
Item 32	.14	.99	−.2	.97	−.5
Item 33	.14	.95	−1.2	.94	−1.5

		INFIT		OUTFIT	
Item Number	Error	MNSQ	ZSTD	MNSQ	ZSTD
Item 34	**.14**	**.98**	**—.5**	**.97**	**—.6**
Item 35	**.14**	**1.02**	**.6**	**1.03**	**.6**
Item 36	.14	.93	−1.7	.92	−1.6
Item 37	**.14**	**1.02**	**.5**	**1.03**	**.5**
Item 38	**.18**	**.97**	**—.2**	**1.01**	**.1**
Item 39	.16	1.10	1.2	1.21	1.8
Item 40	**.14**	**.98**	**—.6**	**.98**	**—.4**

Note : Items in bold were those kept for revised version of the test

Appendix D

Table5 - Infit And Outfit Statistics For Revised Grammar Test - 2005

		INFIT		OUTFIT	
Item Number	Error	MNSQ	ZSTD	MNSQ	ZSTD
Item 1	.22	.95	−.5	.92	−.7
Item 2	.22	1.03	.4	1.00	.0
Item 3	.22	1.01	.2	1.10	.8
Item 4	.20	.93	−1.4	.91	−1.4
Item 5	.21	.99	−.2	.87	−3.2
Item 6	.21	.98	−.3	1.07	1.6
Item 7	.31	1.00	.1	.99	−.1
Item 8	.21	1.02	.4	1.25	1.8
Item 9	.20	.97	−.4	.96	−.6
Item 10	.23	1.02	.2	1.04	.3
Item 11	.20	1.02	.3	1.03	.5
Item 12	.30	.95	−.2	.85	−.5
Item 13	.20	1.24	4.2	1.29	4.3
Item 14	.21	.94	−1.0	.92	−1.2
Item 15	.21	1.03	.5	1.03	.4
Item 16	.24	.90	−.8	.85	−1.0
Item 17	.22	1.03	.4	1.02	.2
Item 18	.22	.92	−.9	.90	−1.0
Item 19	.21	1.06	.8	1.13	1.5
Item 20	.20	.98	−.3	.96	−.6
Item 21	.20	1.02	.5	1.04	.7
Item 22	.28	.99	.0	.94	−.2
Item 23	.21	.99	−.1	1.01	.1

—Abstract—

Little research in second language studies exists to support the claim that shorter versions of the same test can display similar levels of reliability and validity as longer versions. This paper attempts to show, through statistical analysis using the Rasch Model, how the grammar section of a placement test used at a Japanese university could be revised and made shorter (40 to 23 items). This shorter version was piloted on a separate set of students and the results show that the original and revised versions of the test display similarly acceptable levels of reliability, validity and an ability to separate the grammar knowledge of the test-takers.

長いテストと同じように短いテストでも同様の信頼性や妥当性が得られるという証明は、第2言語学習の分野ではほとんど行われていない。本論文では、ある日本の大学のクラス分けテストの文法セクションで用いられた40項目をいかに23項目へと短く改訂できるかを、Rasch Modelを用いた統計的分析を通して検証する。編約版のテストは別の学習者グループで試用され、元々のテストと編約版のテストはほぼ同程度の信頼性と妥当性を備え、学生の文法力を同じように分別できることが示された。