Developing a Sample-Free Grammatical Proficiency Test for SLA Research

Yuko SHIMIZU
Ritsumeikan University
Kenichi YAMAKAWA
Yasuda Women's University
Naoki SUGINO
Ritsumeikan University
Hiromasa OHBA
Joetsu University of Education
Michiko NAKANO
Waseda University

Abstract

This paper summarizes the process of developing a grammar test, which is intended to be used in the context of second language acquisition research. The cyclic process of test development is introduced followed by the delineation of an actual procedure of a grammar test development. Then, the results of pilot tests are provided using both classical item analysis and a two-parameter logistic model of Item Response Theory. We conclude that by refining the items based on classical item analysis, the newly developed test satisfies our criteria to serve as a measurement tool for our future research. However, we need more data to reconfirm its discrimination power.

1.0 Purpose

Our basic concern has been to bridge the two fields of Second Language Acquisition (SLA) and Language Testing (LT) for a deeper understanding of each research area. Historically, each area has developed separately and has devised separate research methodology in its own right. It is true that both areas are contributing in more ways than one to language learning and teaching; however, several problems can also be acknowledged. In data analysis in SLA studies, due heed is not always paid to validity and reliability issues, which are crucial in LT. On the other hand, many studies in LT regard validity and reliability as important, but they sometimes fail to see language ability as a systematic whole, a model of which SLA studies based on a linguistic theory attempt to construct. Also, a modern testing theory such as Item Response Theory (IRT) is not fully contributing to investigating phenomena in language learning and acquisition.

Realizing these issues, we have been carrying out several studies to connect the two areas of

SHIMIZU Yuko is a professor in the College of Economics, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577, Japan. YAMAKAWA Kenichi is an associate professor in the Department of English, Yasuda Women's University, 6-13-1, Yasuhigashi, Asaminami-ku, Hiroshima 731-0153, Japan.

The present study is supported by the Grant-in-Aid from the Japan Society for the Promotion of Science 2004-2007 (16320078).

SLA and LT. Our main concern is in investigating the development of grammatical competence in L2 learners, particularly from the aspects of unaccusative and unergative verbs (Yamakawa et al., 2003, 2005; Nakano et al., 2005), *wh*-questions (Ohba et al., 2005), relative clauses (Ohba et al., 2003), dative alternation (Sugino et al., 2005), and passive forms (Nakano et al., 2005).

For the purpose of identifying the subjects' levels of English proficiency cross-sectionally, we used a grammar test in our previous research, which we compiled from several standardized tests (Shimizu et al., 2003). However, we keenly felt the necessity of developing our own testing instrument, whose items had sample-free indices. In this paper, we will first show the process of developing a new grammar test, which is followed by a report on the results of a pilot study.

2.0 Process of Test Development

Hattie, Jaeger and Bond (1999) view educational testing as a cyclic process of conception models, construction, administration, use and evaluation. The process often starts with the specification of conceptual models of measurement. There are several models of measurement and we initially followed Classical Test Theory (CTT). Then there begins the task of item construction followed by test administration, which we will describe under the headings of *Test Construction* and *Pilot Test Administration #1*. Before we actually use the developed test for specific purposes—SLA research in our case—we must evaluate the results and make the necessary modifications of the items before readopting conceptual models. We will address the second cycle of test development showing the result of data analysis using IRT.

3.0 Test Construction

The stage of test and item development includes the selection of a test format, scoring rules and specification of testing points. Since we needed to develop a grammar test that makes machine scoring possible, we chose a multiple choice type test for this particular test.

Graduate students in the field of Teaching English as a Foreign Language (TEFL), some of them having rich experience in teaching English to Japanese EFL learners, constructed multiple-choice grammar items as a part of the requirements for a language testing course. An instructor of the course assigned grammatical features (e.g., tense-aspect, relative clauses) to each student to write several items with a stem and four answer choices for each item. The items the students wrote were discussed during class and revised to make them more plausible. A total of 186 items was then reviewed by two native speakers of English whose background was in Applied Linguistics. On receiving feedback from the reviewers, we made the necessary changes by modifying and replacing items and distractors. As a consequence, 54 unsuitable items were deleted from the item bank. Using the resulting 132 items, two test forms—Form A with 67 items, and Form B with 65 items—were compiled.

3.1 Pilot Test Administration #1

As a part of the test development process, conducting a pilot test and doing statistical analysis to ensure the quality of the test items are inevitable before the measurement tool is used with the actual subjects in the research. In our pilot study, we used intact groups of students, using English classes at three universities. Three hundred and thirteen students took Form A and 292 took Form B. In among these two groups were 263 students who took both Forms A and B. All participants were Japanese learners of English. Each set required 40 minutes to complete. A three- to seven-day interval was given if the same students were to take both Forms A and B.

3.2 Test Evaluation #1: Classical Item Analysis

Following CTT, which is described as the classical true score model, we obtained item level information to see potential problems with the answer keys and the distractors as well as item characteristics of difficulty and discrimination. The indices we used were item facility, item discrimination, point-biserial correlation coefficient, and distractor efficiency analysis.

Item facility is a measure of the difficulty of an item, arrived at by dividing the number of students answering correctly by the number of students taking the test. Generally speaking, a test's aim is to have an overall facility of approximately .5. However, it is acceptable for individual items to have a higher or lower facility, ranging from .2 to .8. Or, as Brown (1996, pp. 69-70) states, items that fall in arrange between .30 and .70 are considered acceptable.

Item discrimination indicates the degree to which an item distinguishes the test takers who performed well from those who performed poorly. Traditionally, the upper 27% and the lower 27% of the examinee groups are compared to obtain a stable item discrimination index. The item discrimination index can take the values between +1.00 and -1.00. However, it is expected that discrimination will fall in a range between .2 and 1.0.

Obviously, giving two sets of 40-minute test with 132 items has little practical use. Therefore, using the results of item analysis, we reduced the number of items in the test set by adopting refined items. Tables 1 and 2 summarize the basic statistics of item facility and item discrimination of Form A and From B respectively.

Table 1 Item Facility and Item Discrimination of Form A

	Item Facility	Item Discrimination
mean	.543	.325
standard error	.019	.019
median	.560	.381
mode	.430	.405
standard deviation	.156	.159
minimum	.130	-.131
maximum	.870	.607
n	67	67

Table 2 Item Facility and Item Discrimination of Form B

	Item Facility	Item Discrimination
mean	.559	.380
standard error	.021	.020
median	.570	.410
mode	.700	.154
standard deviation	.171	.165
minimum	.170	-.013
maximum	.910	.718
n	65	65

There are some common criteria of item facility and item discrimination. Item facility indexes from .20 to .80 are said to be acceptable by some researchers, while Brown (1996) claims that indexes from .30 to .70 are acceptable as stated above. However, the criteria will depend on the situation and the purpose of the test. Brown (1996, p. 69) notes that ideal items in a norm-referenced test development project have an average item facility of .50 and the highest available item discrimination. Regarding item discrimination, Ebel (1979, p. 267) suggests that test items with item discrimination indexes of .40 and up are considered to be very good, while those below .19 are to be rejected or improved by revision. We basically followed these criteria to select and reject items.

Firstly, calculating the item facility and the item discrimination, we eliminated the most difficult items, whose item facility indexes were below .30, and the easiest items with indexes of .80 and above. Also, we excluded items whose item discrimination indexes were below .30, which we considerably increased the amount of scores cut. Thus, we accepted fewer items.

Then, reexamining the testing points, we excluded redundant items. We also made necessary revisions of the distractors using the information from distractor efficiency analysis, which allowed us to see how each distractor attracts the test takers.

As a result of this careful examination of the items, a new test set with 35 items was compiled. Table 3 shows item level statistics of the new test set.

Table 3 Item level statistics for selected 35 items

	Item Facility	Item Discrimination
mean	.538	.493
standard error	.020	.012
median	.550	.487
mode	.460	.488
standard deviation	.120	.069
minimum	.330	.393
maximum	.800	.718

Seeing the mean of item facility index (.538) and the ranges of item facility index (from .330 to .800) and item discrimination index (from .393 to .718), those 35 items as a whole satisfied our criteria to serve as a complete set of a grammar test. Before printing items in a final form for a fill-fledged pilot test, we rearranged the items according to their testing points as well as difficulty levels from easy to difficult based on the item facility indexes.

3.3 Test Administration #2: Revised Version

The revised version of the grammar test needed to be administered again with the target population to see if it was feasible and if items were functioning appropriately. We used intact groups of students, using 20 minutes of English classes at three Japanese universities from May through June, 2005. All participants were Japanese learners of English, but none of them participated in the first pilot test. The total number of the participants was 355.

3.4 Test Evaluation #2: Revised Version

Firstly, we conducted the classical item analysis to obtain test level information as well as item level information. Tables 4 and 5 provide details of the statistics.

Table 4 Item Analysis: Test Level Statistics

number of examinees	355
number of test items	35
mean	18.45
variance	37.80
standard deviation	6.15
kurtosis	-.74
skewness	.15
range	27
minimum	6
maximum	33
KR-20	.81
standard error (based on KR-20)	2.68

Table 5 Item Analysis: Item Level Statistics

	Item Facility	Item Discrimination
mean	.527	.437
standard error	.023	.019
median	.500	.463
mode	.390	.495
standard deviation	.135	.111
minimum	.240	.200
maximum	.780	.674
n	35	35

In order to see the extent to which the results can be considered stable, the internal-consistency reliability was obtained using the Kuder-Richardson formula 20 (KR-20). The result proved that internal-consistency reliability was considerably high (.81). The mean of item facility indexes (.527) was fairly ideal and the ranges of item facility indexes and item discrimination indexes were still able to fulfill acceptable criteria.

Classical item analysis provides useful information to improve the quality of the test. However, there is an important limitation, which comes form sample-dependency. That is, the statistics we obtain from classical item analysis is sample-based information. If we give the same set of the test to a different group of test takers, it is quite possible that we will obtain totally different statistics. In other words, the difficulty and discrimination estimates we get in CTT depend on the level of ability of the people who we gave the test to. That is, there is an inter-dependence of items and candidates.

It is not problematic if the test is for a classroom-based situation or for the use of specific target participants. However, if we are to compare performances of different groups of test takers, we need more precise analysis to obtain more reliable item information. In order to get a population/ sample free estimate, therefore, we chose IRT as a second conceptual model of measurement.

4.0 Better ways to estimate test data

IRT assumes that there is a correlation between the score gained by a test taker for one item and his/her overall ability on the latent trait which underlies test performance. The characteristics of an item are said to be independent of the ability of the test takers who were sampled.

IRT comes in three forms reflecting the number of parameters: Discrimination Parameter (a_i), Difficulty Parameter (b_i), and Pseudo-chance-level Parameter (c_i). In a one-parameter logistic model, which is often called the Rasch model, only the item difficulty is considered. Difficulty can be defined as the level of ability required to be more likely to correctly answer the question than get it wrong. In a two-parameter logistic model, both difficulty (b_i) and discrimination (a_i) are considered. Discrimination can be defined as how well the question is at separating candidates of similar abilities. In a three-parameter logistic model, chances (c_i) are considered in addition to difficulty and discrimination parameters. Chance is the random factor which enhances a candidate's probability of success through guessing.

Since the number of participants was not sufficient to use a three-parameter logistic model in the present study, we chose a two-parameter model to analyze our data. Table 6 shows the *a* parameter and *b* parameter in IRT and the item facility and point-biserial correlation coefficient followed by the number of test takers who responded to each item. Parameter summaries are provided in Table 7.

Table 6 Final Item Parameter Estimates

item	*a*	*b*	IF	PB	n
1	.67	-1.34*	.75	.46	355
2	.45	-1.70*	.74	.25	355
3	.48	-.88*	.63	.34	355
4	.49	-1.55*	.73	.31	355
5	.37	-.89*	.61	.17	355
6	.43	-1.09*	.65	.27	355
7	.45	-1.68*	.74	.26	355
8	.50	-1.02*	.66	.36	355
9	.44	.28*	.46	.28	355
10	.54	-.61*	.60	.41	355
11	.54	-.86*	.64	.41	355
12	.61	.13	.48	.46	355

13	.55	-.12*	.52	.41	355
14	.44	-.53*	.57	.29	355
15	.54	-.46*	.57	.41	355
16	.52	.15	.48	.38	355
17	.44	-1.99*	.78	.22	355
18	.45	-.01*	.50	.29	355
19	.61	-.38*	.56	.45	355
20	.52	.51	.42	.40	355
21	.71	.02	.50	.51	355
22	.50	.23	.47	.38	355
23	.56	-.77*	.63	.42	354
24	.43	.76	.39	.28	354
25	.52	.38	.44	.39	353
26	.58	-.43*	.57	.44	353
27	.54	.72	.38	.43	353
28	.55	.34	.44	.42	353
29	.53	.68	.39	.40	351
30	.47	.82	.38	.34	351
31	.45	.74	.39	.32	351
32	.59	.34	.44	.46	351
33	.50	1.01	.34	.36	351
34	.46	1.76	.24	.29	350
35	.57	.57	.40	.44	348

IF = item facility index PB= point-biserial correlation coefficient

Table 7 Parameter Summary

	a parameter	*b* parameter
Mean	0.514	-0.196
SD	0.072	0.889
Minimum	0.370	-1.990
Maximum	0.710	1.760

The difficulty parameters (b_i) ranged from -1.99 (item 17) to 1.76 (item 34), where their mean was 0 and their standard deviation was 1. Eighteen items (with asterisks) were below the average logit of zero and 17 items were above the average logit of zero. We concluded that the items were fairly balanced in terms of item difficulty. In addition, easier items were located early so that some items became good lead-in items to the test.

The discrimination parameters (*a* parameter) ranged from .37 (item 5) to .71 (item 21). An *a* value should be above .30 and all the items satisfied this criterion. However, no items had the discrimination parameter of above 1.00, which left us some discussion about why this happened.

4.1 Correlations of IRT and CTT indexes

We obtained test level and item level statistics in both CTT and IRT. Even if each test taker has different latent ability, difficult items should appear relatively difficult, and easy ones should be relatively easy within the given test takers. That is, no matter which analysis we use, we will obtain similar tendency in the results. Therefore, it is natural that we would observe strong correlations between classical item analysis and IRT as shown in Tables 8 and 9. Since each item had a difficulty index in classical item analysis and a *b* parameter in IRT, a correlation coefficient between the two was calculated. Also, since each test taker had his/her total score based on the number right scores in CTT and person ability information in IRT, a correlation coefficient of the two was obtained. The correlations were .996 for the former and .991 for the latter. This proved that the results in CTT and IRT were strongly correlated.

Table 8 Correlation between Difficulty Indexes in CTT and IRT

		CTT
IRT	Pearson Correlation	.996 (**)
	Sig. (2-tailed)	.000

p< .01 n=35 items

Table 9 Correlation between Total Score in CTT and Person Ability in IRT

		CTT
IRT	Pearson Correlation	.991 (**)
	Sig. (2-tailed)	.000

p< .01 n=355 test takers

5.0 Conclusion

The cyclic process of refining test items was outlined in this paper. Starting with the original 186 items, necessary modifications and the discarding of items were performed to reduce the number of items to 35 (18.8%). Classical item analysis provided enough information to prove that the final test set was satisfactory in both difficulty and discrimination criteria as far as the same or similar target samples were concerned. However, as the skewness of .15 indicated (see Table 4), the distribution slightly skewed positively, which means the scores were piled up at the low end of the scale, and tailed off near the high end of the scale. That is, the participants in the current study were primarily concentrated in the intermediate and lower levels learners. This suggests that we should collect more data to avoid the disadvantage of sample dependency in CTT and to effectively utilize the information obtained by IRT, since the more participants we obtain, the closer the CTT information gets to the estimates obtained by IRT. In order to utilize CTT and IRT complementarily, not exclusively, we need follow up the current study to examine the discrimination power of the test set.

References

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. CA: Wadsworth.

Ebel, R.L. (1979). *Essentials of educational measurement (3rd ed.)*. Englewood Cliffs, NJ: Prentice-Hall.

Hattie, J., Jaegaer, R.M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education*, *24*, 393-446.

Nakano, M., Sugino, N., Ohba, H., Yamakawa, K., & Shimizu, Y. (2005). An analysis of grammatical judgment test: Dative constructions, their passive forms, unaccusative and unergative constructions. *Proceedings of the 9th Conference of Pan-Pacific Association of Applied Linguistics*, 386-394.

Ohba, H., Sugino, N., Nakano, M., Yamakawa, K., Shimizu, Y., & Kimura, S. (2003, August). *The development of grammatical competence of Japanese EFL learners: Focusing on relative clause constructions.* Paper presented at the 29th Conference of the Japan Society of English Language Education, Miyagi, Japan.

Ohba, H., Yamakawa, K., Sugino, N., Shimizu, Y., & Nakano, M. (2005, August). *The acquisition of* wh-*questions by adult Japanese EFL learners.* Poster session presented at the 10th Conference of Pan-Pacific Association of Applied Linguistics, Edinburgh, UK.

Shimizu, Y., Kimura, S., Sugino, N., Yamakawa, K. Ohba, H., & Nakano, M. (2003). Eibumpou nouryoku hyoujun tesuto no shinraisei･datousei no kenshou to shin-eibumpou nouryoku tesuto Measure of English Grammar (MEG) [The validity and reliability of standardized English tests and compilation of “Measure of English Grammar (MEG)”]. *Policy Science, 10* (3), 59-68.

Sugino, N., Nakano, M., Ohba, H., Kimura, S., Yamakawa, K., & Shimizu, Y. (2005). The development of grammatical competence of Japanese EFL learners: Focusing on dative alternation. *Proceedings of the 9th Conference of Pan-Pacific Association of Applied Linguistics*, 322-331.

Yamakawa, K., Sugino, N., Kimura, S., Nakano, M. Ohba, H., & Shimizu, Y. (2003). The development of grammatical competence of Japanese EFL learners: Focusing on unaccusative/unergative verbs. *Annual Review of English Language Education in Japan, 14,* 1-10.

Yamakawa, K., Sugino, N., Kimura, S., Nakano, M., Ohba, H., & Shimizu, Y. (2005). Nihonjin eigo gakushuusha ni yoru hitaikaku-doushi to hinoukaku-doushi no shuutoku. [Acqusition of naccusative/unergative verbs by Japanese EFL learners]. *JACET Chuugoku-Shikoku Chapter Research Bulletin, 2,* 91-110.