コミュニケーション学科論集
茨城大学人文学部紀要　No.15 2004. 3

Measuring English Language Proficiency Using the Criterion-Referenced Test (1)

— analyses by the classical test theory —

Taiko Tsuchihira

Introduction

Background

Recently, curriculum reforms are taking place at English language programs of many universities. This reflexive action might have come from the criticisms of ineffective teaching in the last era. Schoolteachers, newspapers, and even students pointed out that they could not attain adequate English proficiency even though most Japanese students study English for many years at school. Now, English language programs are required to demonstrate more effective teaching, and the importance of the assessment is rapidly being recognized.

Similarly to some other universities in Japan, a new language program called the Integrated Language Program (IEP) has started at Ibaraki University in Japan as a curriculum reform. The goal of the program is to develop the language skills required for Japanese university students. It includes training four skills (i.e. listening, speaking, reading, and writing) and the skills in academic situations (English for Academic Purposes). The program employs the language test called G-TELP (General Test of English Language Proficiency) at some stages of the year just as other programs. However, since G-TELP is a criterion-referenced test (CRT), it requires us to use different methods of analyses to examine the results from a norm-referenced test (NRT). In this paper, several aspects of CRT are described first, and the processes of analyzing the CRT scores are presented using G-TELP scores. Finally, the results of the analyses are discussed and summarized. Since techniques of analyzing the CRT results are still being developed, the present study will be able to present interesting examples of the analyses.

Literature Review

Criterion-referenced test (CRT) vs. norm-referenced tests (NRT). According to Henning (1987), criterion-referenced tests (CRT) are devised before instruction is designed.

CRT tests match the teaching objectives perfectly, and a criterion or cut-off score is set in advance (usually 80 to 90 percent of the total possible score). Students are not evaluated by comparison with the other students, but their achievement is measured with respect to the degree of their learning or mastery of the content domain. The positive aspects of CRT are (a) they help clarify the achievements or problems of course objectives, (b) they are useful with small groups where norms are not available, and (c) students' test anxiety is believed to be low because they know what is on the test. The negative sides of CRT are (a) the content domain is too limited, (b) the comparison with other students are typically unavailable for students, (c) the standards or cut-off scores are in practice rather arbitrary, and (d) techniques of estimating reliability and validity of CRT are still being developed.

Norm-referenced tests are quite different. Henning (1987) defines that norm-referenced tests must have been administered to a large sample. The standards of achievement are found by reference to the mean or the mean score of other students. In order to obtain a broad distribution, items at various levels of difficulty are included. Their strengths are (a) the comparison can easily be made with the performance or achievement of other students, (b) the standards of achievement are less arbitrary since they are based on the performance of other students, and (c) more comparative information is provided than CRT. Their weaknesses are, on the other hand, (a) they are valid only the population are normed, (b) it is difficult to match the result with the instructional objectives, (c) test anxiety may be fostered in these tests, and (d) they are said to be insensitive to fluctuations in the individual. The reasons for choosing CRT will be described in the next session.

Why criterion-referenced tests? There were several reasons for choosing the CRT for the IEP. As for practical reasons, factors such as test time, scoring time, and test fees were considered. However, there is also a theoretical advantage for choosing CRT for this program. Especially when we wish to examine the students' improvement or the effectiveness of the program, CRT works far better than NRT.

W. James Popham, who is one of the most important researchers and advocates for CRT, continuously describes that a standardized achievement test, which has shares most of the characteristics with NRT, provides a misleading estimate of students' improvement or a school staff's effectiveness (Popham, 1977, 1978, 1999, 2001). He lists three reasons as follows:

1. To sell standardized achievement tests, their descriptors should be general, and it is

extremely difficult to match what on the test and what is being taught.

2. Standardized achievement tests usually consist of mostly middle difficulty items because they exclude easy or difficult items in the process of improvements. However, items on which students do well are often the items which teachers stress.

3. Standardized achievement tests measure students' native intellectual ability and students' out-of-school learning other than what's taught in school.

(Popham, 1999)

For these reasons, he states that standardized achievement tests work better in supplying the evidence needed to make comparative inference of a student's ability to other students than checking the improvement or achievement. As he says, employing standardized achievement tests to ascertain educational quality is like "measuring temperature with a tablespoon" in a sense that the tools are used for wrong missions (Popham 1999, p. 10).

In the IEP, measuring the effectiveness of the program is one of the important roles that the test had to play, and its effectiveness were demonstrated using CRT as in Tsuchihira and Kumazawa (2003). This is one of the biggest advantages of CRT.

Previous studies on CRT. CRT had deeply studied since 1960s, and there have been many articles and books. In educational measurement journals, the studies on CRT were originated in the work of Robert Glaser and William J. Popham. However, as Popham (1978) points out, some of them are not practical, and we still need more practical examples, especially on language testing not on education general. Here, I would like to introduce some of the CRT studies focusing on language teaching.

Lynch and Hudson (1984) present an example of criterion-referenced language test development (CALTD). They first introduce the criterion-referenced measurement principles and compare them with norm-referenced approaches in terms of the types of decisions that result from either approach. CRLTD utilizes the test specifications as Popham (1978) proposed, and produces items and tasks from them. They introduce CALTD processes for deciding specifications, their refinement, and the way they link testing and teaching. They examine the data from teachers' workshops and feedback, and conclude its benefits of interactive processes between teaching and testing.

Cartier (1968) seems to be the first article on CRT appeared in *TESOL Quarterly*. He first describes the contrast between NRT and CRT, and discusses the possibility of applying the criterion-referenced measurement in language testing. According to him, it is easier to introduce CRT in simpler, mechanical jobs than language teaching, since they at least know their mission clearly. He describes the idea of what a criterion test is like,

and especially emphasizes that we need to elicit the actual behavior in CRT. He points out that the language teachers tend to set up vague, general, idealistic objectives, and presents the critical view that we will not succeed in CRT unless we have a systematic appraisal of the students' real needs and the actual language behavior.

Cziko (1982) presents the study attempting to modify the existing ESL dictation tests into the psychometric, criterion-referenced tests with practical qualities. By "practical qualities", he means the test to (a) be applicable to a wide range of ability, (b) be easy and fast to score, (c) be consisted of set items with a unidimensional and cumulative scale, and (d) yield scores that are interpretable with specified levels of English proficiency. With descriptive statistics, correlational analyses and the Guttman analysis, the nature and the validity of the dictation test were examined. In addition, they proved that the modified dictation test with the visual modality to be a measure of general second language proficiency in the same way as that with the audio modality with correlational statistics.

Furthermore, the recent book by Brown and Hudson (2002) provides a precious, comprehensive views on criterion-reference language testing. It first describes (a) historical backgrounds and definitions of CRT, and emphasize (b) the relationships between curriculum and testing. The authors, then, continue to discuss (c) the nature of CRT items, and introduce the analyses of CRT results in terms of (d) descriptive statistics, (e) reliability, dependability, and unidimensionality, and (f) validity. Finally, they present the ways and problems in (g) administering, giving feedback, and reporting the results in CRT. It seems that this is the first and most comprehensive book on CRT analyses in language testing, and the author considers it meaningful to analyze the existing data of G-TELP with the methods presented.

The Purpose of the study

In the present situation of the IEP, G-TELP is being used with two conflicting purposes: (a) as a placement test, and (b) to examine the effectiveness of the program. Theoretically, it is extremely difficult to fulfill these two purposes with one test though we need to reconcile with the practical situation. In other words, it is a big burden for one language program to conduct two different kinds of tests in multiple occasions to hundreds of students in a year. In sum, from financial and practical reasons, the roles of the two types of tests, NRT and CRT, are being required to G-TELP.[1]

In spite of the present situation, however, it is still essential to examine how G-

TELP plays its roles in the program. Especially, it is meaningful to provide additional examples of analyzing CRT scores. Different data and methods of analyses might bring different results. How should we interpret them? The purpose of the study is to answer the following questions:

(a) To what extent G-TELP plays its role as the CRT?

(b) Do different methods of analyses bring different results? Do they not?

(c) What do the differences/similarities of the results tell us? What are the implications for the further use of the test?

Method

Participants

All the participants are the students enrolled in the IEP classes. IEP started at Ibaraki University in Japan as a curriculum reform in 2002. The goal of the program is to develop the language skills required for Japanese university students. As explained earlier, it includes training four skills (i.e. listening, speaking, reading, and writing) and the skills in academic situations (English for Academic Purposes). The participants are all first-year students at Ibaraki University majoring either in agriculture or social sciences. Though 376 students took the test in April, 2003 as a pretest, after matching the data with the posttest (July, 2003), the scores of 353 students were used for the analyses. Their English levels can be regarded as the average of those of Japanese university students.

Material

G-TELP was developed at the International Testing Center by Dr. Robert Lado, Francis Henofotis and their colleagues. It is a criterion-referenced test which assesses the English language proficiency of nonnative speakers in real-world situations. It provides detailed, task-referenced information on the examinees' performance according to their test levels. It provides the examiness with diagnostic reports indicating their strengths and weaknesses, and what they can do to improve their English further. The test forms used in the study are Level 3 which is at TOEIC 300-400 level. The structure of the test is shown in Table 1. As a common practice in CRT, scores are culculated in percentages for each section and added up. Therefore, 300 is the maximum. Futher information and test descriptors are available from G-TELP Japan.

Table 1. Structure of G-TELP Level 3

	Grammar	Listening	Reading & Vocabulary
No. of items	22	24	24
Time	20	20	35

Procedures

Participants took the test twice. They took once at the beginning of the semester (April, 2003) as a pretest, and at the end of the semester (July, 2003) as a posttest. The tests were administered in several classrooms as a part of ordinary classes.

Methods of analyses

After matching the examinees, all the data are to be analyzed according to the procedures introduced in Brown and Hudson (2002). However, because of its volume and variability, the present study limits its focus to descriptive statistics and item analyses based on the classical test theory. Therefore, it deals with (a) descriptive statistics, (b) the B-index, (c) the agreement index, and (d) the item phi (ϕ). The analyses on discrimination, validity, and those using other testing theories will be covered in the following studies.

Results

Descriptive statistics

First of all, let us view the score distributions descriptive statistics. According to the criterion suggested by Tachnick and Fidell (2001), if we look at Table 2, we will notice that the values of skewness and kurtosis in the pretest are less than twice as much as the standard errors, which suggests the normality of the distribution. If we look at the distribution of the G-TELP score in the pretest, we can see that it is the normal distribution. On the other hand, for the posttest, we can see that the values of skewness and kurtosis exceed the twice of the value of the standard errors. This means that the distribution is skewed and has kurtosis. And from the values of the skewness and kurtosis, and the distribution in Figure 2, we can tell that the distribution of the posttest is negatively skewed and has leptokurtosis.

Table 2. Descriptive statistics for G-TELP scores

	Range	Min.	Max.	Mean		SDVariance	Skewness		Kurtosis	
				Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Pretest	193.00	43.00	236.00	149.30	1.59	29.80	887.95	−0.21	0.13	0.32	0.26
Posttest	220.00	64.00	284.00	182.51	1.68	31.62	999.97	−0.32	0.13	0.61	0.26

N=353

Figure 1. Distribution of G-TELP score (pre-test, April)

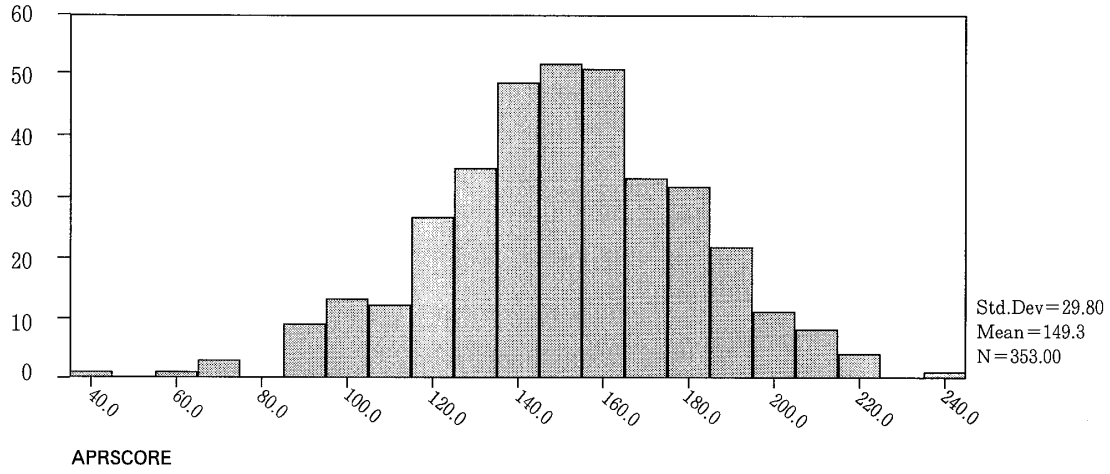
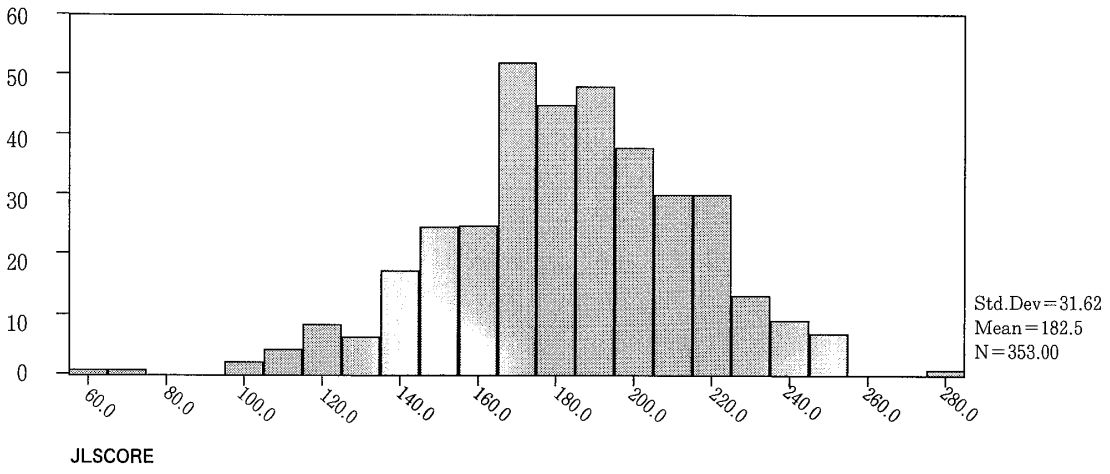


Figure 2. Distribution of G-TELP score (post-test, July)



Brown and Hudson (2002) present the essential contrast between NRT and CRT. According to Brown and Hudson (2002). a CRT achievement test in a course where the students all performed reasonably well and learned the material that was being tested might be expected to produce a negatively skewed distribution. On the other hand, NRT should produce the normal distribution. In this sense, the fact that this test form of G-

TELP formed the normal distribution is a good evidence to use it as a placement test at the beginning of the semesters.

According to Brown (1996), this negatively skewed distribution is ideal for the CRT posttest and in this aense this form of G-TELP is suitable to the present situation. However, the most ideal pattern is to have the contrast of positive skew (pretest) and negative skew (posttest). Brown (1996) states that an ideal pattern is creating a negatively skewed distribution after the instruction and a positively skewed distribution before the instruction took place. This is because G-TELP is a commercial test, it does not exactly reflect what has been taught in the class. Moreover, the items in each test should be matched by the content to compare the exact improvement made by instruction. Since the test forms used in the pretest and posttest are different, each section in each test form might be assessing different realms of English. In order to prove the positive teaching effect with more confidence, the linguistic knowledge or skills that each item assessed need to be matched more closely.

Item analyses

B-index. According to Brown and Hudson (2002), what are mostly involved in CRT analyses are to measure achievement, categorical status, and mastery levels. Here, the present study limits its analysis to the cut-off score indices. First of all, as a measure of the cut-off score, the B-index was calculated by the following formula:

$$\text{B-index} = IF_{pass} - IF_{fail}$$

where:

B-index = difference in IF between students who passed and failed a test

IF_{pass} = item facility for students who passed the test

IF_{fail} = item facility for students who failed the test

(Brown & Hudson 2002, p. 123)

The values of The B-index for each item are summarized in Table 3. In the statistics in Table 3, we can find some items have extremely low values, and others have moderate values. In Table 4, the number of the items and their percentages for each value range are shown.

Table 3. The result of B-index analysis

Item	IF_{pass}	IF_{fail}	B	Item	IF_{pass}	IF_{fail}	B
No.1	0.915	0.627	0.289	No.36	0.366	0.391	-0.025

No.2	0.873	0.546	0.327 *	No.37	0.197	0.155	0.042
No.3	0.972	0.817	0.155	No.38	0.592	0.595	-0.004
No.4	0.859	0.567	0.292	No.39	0.676	0.518	0.158
No.5	0.958	0.866	0.092	No.40	0.859	0.806	0.053
No.6	0.958	0.796	0.162	No.41	0.606	0.345	0.261
No.7	0.930	0.761	0.169	No.42	0.775	0.743	0.032
No.8	0.986	0.673	0.313 *	No.43	0.183	0.180	0.004
No.9	0.437	0.292	0.144	No.44	0.366	0.352	0.014
No.10	0.915	0.694	0.222	No.45	0.507	0.356	0.151
No.11	0.915	0.634	0.282	No.46	0.479	0.468	0.011
No.12	1.000	0.933	0.067	No.47	0.324	0.218	0.106
No.13	0.930	0.750	0.180	No.48	0.845	0.634	0.211
No.14	0.761	0.479	0.282	No.49	0.746	0.616	0.130
No.15	0.986	0.877	0.109	No.50	0.930	0.729	0.201
No.16	0.789	0.623	0.165	No.51	0.606	0.363	0.243
No.17	0.930	0.778	0.151	No.52	0.930	0.725	0.204
No.18	0.817	0.511	0.306 *	No.53	0.606	0.408	0.197
No.19	0.775	0.415	0.359 *	No.54	1.000	0.915	0.085
No.20	0.817	0.746	0.070	No.55	0.845	0.637	0.208
No.21	0.690	0.380	0.310 *	No.56	0.901	0.507	0.394 *
No.22	0.592	0.352	0.239	No.57	0.873	0.739	0.134
No.23	0.873	0.824	0.049	No.58	0.972	0.880	0.092
No.24	0.634	0.556	0.077	No.59	1.000	0.933	0.067
No.25	0.451	0.423	0.028	No.60	0.944	0.803	0.141
No.26	0.577	0.440	0.137	No.61	0.972	0.937	0.035
No.27	0.437	0.342	0.095	No.62	0.944	0.789	0.155
No.28	0.592	0.475	0.116	No.63	0.761	0.560	0.201
No.29	0.183	0.225	-0.042	No.64	0.718	0.518	0.201
No.30	0.620	0.514	0.106	No.65	0.732	0.384	0.349 *
No.31	0.507	0.327	0.180	No.66	0.915	0.585	0.331 *
No.32	0.676	0.644	0.032	No.67	1.000	0.912	0.088
No.33	0.859	0.817	0.042	No.68	0.944	0.595	0.349 *
No.34	0.296	0.218	0.077	No.69	0.789	0.444	0.345 *
No.35	0.211	0.151	0.060	No.70	0.634	0.377	0.257

* —very good items in the B-index

Although more than half of the items show moderate values in the B-index, there are some items with extremely low values. Especially, item 29, 36, and 38 have negative values. Interestingly, all of them are in the listening section. And if we look at the B-index averaged by section, we can see that the B-index for the listening section is much lower than other sections. Since the detailed content analyses have not done yet, it is not clear if it is because of the bad-written items. Further analyses on the content will reveal what is exactly happening in the listening section.

Table 4. Distribution of the items by the B-index

	B< 0	0 <B<0.1	0.1<B<0.2	0.2<B<0.3	0.3<B
No. of item	3	22	20	15	10
Percentage	4.29	31.43	28.57	21.43	14.29

Table 5. B-index means for each section

	Grammar	Listening	Reading & Vocabulary
B-index means	0.21	0.07	0.20

Agreement statistic. Brown and Hudson (2002) present other statistics as measures of the cut-off score, the agreement statistic (A) was calculated by the following formula:

$$A = 2P_{iT} + Q_i - P_T$$

where:

P_{iT} = proportion of total examinees who answered the item correctly and passed the test

Q_i = proportion of examinees who answered the item incorrectly

P_T = proportion of examinees who passed test

(Brown & Hudson 2002, p. 125)

The values of the agreement statistic for each item are summarized in Table 8 with other statistical measures. In Table 6, the number of the items and their percentages for each value range are shown. As in the table, more than half of the items have values higher than 0.5. Moreover, it is noticeable that most of the items have higher values compared to the B-index results. The reason for this difference will be discussed in the later section.

Table 6. Distribution of the items by the agreement statistic

	0.2<A<0.3	0.3<A<0.4	0.4<A<0.5	0.5<A<0.6	0.6<A<0.7	0.7<A
No. of item	6	15	13	18	16	2
Percentage	8.57	21.43	18.57	25.71	22.86	2.86

Item phi (ϕ). The index called the item phi was also calculated. According to Brown and Hudson (2002), this index is essentially a Pearson correlation between examinees, items and test performance outcomes. In other words, it examines the relation between the mastery of the item and the mastery of the test. The agreement statistic (A) was calculated by the following formula:

$$\phi = (P_{iT} - P_i P_{iT}) / SQRT(P_i Q_i P_T Q_T)$$

where:

P_T = proportion of examinees who passed test

Q_i = proportion of examinees who answered the item incorrectly

P_T = proportion of examinees who passed test

Q_T = proportion of examinees who failed the test, or (1- P_T)

P_{iT} = proportion of total examinees who answered the item correctly and passed the test

(Brown & Hudson 2002, p.126)

Table 7. Distribution of the items by the item phi (ϕ)

	$\phi < 0$	$0 < \phi < 0.1$	$0.1 < \phi < 0.2$	$0.2 < \phi < 0.3$	$0.3 < \phi$
No. of item	2	19	30	17	2
Percentage	2.86	27.14	42.86	24.29	2.86

Table 8. Summarized table of the indexes in CRT item analyses

Item	B	A	ϕ	Item	B-index	A	ϕ
No.1	0.29	0.48	0.25	No.36	-0.02	0.56	-0.02
No.2	0.33	0.54	0.27	No.37	0.04	0.72	0.05
No.3	0.15	0.34	0.17	No.38	0.00	0.44	0.00
No.4	0.29	0.52	0.24	No.39	0.16	0.52	0.13
No.5	0.09	0.30	0.11	No.40	0.05	0.33	0.05
No.6	0.16	0.35	0.17	No.41	0.26	0.65	0.21

No.7	0.17	0.38	0.17	No.42	0.03	0.36	0.03
No.8	0.31	0.46	0.28	No.43	0.00	0.69	0.00
No.9	0.14	0.65	0.12	No.44	0.01	0.59	0.01
No.10	0.22	0.43	0.20	No.45	0.15	0.62	0.12
No.11	0.28	0.48	0.24	No.46	0.01	0.52	0.01
No.12	0.07	0.25	0.12	No.47	0.11	0.69	0.10
No.13	0.18	0.39	0.18	No.48	0.21	0.46	0.18
No.14	0.28	0.57	0.23	No.49	0.13	0.46	0.11
No.15	0.11	0.30	0.14	No.50	0.20	0.40	0.19
No.16	0.17	0.46	0.14	No.51	0.24	0.63	0.20
No.17	0.15	0.36	0.15	No.52	0.20	0.41	0.19
No.18	0.31	0.55	0.25	No.53	0.20	0.59	0.16
No.19	0.36	0.62	0.29	No.54	0.08	0.27	0.13
No.20	0.07	0.37	0.07	No.55	0.21	0.46	0.18
No.21	0.31	0.63	0.25	No.56	0.39	0.57	0.32
No.22	0.24	0.64	0.20	No.57	0.13	0.38	0.13
No.23	0.05	0.32	0.05	No.58	0.09	0.29	0.12
No.24	0.08	0.48	0.06	No.59	0.07	0.25	0.12
No.25	0.03	0.55	0.02	No.60	0.14	0.35	0.15
No.26	0.14	0.56	0.11	No.61	0.04	0.25	0.06
No.27	0.10	0.61	0.08	No.62	0.15	0.36	0.16
No.28	0.12	0.54	0.09	No.63	0.20	0.50	0.16
No.29	-0.04	0.66	-0.04	No.64	0.20	0.53	0.16
No.30	0.11	0.51	0.08	No.65	0.35	0.64	0.28
No.31	0.18	0.64	0.15	No.66	0.33	0.52	0.28
No.32	0.03	0.42	0.03	No.67	0.09	0.27	0.14
No.33	0.04	0.32	0.04	No.68	0.35	0.51	0.30
No.34	0.08	0.68	0.07	No.69	0.35	0.60	0.28
No.35	0.06	0.72	0.06	No.70	0.26	0.63	0.21

The values of the item phi for each item are summarized in Table 8 with other statistical measures. In Table 7, the number of the items and their percentages for each value range are shown. If we look at Table 7, we can notice that some of the items show

low values as they were in the B-index. Moreover, if we look at Table 8, we can easily see that the values are very different by indexes. For example, item no. 1 shows .29 in the B-index, .48 in the agreement statistic, and .27 in the item phi. Why are they so different? What do these differences tell us?

First of all, the values of the B-index are similar to those of the item phi. Brown and Hudson (2002) present the same phenomena. According to them, the values obtained for these two statistics are similar in most cases. What does this mean? Considering the B-index is an item statistic based on the differences in the item facilities of those students who passed the test as opposed to those who failed it and that the item phi is basically a Pearson correlation between examinee item and test performance, it is possible to say that both of them look at the item and masters/non-masters relationship, and this similarity of the values sounds natural.

How about the differences between the B-index and the agreement statistic? As in Table 8, the values obtained for the agreement statistic are quite different from those obtained using the B-index. According to Hudson and Brown (2002), this is because the B-index indicates the degree to which an item distinguishes between the students who passed the test and those who failed, while the agreement statistic indicates the degree to which those answering the item correctly are the same as those who passed the test. In other words, while the B-index and the item phi examine the relationships between the items (test) and masters/non-masters decision, the agreement index looks at that of the items (test) and the masters. Therefore, in this version of the test, some students answer some items right consistently and pass the test, but when it comes to distinguishing between masters and non-masters, the decision was not always consistent.

The reason for this may be the fact that the content of G-TELP does not always match with what was taught in the class. Considering that all the items analyses shown in the present study use the data from the posttest, we probably cannot observe the students' improvement in a clear-cut result because the content of the test does not match what has been taught. This might especially because of the items in the listening section. As is summarized in Table 8, the mean values for each statistic are especially different just for the listening section. Though two other sections show higher values for the B-index and the item phi, the listening section shows a higher value for the agreement statistics, which implies this strange mismatch tends to be rather prominent in the listening section. In sum, this result suggests that it is necessary to check if what is tested matches what has been taught, especially in the listening section.

Table 9. Summary table of the means of the statistics by section

Sections	B	A	ϕ
Grammar	0.21	0.46	0.19
Listening	0.07	0.54	0.06
Reading & Vocabulary	0.20	0.46	0.18

Conclusion

I would like to draw my conclusions by answering the questions posed in the purposes of the present study.

To what extent G-TELP plays its role as the CRT?

As long as we could observe the negatively skewed distribution, we can say that it is ideal for the CRT posttest. However, in order to obtain the most ideal pattern, the pretest should be positively skewed, but this might be because G-TELP is a commercialized test and does not exactly reflect what has been taught in the class.

Do different methods of analyses bring different results? Do they not?

As it was seen in Table 8, three item statistics, the B-index, the agreement statistic, and the item phi showed different values. Especially, the agreement statistic showed different values though the B-index and the item phi showed rather similar results.

What do the differences/similarities of the results tell us? What are the implications for the further use of the test?

The differences and similarities among the indexes suggested that the agreement statistic looks at the relationship betwwen the items (test) and the masters while the B-index and the item phi examine that of the items (test) and masters/non-masters decision. This result implied the possibility that the content of the test might not match what has been taught in the classrooms. The analysis by section suggests that this tendency is especially prominent in listening section though further analyses are required.

Implications for the further improvements

The present study showed the process of basic analyses of CRT items. As a result, we could obtain some meaningful insights on CRT items. As it was mentioned in the conclusion, in order to pursue the findings, a closer content analysis on the test items and the classroom instruction is necessary. As Popham (1999) states, studying test items will provide significant progress for the further study.

Furthermore, the analyses demonstrated in the study are only on CRT item analyses using the classical test theory. The author can introduce the method using item response theory (IRT) in the analyses as is also done in Brown and Hudson (2002). Moreover, the analyses should be continued on the validity, reliability of the test.

Note

1. Although G-TELP is a CRT, however, since G-TELP is a commercialized, its data might show some characteristics of NRT. As Popham (1977) pointed out, commercial CRT tends to have vague and general objectives, and since its descriptions do not always match the course objectives, the ideal CRT would be the customized ones. However, in this case, this nature of commercialized CRT might allow us to use G-TELP as a placement to some extent. For the basic analyses of G-TELP as a placement data, please refer to Tsuchihira and Kumazawa (2003).

References

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

Brown, J. D, & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Cartier, F. A. (1968). Criterion-referenced testing of language skills. *TESOL Quarterly, 2(1)*, 27-32.

Cziko, G. A. (1982). Improving the psychometric, criterion-referenced, and practice qualities of integrative language tests. *TESOL Quarterly, 16(3)*, 367-379.

Lynch, B. K., & Davidson, F. (1994). Criterion-referenced language test development: linking curricula, teachers, and tests. *TESOL Quarterly, 28(4)*, 727-743.

Popham, W. J. (1977). Customized criterion-referenced tests. *Educational Leadership, 34(4)*, 258-259.

Popham, W. J. (1999). Why standardized test don't measure educational quality. *Educational Leadership, 56(6)*, 8-15.

Tabachnick, B. G.., & Fidell, L. S. (2001). *Using multivariate statistics (4th ed.)*. Needham Heights, MA: Allyn & Bacon.

土平泰子・熊澤孝昭 (2003).「G－TELPを用いた総合英語プログラムの評価」『茨城大学人文学部紀要（コミュニケーション学科論集), 第14号, 47－71.