Revision of a Criterion-Referenced Vocabulary Test Using Generalizability Theory

Takaaki Kumazawa
Kanto Gakuin University

Classical test theory (CTT) has been widely used to estimate the reliability of measurements. Generalizability theory (G theory), an extension of CTT, is a powerful statistical procedure, particularly useful for performance testing, because it enables estimating the percentages of persons variance and multiple sources of error variance. This study focuses on a generalizability study (G study) conducted to investigate such variance components for a paper-pencil multiple-choice vocabulary test used as a diagnostic pretest. Further, a decision study (D study) was conducted to compute the generalizability coefficient (G coefficient) for absolute decisions. The results of the G and D studies indicated that 46% of the total variance was due to the items effect; further, the G coefficient for absolute decisions was low.

古典的テスト理論は尺度の信頼性を測定するため広く用いられている。古典的テスト理論の応用である一般化可能性理論（G理論）は特にパフォーマンステストにおいて有効な分析手法であり、受験者と誤差の要因となる分散成分の割合を測定することができる。本研究では診断テストとして用いられた多岐選択式語彙テストの分散成分を測定するため一般化可能性研究（G研究）を行った。さらに、決定研究（D研究）では絶対評価に用いる一般化可能性係数を算出した。G研究とD研究の結果、項目の分散成分が全体の分散の46%を占め、また信頼度指数は高くなかった。

Keywords: G theory, G study, D study, reliability, criterion-referenced test, diagnostic testing

Classical Test Theory

Classical test theory (CTT) is based on the theoretical foundation that an observed test score is conceptually composed of true score variance and error variance.[1] In other words, the test score variance includes the examinees' true abilities for a target construct, which the test is designed to measure, and measurement error, which creates noise in the testing. The underlying concept of the reliability theory states that if the test succeeds in spreading the examinees' test scores relatively along a continuum or exhibits a large degree of variance, the reliability coefficient is likely to be high. Therefore, the test can be said to estimate their true ability with relative accuracy. That is, the observed test scores vary because the examinees behave differently on the target construct being measured, not because of random noise in the test (Strube, 2000).

Based on the theoretical foundation, actual mathematical formulas were developed to estimate reliability coefficients. The core of the reliability formula is derived by dividing the true score variance by the observed score variance.[2] The most widely reported reliability coefficient is the Cronbach alpha internal consistency reliability formula.[3] If the reliability of a measurement is found to be .80, it indicates that 80% of the observed test score variance represents the examinees' true abilities and 20% is the result of random error creating inconsistency in estimating the examinees' true scores. Such error may be caused by examinee carelessness, testwiseness. or other factors that can result in inconsistency (Brown, 1996). Reliability is also indicative of consistency. For example, sometimes we can draw the inference from a reliability estimate that approximately 80% of the time, the examinees' test scores will vary in the same ways even if they repeatedly take the same test.

In CTT, a standard error of measurement (*SEM*) value for the entire test can also be calculated to show a range within which examinees would probably score if they repeatedly took the same test. Based on the reliability coefficient and standard deviation derived from the test scores, the *SEM* is easy to estimate[4] and interpret. For instance, if the *SEM* was found to be 2.00 and a particular examinee's score was 50.00, the *SEM* indicates that the examinee's test scores would fall between 48.00 and 52.00 about 68% of the time, if the same test were taken repeatedly.

In short, reliability breaks down a set of observed test scores into true score and error variances. However, CTT can only deal with error variance as a single entity and therefore cannot deal with multifaceted sources of error variance. This CTT notion is rather simplistic and not maximally useful

because it is impossible to define the sources of error. In an actual testing situation, numerous facets—such as the number of tasks, passages, and raters—may cause measurement errors. Further, examinees may respond to such facets in complex ways. Therefore, when numerous facets are inherent in a testing situation, the sources of measurement error should be investigated cautiously.

Generalizability Theory

This section introduces the background of G theory and discusses its advantages over CTT. G theory, introduced by Cronbach, Rajaratnam, and Gleser (1963), was extended by Cronbach, Gleser, Nanda, and Rajaratnam (1972) and has been discussed in numerous books on psychological measurement (Brennan, 1983, 2001; Fyans, 1983; Shavelson & Webb, 1991; Strube, 2000; Suen, 1990; Thompson, 2003). This theory was developed as an extension of CTT to investigate the sources of variance in the facets of measurement and to generalize the universe score or true score in CTT obtained from one observation to numerous observations (Brown & Hudson, 2002).

One of the powerful features of G theory lies in the first phase of the investigation called a G study. The multifaceted nature of testing can be broken down into each of the relevant facets of variance, enabling the study of the degree to which the facet variances contribute to the total variance of the test scores. The facets to be examined will depend on the testing situation involved. In performance testing, typical facets include examinees' abilities, rater severities, item difficulties, and occasion difficulties. The variance components for each facet in a particular testing situation can be estimated using an analysis of variance (ANOVA) procedure.

Another advantage of G theory over CTT is that it provides a more adequate estimate of reliability for criterion-referenced tests (CRTs). In CTT, the variability of the test scores is often highly related to the reliability of the test. Since the purpose of norm-referenced tests (NRTs) is to spread examinees' test scores out along a continuum, such variance is appropriate for determining the reliability for NRTs. In contrast, with CRTs, the variance may be suppressed due to three main factors: (a) small sample size, (b) homogeneity of students' proficiency levels, and (c) negatively skewed distributions of test scores at the end of a course. In CRTs, the sample size is relatively small because a limited number of students take classroom-level tests. In a language program where placement tests are administered to create homogeneous classes, that homogeneity is likely to suppress the variance in test scores. Ideally, criterion-referenced items have to be developed

based on class content, such that if all the students learn all the content, they should all score 100% on the test. This can create a negatively skewed distribution that is perfectly logical and as a result suppress the variance. In sum, the CRT's purpose remains to estimate students' achievement in a specific domain. Thus, CTT reliability does not fit the purpose of estimating criterion-referenced measurement consistency; therefore, G theory should be applied to estimate dependability for CRTs. CRT dependability is analogous to NRT reliability in CTT (Brennan, 1980).

Another advantage of G theory over CTT is found in the second phase of the investigation, called a decision study (D study). In CTT, the Spearman-Brown prophecy formula[5] can be employed to estimate reliability with different numbers of items. However, this formula cannot deal with multi-faceted sources of error in a measurement. To estimate the dependability of CRTs in different parallel tests, the index has to be determined based on the multiple sources of error estimated in a G study. The result of a D study is extremely useful in deciding how to revise or redesign a CRT. For instance, let us assume that sections and items are the facets in a given testing situation. The D study allows for calculating the degree of dependability for different hypothetical scenarios, that is, based on different hypothetical numbers of sections and/or items. This constitutes the most practical application of G theory.

A G study should be carefully designed and conducted to investigate the variance components for facets in a given test. Depending upon the testing situation and the measurement design adopted, the study can be designed as crossed or nested and balanced or unbalanced. If all the levels of one facet are the same in the levels of another facet, the two facets are considered crossed. For example, if the five different categories (say Content, Organization, Grammar, Mechanics, and Vocabulary) are scored by three raters, the categories facet is said to be crossed with the raters facet. Alternatively, if all the levels of one facet are different within the levels of another facet, the first facet is said to be nested within the second one. For example, if 10 items in each of three subtests are all different (i.e., items 1-10 are in subtest A, items 11-20 in subtest B, and items 21-30 in subtest C), the items are said to be nested within the subtests.

If all levels of all facets have the same number of observations per facet the design is considered balanced. For example, if all three subtests have 10 items each, it is a balanced design. Conversely, if the levels of even one facet have unequal numbers of observations, the design is considered unbalanced. For instance, in a performance test, if three subtests have different numbers of items (say 8, 12, and 18), it is an unbalanced design.

Based on variance components that can be extracted using an ANOVA procedure in a G study, a G coefficient can be estimated. A G coefficient in G theory is analogous to a reliability coefficient in CTT. Therefore, a G coefficient for norm-referenced (i.e., relative) decisions for a G study design of *p X i* can be estimated by dividing the persons variance component by persons variance component plus persons-by-items interaction variance component (divided by the number of items).[6] True score variance in CTT is analogous to the variance component for persons in G theory, while error variance in CTT is analogous to the variance component for the persons-by-items interaction in G theory. Therefore, G theory is an extension of CTT, but G theory has the additional benefit of making possible the estimation of separate variance components for all possible facets in a testing situation. Under identical conditions, the magnitude of a Cronbach alpha reliability coefficient and G coefficient for relative decisions should be nearly equivalent.

However, G theory can also be used to help in making criterion-referenced (i.e., absolute) decisions based on the extent to which students have mastered a certain domain. In this case, the equation is slightly different from the equation for relative decisions: here, the persons variance component is divided by the persons variance component and items variance component (divided by the number of items) plus persons-by-items interaction variance component (divided by the number of items).[7] The difference between the equations for relative and absolute decisions lies in how error variance is defined. For relative decisions, in the present case, the error variance is defined as the persons-by-items variance component (divided by the number of items). However, in the equation for absolute decisions, error variance includes both the persons-by-items interaction component (divided by the number of items) and the items variance component (divided by the number of items). With NRTs, administrators aim to estimate an examinee's true ability relative to a norm using the test; therefore, the focus is on persons and the interaction of persons with items, and items variance itself is excluded from the equation. However, in CRTs, teachers aim to estimate students' mastery over the item content or domain; therefore, the items variance is included in the equation.

A D study is used to answer a "what-if" question in that it is used to estimate the expected G coefficients if the numbers of items or raters are set at various levels. In other words, a D study generalizes the expected G coefficients under different hypothetical scenarios based on the extracted variance components in the G study. The D study can be conducted by changing the number of items for either relative or absolute decisions. In CTT, after

estimating a reliability coefficient, the Spearman-Brown prophecy formula can be employed to estimate the expected reliability coefficient by increasing and decreasing the number of items in the equation. Although a D study is analogous to the Spearman-Brown prophecy formula, the former can only estimate reliability for changes in one facet (usually items). In contrast, a D study can estimate the expected G coefficients along one, two, or more facets (e.g., items, raters, subtests, and occasions) by setting different numbers of facets at the same time (Suen, 1990).

In the field of educational measurement, numerous articles have been published that apply G theory, particularly for performance testing (Brennan, 2000; Brennan, Gao, & Colton, 1995; Cronbach, Linn, Brennan, & Haertel, 1997). With regard to language testing, only a few books refer to G theory (Bachman, 1990, 1997, 2004; Brown & Hudson, 2002). Brown (1982) first applied G theory to ESP testing. Brown (1993) and Kunnan (1992) investigated CRTs' dependability and employed criterion-referenced item analyses. Lynch and McNamara (1998) applied G theory and the multi-faceted Rasch model to develop ESP speaking tests. They contrasted the two analytical techniques. Employing a large data set from TOEFL, Brown and Ross (1996) and Brown (1999) investigated variance components for the test takers' number, items, subsections, and nationalities. They discovered that the interaction effect caused the most error variance.

In Japanese contexts, few studies have applied G theory (e.g., Yamanishi, 2004). Apart from Griffee's study (1995), which demonstrates the design and evaluation of CRTs using criterion-referenced item analyses, no other study has analyzed teacher-made, criterion-referenced language tests.

Research Questions

In this study, a vocabulary test was developed for a particular class and criterion-referenced item analyses were conducted. What makes the study different is that the test's dependability was estimated by conducting a G study followed by a D study to investigate the optimal number of items and sections needed to achieve a certain magnitude of the G coefficient. In the process, the following two research questions were raised:

1. To what extent is the vocabulary test dependable in terms of the G coefficients for absolute decisions?
2. How many items and subsections are optimal to achieve a certain magnitude of the G coefficient for absolute decisions?

Method

Participants

One hundred thirty-one 1st-year university students enrolled in a required general English course majoring in literature, law, or economics at a high-ranking private university in the Kanto area participated in this study. Four reading and listening classes taught by two instructors were selected. Their goals included improving students' listening comprehension so that they could understand English instructions when taught by native or nonnative teachers in the institution and improving their reading skills and speed. An additional goal included vocabulary development. The teachers set the following goal for vocabulary development: to get approximately 70% of the multiple-choice items correct. The test was designed to gauge the extent to which students learned the receptive meaning of the target words that appeared in the assigned textbook. At the beginning of the first semester, all students were placed into homogeneous groups according to their level.

Materials

To estimate students' mastery of the vocabulary items in the assigned textbook—developed by the English program for a particular course—a vocabulary achievement test was designed and developed. Five chapters were randomly selected from 10 and the items were prepared. Five target words were also selected at random from each chapter in the process of preparing the items; that is, 5 items, from a total of 25, were nested within each section. A sentence identical to one given in the textbook was provided with an underlined target word. All the items were multiple-choice questions, and the choices were written in English. The students were required to choose the answer closest in meaning to the target word. A sample item is as follows:

1. The idea of the need for a common language across the world has become <u>prominent</u> in the twentieth century.

a. Important
b. Common
c. Nonsense
d. Problematic

In the above example, the distractors are *common*, *nonsense*, and *problematic*, which were selected from high-frequency or academic word lists. The test mainly estimates students' receptive knowledge and their ability to gauge meanings from a given context.

Procedure

First, all the textbook passages were scanned and digitalized, following which WordClassifier (Denies, 2004) was employed to classify all the words in the passages in the order of frequency. For each chapter, the target words were selected based on the results of the frequency count. Preceding the test development, test specifications were prepared to clarify the test's purpose and to set a sample test item. During the first week of class, the teachers clearly explained the syllabus, including its goals, objectives, and grading system. While explaining the grading system, they announced that two tests would be administered, at the beginning and end of the course. While the pretest encouraged the students to perform well, it did not affect the students' grades; however, the posttest score accounted for 15% of their final grades. After the procedure was explained, the test was administered in the second week of the second semester in 2005; this was a diagnostic test to gauge the students' knowledge of the target vocabulary items before instruction. The test scores were to be used for the pedagogical purpose of allowing teachers to focus on helping those students with low scores. For vocabulary instruction, the teachers presented a list of vocabulary words for every chapter and provided the Japanese translations and synonyms. An alternative test form was planned to be administered as an achievement posttest for the final assessment.

Analysis

All the items were dichotomously scored, with any missing data treated as an incorrect item. ITEMAN (Assessment Systems Corporation, 1996) was used for the descriptive statistics, distracter analysis, and norm-referenced item analyses such as item facility (IF), item discrimination (ID), and reliability. All the responses were entered into Excel spreadsheet format for conducting criterion-referenced item analyses such as the *B*-index, agreement statistic (A-statistic), and item phi (φ-index). The *B*-index indicates the degree to which a criterion-referenced item differentiates mastery from nonmastery students. The A-statistic indicates the degree to which students answering the item correctly are identical to those who passed the test (Brown & Hudson, 2002). The φ-index essentially refers to the correlation "between examinee item and test performance outcome, their mastery of the item to their mastery of the test" (Brown & Hudson, 2002, p. 126). These statistics are a family of cut-point indices. Based on the cut-point of the test, which was set at 70%, the students scoring higher or lower than 18 were identified as belonging to the mastery or nonmastery groups. XCalibre (As-

sessment Systems Corporation, 1995) is a software program based on the three-parameter logistic model belonging to item response theory. It was used to estimate the KR-21 reliability of the vocabulary test. This software's command file follows the same format as that followed by ITEMAN. Subsequently, GENOVA (Crick & Brennan, 1983) was used to conduct the generalizability and decision (G and D) studies. GENOVA enables users to conduct balanced design G and D studies for random and fixed effects. Here, the G study was a *p* X (*i: s*) balanced design. I treated sections in the textbook as a facet for investigating the extent to which sections variance contributed to the total variance. This design was adopted because five items were nested in each section. After extracting the variance components for all the effects, a D study was conducted to investigate the dependability of the test. Then, the results were processed in Excel spreadsheets.

Results

Table 1 provides the descriptive statistics. The mean of the vocabulary test was 12.37 out of 25; this is desirable because it reveals that the examinees have not yet mastered all the vocabulary words. However, it would have been more desirable if the mean had been lower with a positively skewed distribution. This would reveal that most of the examinees had little knowledge of the target words. Based on the Cronbach alpha and the KR-21, the reliability coefficients for the vocabulary test were found to be .64, indicating that the CRT spread out the examinees' abilities fairly well. Or put another way, the test consistently measured 64% of the examinees' abilities, with the remaining 36% occurring due to error. The *SEM* derived from the Cronbach alpha reliability coefficient was 2.29, indicating that approximately 68% of the time, the examinees' scores would remain in a band that was 2.29 points above or below their observed scores. However, the coefficients and *SEM* are mainly used for interpreting the NRTs' results.

Table 2 summarizes the item analyses. Despite the fact that the IF and ID statistics are norm-referenced item statistics, they provide insightful information for criterion-referenced items. For diagnostic tests, IF values should be low enough to enable students to participate in class and then perform well on achievement tests. For instance, the IFs for items 8 and 18 are extremely high at .80 and .81, indicating that most of the students had already learned the target words before instruction. The mean proportion for the correct items was .49, which is desirable for norm-referenced purposes; however, it would have been more desirable for diagnostic purposes if the value had been slightly lower. Apart from items 8, 18, and 21, which had high

Table 1. Descriptive Statistics for the Total Score

k	*N*	*M*	Variance	*SD*	Skew	Kurtosis	Min	Max	Alpha	*SEM*
25	131	12.37	14.39	3.79	-0.27	-0.45	4	21	0.64	2.29

Notes. Skew = skewness; Min = minimum; Max = maximum; Alpha = Cronbach alpha; *SEM* = standard error of measurement

or low IF values, the remaining items had a large degree of variation. The items with IF values above .50 tend to be negatively skewed; those with IF values below .50 tend to be positively skewed. Most of the items have negative kurtosis values, indicating a flat distribution. Except for items 6, 10, and 20, all ID values were above .20 with a mean ID of .32. In other words, the items discriminated among the examinees' abilities. Ten out of 131 students scoring above the set cut-point were identified as mastery students.

The values of ID and the *B*-index are quite different. In particular, although item 20 was a potential candidate for revision from a norm-referenced perspective, it was a suitable item from a criterion-referenced perspective. Notice that the values of *B*- and φ-indices were nearly equivalent. Items 1, 2, 6, and 25, which have low *B*-index and φ-index values appeared to be problematic. Notice also that the values of the *B*-index and A-statistic are quite different. The A-statistic indicates agreement between answering correctly or incorrectly and passing or failing the test, while the *B*-index indicates the items' capacity to differentiate between students who passed and failed the test. Although item 20 is inappropriate from a norm-referenced perspective, it is suitable from a criterion-referenced perspective because most students who passed the test got this item correct.

Table 3 shows that the items effect and interaction effect accounted for 46% and 52% of the variance, respectively, accounting together for 98% of the total variance. Therefore, the total variance was mainly due to items and interaction effects. The universe score or persons effect included only 2% of the total variance.

A D study was conducted by using the variance components extracted in the G study. In Table 4, the dependability of the vocabulary test with the five sections per five items (k = 25) was found to be .30, which is very low. If the test were to be revised to contain six sections per five items (k = 30), the dependability would be .34, a slight increase. Similarly, if the test were increased to six sections with ten items (k = 60), the dependability would increase by .21. This reveals that a lower number of items and sections results in unsatisfactory dependability.

Table 2. Criterion-Referenced Item Analyses

	Item	IF	Variance	Skew	Kurtosis	ID (Rpbi)	*B*-index	A-statistic	φ-index
1	Prominent	0.35	0.23	0.63	-1.63	0.28	-0.06	0.62	-0.03
2	Guarantee	0.66	0.22	-0.70	-1.53	0.51	0.15	0.38	0.08
3	Emergence	0.39	0.24	0.46	-1.82	0.27	0.34	0.64	0.18
4	Mutual	0.51	0.25	-0.05	-2.03	0.24	0.42	0.55	0.22
5	Diversity	0.68	0.22	-0.78	-1.42	0.46	0.35	0.40	0.20
6	Civilization	0.47	0.25	0.14	-2.01	0.19	0.04	0.53	0.02
7	Ethnicity	0.50	0.25	0.02	-2.03	0.34	0.22	0.53	0.12
8	Clash	0.80	0.16	-1.53	0.34	0.43	0.21	0.27	0.14
9	Scarce	0.55	0.25	-0.20	-1.99	0.36	0.27	0.50	0.14
10	Tremble	0.39	0.24	0.46	-1.82	0.18	0.34	0.64	0.18
11	Equivalent	0.27	0.20	1.02	-0.97	0.32	0.68	0.79	0.40
12	Clinging	0.67	0.22	-0.74	-1.48	0.37	0.25	0.39	0.14
13	Dwelling	0.41	0.24	0.36	-1.90	0.32	0.53	0.65	0.28
14	Excavation	0.38	0.24	0.49	-1.78	0.23	0.45	0.66	0.25
15	Glimpse	0.34	0.23	0.67	-1.58	0.20	0.49	0.70	0.28
16	Restrict	0.65	0.23	-0.63	-1.63	0.58	0.38	0.43	0.21
17	Intimate	0.60	0.24	-0.39	-1.87	0.33	0.22	0.45	0.12
18	Domestic	0.81	0.16	-1.59	0.54	0.47	0.21	0.27	0.14
19	Bury	0.37	0.24	0.53	-1.75	0.35	0.35	0.66	0.19
20	Gullible	0.50	0.25	-0.02	-2.03	0.45	0.21	0.53	0.11
21	Intimidate	0.15	0.12	2.04	2.19	-0.01	0.17	0.82	0.13
22	Distinct	0.29	0.21	0.94	-1.14	0.27	0.23	0.71	0.13
23	Substitute	0.56	0.25	-0.23	-1.98	0.34	0.15	0.47	0.08
24	Sophistication	0.59	0.24	-0.36	-1.90	0.32	0.23	0.46	0.12
25	Ignorance	0.48	0.25	0.08	-2.03	0.22	0.13	0.53	0.07
	M	0.49	0.23	0.02	-1.41	0.32	0.28	0.54	0.16

Note. Skew = skewness; Rpbi = point-biserial correlation

Table 3. Variance Components for the G Study

Source	Variance components	Standard error	Percentage
p	0.003534	0.001016	2%
s	0.000000*	0.006504	0%
i X s	0.091668	0.027881	46%
p X s	0.001102	0.001486	1%
p X i:s	0.104974	0.002910	52%
Total	0.201277		100%

*After Brennan, (1983, pp. 47–48), the negative variance component found for this facet was rounded to zero.

Table 4. Dependability for D Study

	Items									
Sections	1	2	3	4	5	6	7	8	9	10
3	0.05	0.10	0.14	0.17	0.21	0.24	0.27	0.29	0.32	0.34
4	0.07	0.12	0.17	0.22	0.26	0.29	0.33	0.36	0.38	0.41
5	0.08	0.15	0.21	0.26	0.30	0.34	0.38	0.41	0.43	0.46
6	0.10	0.18	0.24	0.30	0.34	0.38	0.42	0.45	0.48	0.51
7	0.11	0.20	0.27	0.33	0.38	0.42	0.46	0.49	0.52	0.54

	Items									
Sections	11	12	13	14	15	16	17	18	19	20
3	0.36	0.38	0.40	0.41	0.43	0.44	0.46	0.47	0.48	0.49
4	0.43	0.45	0.47	0.48	0.50	0.51	0.53	0.54	0.55	0.56
5	0.48	0.50	0.52	0.54	0.55	0.57	0.58	0.60	0.61	0.62
6	0.53	0.55	0.57	0.58	0.60	0.61	0.63	0.64	0.65	0.66
7	0.57	0.59	0.60	0.62	0.64	0.65	0.66	0.67	0.68	0.69

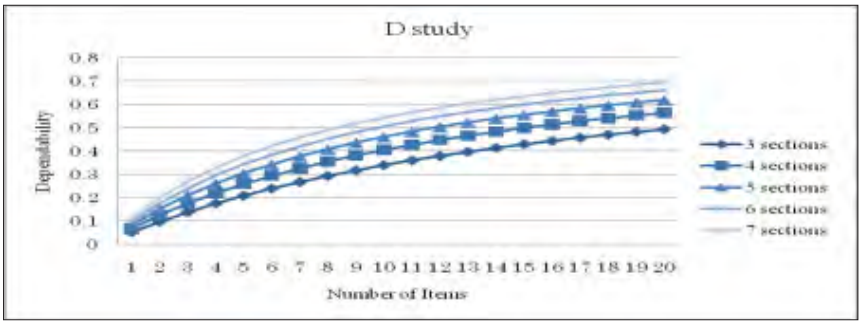

Figure 1. Graphic Representation of the Dependability for the D Study

Discussion

This section discusses the study's research questions, mainly based on the results of the G and D studies.

1. To what extent is the vocabulary test dependable in terms of the G coefficients for absolute decisions?

Two potential reasons for the lack of variability in the persons effect are: (a) sample size and (b) group homogeneity. Nationwide NRTs or placement tests are administered to numerous examinees; however, CRTs are usually administered to relatively small, homogeneous groups of students. The sample size in this study was 131, which is comparatively small from a norm-referenced perspective. Further, examinees with different backgrounds and proficiency levels take NRTs; however, a nearly homogeneous student group, similar in educational backgrounds and proficiency levels, take CRTs. In this study, the test was administered in four classes that two teachers were in charge of. With the exception of one class—identified as a high proficiency group based on a placement test—the proficiency levels of the classes were similar.

Because five items were nested within the corresponding chapters, a G study design of *p X (i:s)* had to be adopted. The results showed that no sections effect was observed. Some students studying only particular chapters of the assigned textbook and not the other chapters might yield sections variance in the posttest score. However, at this time, the students had not studied the textbook. Thus, it was reasonable that no sections variance was observed in this pretest because it did not test how many students had

learned the target words in each chapter. Another possibility was that this multiple-choice vocabulary test was context-independent (Read, 2000). That is, the examinees were able to answer the test items correctly without referring to the context or reading the embedded sentences.

The large variability in the items effect was an interesting result because, thus far, no studies have yielded a similar degree of variability. For NRTs, the persons effect should be large, whereas the items effect should be about one-third less than the persons effect (Brown & Ross, 1996). However, for CRTs, because students are homogeneous in terms of their proficiency level, the persons variance may be low. In addition, since a CRT should be based on items that measure a certain criterion or objective, the large amount of item variance found here may be desirable.

Due to low dependability, the items should be revised. Table 2 shows that the IF values range from .15 to .81. In a diagnostic test, it is desirable that the IF values be generally low, indicating that students have not yet learned the words. The IF values for items 8 and 18 were relatively high compared with the other items, and therefore, they should be excluded from the test. Further, based on the values of the *B*- and φ-indices, the items with low values should be revised. Items 1, 2, 6, 17, and 25 are the candidates for revision. For example, in item 1, students were made to choose the word closest in meaning to *prominent* from the following four choices: (a) important, (b) common, (c) nonsense, and (d) problematic. The correct response is (a). The distractor analysis reveals that the IF values for the four choices were .35, .19, .06, and .40 and the ID values were .28, .42, .03, and .08, respectively. Although ID is a norm-referenced statistic, it can prove useful during the revision of the items. The ID for choice (b) is higher than that for (a), indicating that students with higher scores are more likely to choose (b). The two words, important and common, are synonyms for *prominent*; therefore, both may be correct. However, students with high scores selected (b). Thus, (a) should be replaced with another word so as to function as a distractor and (b) should be the correct choice.

2. How many items and subsections are optimal for achieving a certain magnitude of the G coefficient for absolute decisions?

The Cronbach alpha was moderate, at .64. Although the purpose of this CRT was not to spread students' test score, there was moderate consistency in the test scores. As can be seen in Table 3, the dependability for this CRT was .30. Similar to the classroom tests analyzed in Brown (1993) and Kunnan (1992), this test, too, was not as dependable as expected. First, the G

coefficients for absolute decisions are generally lower than the Cronbach alphas and G coefficients for relative decisions (Brennan, 1980). Second, since these test scores did not affect their final grades, some students may not have taken the diagnostic test seriously; this was a low-stakes test.

Table 2 reveals that the variance component for sections effect was zero. In other words, adding another section to the test would have no effect on its dependability. The results of the D study are presented in Table 4; they reveal that increasing the number of items could contribute to the variability in the students' test scores and produce a higher dependability because a large variability was observed in the items effect. However, the administration time would be longer. In this testing situation, the test should not take over 20 minutes. While developing a test, teachers have to consider dependability and practicality. Finding the "happy medium" (Brown, 1996, p. 34) is the key for revising the test to ensure that it is dependable and practical. The maximum number of items that can be incorporated in the test are 40 because of time constraints in this testing situation. Otherwise, students would not be able to complete the test within the stipulated 20 minutes. If the test contained 40 items, then, based on the D study, the dependability would be .41.

Conclusion

In this study, G theory and criterion-referenced item analyses were applied to revise a CRT. While NRTs are used to spread examinees' test scores out, CRTs are designed to estimate students' mastery of specific objectives or language points. The *B*-index, A-statistic, and φ-index were used for the criterion-referenced item analyses; G theory was also applied to estimate the dependability of the domain score. In addition, the study showed that a G study can capture the multifaceted nature of testing by examining the degree to which the facets (sections and items nested within sections in this study) contribute to the total variance. A D study was applied to determine the optimal numbers of items and sections needed to make the test more dependable and practical in a revised version.

Before developing the test, it is crucial for teachers to thoroughly conceptualize its design in terms of purpose, content, procedure, target domains, number of items, sections, constraints, and analyses. Test specifications (a) are a good way to describe the design, (b) can guide test development, and (c) can serve as the basis for validity arguments to defend the diagnostic or achievement decisions that affect students' lives.

Often, preparing and marking a test is a cumbersome process that causes teachers to lose interest in analyzing their own tests. Teachers who neglect this procedure as part of teaching practice should recognize the importance of learning from the data. Sometimes, the expected result can differ completely from the actual results; therefore, the data analysis should be considered as part of good practices that confirm the extent to which expectations and results match.

The classroom tests must be developed before the actual teaching occurs so as to enable teachers to be aware of what is going to be tested; this will lead to the implementation of successful teaching-to-test instruction with the objective of maximizing students' achievement. Further, diagnostic tests are not often administered as part of teaching practices because the administration of tests takes up class time. However, the results can provide a great source of information, helping to identify misplaced students or mismatches between the students and the class objectives. In this study, 10 students scored higher than the stipulated cut-off based on the diagnostic test administered in the second semester. However, for reasons yet unknown, the students did not perform well in the placement test and were therefore not placed in the correct class levels. It is possible that they effectively learned vocabulary during the first semester or the summer vacation. If a large proportion of students scored above the cut-point, it is possible that the objectives were not set properly. Here, most of the students were nonmastery students; therefore, it was not necessary to change the materials or redesign the objectives.

The result of diagnostic tests can also be used for pedagogical purposes: to identify students' strong and weak points. The teaching should focus on the objectives that were not attained by students to enable them to achieve a high score on a posttest. In order to examine the score gain, it is recommended that the students' pretest scores be compared with their posttest scores. This procedure is termed intervention strategy (Brown, 2005). Study of the score gains can serve as empirical support showing that learning has taken place. Conversely, if gains are not observed for certain objectives, teachers should reconsider their teaching plans to better enable effective learning.

Five different kinds of software were used in this study. Apart from GENOVA, the other four software programs are quite user-friendly. The teachers can refer to the output to confirm whether their experience-derived teaching is suitable for the actual outcome of the teaching. Although this requires hard work, it is definitely beneficial in terms of improving teaching.

Two limitations are inherent in this study. First, further investigation is required to determine which criterion-referenced, multiple-choice vocabu-

lary test items are valid. Second, replication studies should be conducted to investigate how the magnitude of the G coefficient for absolute decisions in criterion-reference language tests would change in different testing situations. In spite of the fact that CRTs are frequently used by many teachers, studies on CRTs are rarely conducted. Additional studies on this issue are needed of other language programs in Japanese university contexts to reveal ways to prepare dependable and valid CRTs.

Acknowledgement

I would like to thank Dr. J. D. Brown for his invaluable lessons on language testing. I am also grateful to H. P. L. Molloy, two anonymous reviewers, and the editor of *JALT Journal* for reading earlier drafts of this paper and providing me with insightful comments. I am solely responsible for any remaining errors.

References

Assessment Systems Corporation (1995). *XCalibre* (version 1.1e) [computer software]. St. Paul, MN: Assessment Systems Corporation.

Assessment Systems Corporation (1996). *ITEMAN* (version 3.6) [computer software]. St. Paul, MN: Assessment Systems Corporation.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (1997). Generalizability theory. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 255-262). Norwell, MA: Kluwer Academic Publishers.

Bachman, L. F. (2004). *Statistical analysis for language assessment*. Cambridge: Cambridge University Press.

Brennan, R. L. (1980). Applications of generalizability theory. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of art* (pp. 186-232). Baltimore, MD: Johns Hopkins University Press.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339-354.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analysis of work keys listening and writing tests. *Educational and Psychological Measurement, 55*(2), 157-176.

Brown, J. D. (1982). *Testing EFL reading comprehension in engineering English*. Unpublished doctoral dissertation, University of California, Los Angeles.

Brown, J. D. (1993). A comprehensive criterion-referenced testing project. In D. Douglas, & C. Chapelle (Eds.), *A new decade of language testing: Collaboration and cooperation* (pp. 163-184). Ann Arbor, MI: University of Michigan.

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

Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TEOFL test variance. *Language Testing, 16*(2), 217-238.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York: McGraw-Hill.

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

Brown, J. D., & Ross, J. A. (1996). Decision dependability of subtests, tests and the overall TOEFL test battery. In M. Milanovic, & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 231-265). Cambridge: Cambridge University Press.

Crick, J. E., & Brennan, R. L. (1983). *GENOVA* (version 2.1) [computer software]. Iowa City, IA: The American College Testing Program.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*, 137–163.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*(3), 373-399.

Denies, J. (2004). *WordClassifier* (version 2.5). [computer software]. Belgium: Michael Goethals and EET Project Team.

Fyans, J. L. (Ed.). (1983). *Generalizability theory: Inferences and practical application*. San Francisco: Jossey-Bass Inc.

Griffee, D. (1995). Criterion-referenced test construction and evaluation. In J. D. Brown & S. Yamashita (Eds.), *Language testing in Japan* (pp. 20-28). Tokyo: The Japan Association for Language Teaching.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analyses. *Language Testing*, *9*(1), 30-49.

Lynch, B. K., & McNamara, T. F. (1998). Using G theory and many facet Rasch measurement in the development of performance assessment of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Strube, M. J. (2000). Reliability and Generalizability theory. In L. G. Grim, & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 23-66). Washington, DC: American Psychological Association.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thompson, B. (2003). A brief introduction to Generalizability theory. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 43-58). Thousand Oaks, CA: Sage.

Yamanishi, H. (2004). How are high school students' free compositions evaluated by teachers and teacher candidates: A comparative analysis between analytic and holistic rating scales. *JALT Journal, 26*(2), 187-205.

Endnotes

1. $X = T + E$
   where: X = observed score
   T = true score
   E = error

2. $r_{xx} = \sigma^2(t) / (\sigma^2(t) + \sigma^2(e))$
   where: r_{xx} = reliability
   $\sigma^2(t)$ = true score variance
   $(\sigma^2(t) + \sigma^2(e))$ = observed score variance

3. $\alpha = (k/k{-}1)\ (1{-}\sum\sigma^2(i)/(\sigma^2(t) + \sigma^2(e))$
   where: α = Cronbach alpha internal consistency reliability
   k = number of items
   $\sum\sigma^2(i)$ = sum of items variance
   $(\sigma^2(t) + \sigma^2(e))$ = observed score variance

4. $SEM = SD_x\sqrt{}\ (1 - r_{xx})$
   where: SEM = standard error of measurement
   SD_x = standard deviation of the test score
   r_{xx} = reliability

5. $r_{kk} = kr_{xx}/(1 + (k - 1)\ r_{xx})$
   where: r_{kk} = estimated reliability when the multiple of test items is set at k
   k = number of items
   r_{xx} = reliability

6. $E\rho^2(\delta) = \sigma^2(p)/(\sigma^2(p) + \sigma^2(pi)/ni)$
   where: $E\rho^2(\delta)$ = G coefficient for relative decisions
   $\sigma^2(p)$ = persons variance
   $\sigma^2(pi)$ = persons-by-items interaction
   ni = number of items

7. $E\rho^2(\Delta) = \sigma^2(p)/(\sigma^2(p) + (\sigma^2(i)/ni) + (\sigma^2(pi)/ni))$
   where: $E\rho^2(\Delta)$ = G coefficient for absolute decisions
   $\sigma^2(p)$ = persons variance
   $\sigma^2(i)$ = items variance
   $\sigma^2(pi)$ = persons-by-items interaction
   ni = number of items