SYSTEMATIC CRITERION-REFERENCED TEST DEVELOPMENT IN AN ENGLISH-LANGUAGE PROGRAM

A Dissertation
Submitted
to the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree of
Doctor of Education

By
Takaaki Kumazawa

Janurary, 2011

Examining Committee Members

David Beglar, Advisory Chair, CITE/TESOL
Marshal Childs, CITE/TESOL
James Dean Brown, External Member, University of Hawai'i at Manoa
James Sick, External Member, J. F. Oberlin University
Edward Schaefer, External Member, Ochanomizu University

ABSTRACT

Systematic Criterion-Referenced Test Development in an English-language Program
Takaaki Kumazawa
Doctor of Education
Temple University Japan, 2011
Doctoral Advisory Committee Chairperson: Dr. David Beglar

Although classroom assessment is one of the most frequent practices carried out by teachers in all educational programs, limited research has been conducted to investigate the dependability and validity of criterion-referenced tests (CRTs). The main purpose of this study is to develop a criterion-referenced test for first-year Japanese university students in a general English program. To this end, four research questions are formulated: (a) To what extent do the criterion-referenced items function effectively?; (b) To what extent do the facets of persons, items, sections, classes, and subtests contribute to the total score variation in two CRT forms?; (c) To what extent are two CRT forms dependable when administered as pretests and posttests?; and (d) To what extent are two CRT forms valid when administered as pretests and posttests? Two CRT forms made up of vocabulary (k = 25), listening (k = 20), and reading (k = 25) subtests were administered to 249 students using a counterbalanced design. Criterion-referenced item analyses showed that most items were working well for criterion-referenced purposes. Both univariate and multivariate generalizability studies indicated that the most of the variance was accounted for by the interaction effect, followed by the items effect, and then by the persons effect. FACETS analyses showed the separation for all the facets accounted for in the analyses and showed that item separation was greater than person separation. This indicated that the students' ability estimates were similar due to their having taken a placement test, whose results were used to form proficiency-based classes. Both univariate and multivariate decision studies indicated that the CRT forms were

moderately to highly dependable. The content validity of the CRT forms was supported because the test content was strongly linked to what was taught in class. The construct validity was supported mainly because a fair amount of score gain was observed. This study elucidates how the statistical analyses used in this study can be applied to CRT development, and how CRT development can be carried out as part of curriculum development.

ACKNOWLEDGEMENTS

I could not have completed this dissertation without support from a number of people. First, I would like to thank Dr. David Beglar for his patience and support for carefully going over my manuscript a number of times and giving me insightful comments. Second, I would like to thank Dr. J. D. Brown for getting me interested in language testing and guiding me through this dissertation process. Third, I would also like to thank the other members of my dissertation committee, Drs. James Sick, Edward Schaefer, and Marshal Childs, for the time they took to read this lengthy study and for giving me many helpful comments. I would especially like to thank Dr. Sick for his help on my analyses. I am also thankful to Dr. Schaefer, the former director of the TESOL program for his constant encouragement and being on my proposal defense committee. Without his support I could not have gotten through the M.Ed. program. I would like to thank professors Akiko Kawasaki and Kazuo Ichinose for allowing me to conduct this project in their institution, and I am grateful to the teachers and students who participated in this project. I would also like to thank my cohort-mates for going through the demanding coursework with me. Finally, my special thank goes to my wife, Masako, and my daughters, Sae and Kie, for their patience and support during this entire busy and *taihen* process of this dissertation writing. Without them, this dissertation would not have been completed.

DEDICATION

This dissertation is dedicated to my father and mother who got me into the field of education.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The Background of the Issue

An educational curriculum should aim to maximize students' learning. In order to help students achieve a certain proficiency level efficiently in a foreign language program, it is necessary that the curriculum be coordinated with a well-defined system. In the systematic curriculum model (Brown, 1995), administrators and teachers design the curriculum on the basis of all the information obtained from a needs analysis, goal/objective setting, testing, material development, teaching, and program evaluation. In other words, they arrange all these components coherently in order to maximize learning based on a clearly defined system.

In the process of curriculum development, criterion-referenced tests (CRTs) play a crucial role in linking all parts of the curriculum. This occurs because CRTs enhance students' achievement by ensuring that they study the course materials and teachers teach to the tests based on the materials. CRTs have to be developed systematically in order to assess students' achievement and make valid achievement decisions. Therefore, developing dependable and valid CRTs is a crucial component in systematic curriculum development.

Statement of the Problem

The language testing literature indicates that despite the crucial role played by CRTs in curricula designed to maximize students' learning, few studies have been conducted on CRT development (e.g., Brown, 1993; 2001; Kumazawa, 2009). In the Japanese context, there have been criticisms of relative evaluation (e.g., Matsuzawa,

2002), which is defined as a form of assessment in which each examinee's score is determined by comparing it to others examinees' scores. The Ministry of Education announced plans to replace relative evaluation with absolute evaluation in which each examinee's score is based on a set criterion as part of the educational innovations implemented in Japanese secondary schools in 1989 (Sasaki, 2008). Since then, secondary school teachers have been urged to be accountable for classroom tests and the grading system; consequently, teachers became involved in this issue as they were responsible for making decisions about their students' grades and creating classroom tests each time they taught a course. However, the debate surrounding criterion-referenced testing died down after a few years and it is rarely discussed at present (Negishi, 2005). Currently, it is difficult to know how much criterion-referenced testing is used in Japanese secondary schools due to a lack of published studies.

In Japanese higher education, criterion-referenced testing has not been fully introduced for a number of reasons, possibly including: (a) teachers' ignorance of testing, (b) testing as a ritual, (c) the lack of guidelines for testing, and (d) no set goals or curriculum. Most university teachers do not have to take testing classes; as a result, they lack the basic knowledge and skills necessary to statistically analyze criterion-referenced items and develop sound CRTs, and their ignorance of sound testing practices can lead them to be relatively unconcerned about their testing practices (Wakabayashi & Negishi, 1993). Some university teachers appear to believe that testing is simply a "ritual" or formality that they have to perform in order to give credits to the students. This thinking is encouraged because Japanese universities do not have guidelines that teachers have to follow in developing and administering classroom tests. Moreover, most foreign language programs in Japanese universities

are not well-coordinated and do not have clearly stated goals nor a coherent curriculum. Some call these programs “English for no particular purposes.” To develop CRTs appropriately, teachers must determine what to test by developing specific course objectives. However, because there are no unified syllabuses, assessment guidelines, and coordinated curricula in most Japanese universities, teachers can develop CRTs for their classes, but they cannot develop sound CRTs that are specifically designed for their programs.

Due to the above reasons, the misuse of criterion-referenced testing often occurs in Japanese universities. For instance, some teachers administer achievement tests in which they ask the students to write a Japanese summary of English reading passages to test the students’ comprehension of the passages. A second example occurs when teachers give tests that have no relationship to the class content, so students cannot study for the tests. Finally, I have seen teachers use TOEFL items to estimate students’ mastery of certain listening materials taught in class.

My initial motivation for conducting this CRT development project goes back to my teaching experience at a university I worked for previously. The general English program was well-coordinated and unified in the sense that an in-house placement test was given to form a proficiency-based curriculum, teachers used the same textbooks to teach to the program goals, and the teachers administered the same achievement test at the end of the semester. Although this well-controlled, systematic curriculum helped make this one of the best university English programs in Japan, the unified achievement test was problematic because although it was supposed to be an achievement test, it was a norm-referenced test that had little connection to the class content. Both my students and I found the unified test demotivating because though I tried to teach well, it was difficult for my lower proficiency students to score well on

the test. A CRT would have been fairer and more motivating because even lower proficiency students who had studied hard could have gotten a high score on the unified test. All these issues concerning the misuses of tests should be addressed.

Purposes of the Study

The main purposes of this study are to investigate the extent to which: (a) the CRT items function effectively, (b) the facets inherent in criterion-referenced testing contribute to the total score variation, (c) the CRT forms used in this study are dependable when administered as pretests and posttests, and (d) those CRT forms are valid. In order to address these issues, two CRT forms were administered twice as pretests and posttests in a counterbalanced design. Criterion-referenced item analyses and many-faceted Rasch measurement were used to analyze the CRT forms. Univariate and multivariate generalizability theory (G theory) was applied to examine the sources of variance and to estimate the dependability of the CRT. The gain scores were also examined to argue for the validity of the tests. Based on the results, I discuss writing, piloting, and editing criterion-referenced items, administrating CRTs, performing statistical analyses of test results, and using the CRT results for program evaluation and curriculum innovation in a required English program in a Japanese university.

Significance of the Study

This study can contribute to the field of language testing in Japan for the following reasons. First, in the field of language testing, few studies have been conducted on CRT development as a part of curriculum development (e.g., Brown, 1993, 2001). Even fewer studies have been conducted in the context of EFL classes in

Japanese universities (e.g., Kumazawa, 2007b, 2009). The present study is the first empirical study designed to shed light on the development of a CRT in a Japanese university. Second, in the Japanese context, criterion-referenced testing is a growing concern, especially to teachers involved in secondary education because a large number of teachers do not properly use CRTs as part of their teaching practices. In some institutions norm-referenced tests are administered to estimate students' achievement; in these cases, only a small proportion of students can get an "A" and the test content often has little or no connection with the content being taught. This inefficiency can not only demotivate students but also decreases curriculum effectiveness. By demonstrating how CRT development can be part of overall curriculum development, the present study can serve as a means to show that CRT development can be a crucial component in foreign language programs.

The Audience for the Study

Several groups of people can benefit by this study: teachers, researchers, second language learners, and curriculum designers. The main audience for this study is teachers. Criterion-referenced testing is still not used effectively by most teachers, mainly due to their lack of training in classroom assessment. This study demonstrates how criterion-referenced testing can be successfully conducted, how sound CRTs can be developed as part of overall curriculum development, and how criterion-referenced item analyses can be used to develop sound criterion-referenced items. As a consequence of administering sound CRTs, teachers' effectiveness might increase as a result of better understanding their students' mastery of the course content and the degree to which specific objectives have been learned.

The second audience for this study is researchers interested in criterion-referenced testing or testing in general. Researchers might gain insights into how CRTs can be analyzed and interpreted. In the field of language testing, univariate and multivariate generalizability theory is still not often used to analyze criterion-referenced multiple-choice tests. In addition, one type of Rasch model, the many-faceted Rasch model, is not often used for the analysis of CRTs. However, by applying these analyses, I hope to demonstrate that the results can be informative in illuminating the strengths and weaknesses of CRTs.

Second language learners can more effectively learn the target language if criterion-referenced testing is carried out properly. Student motivation to learn the language can be enhanced when they know what to study for in terms of well-defined objectives, their non-mastery of the objectives before learning the class content, and their mastery of the objectives as a result of learning the class content.

For curriculum designers, this study shows that CRT development is a crucial component of curriculum that has close links with other components of the curriculum, such as course objectives and materials development. For instance, in a proficiency-based program, they can identify their students' needs and mastery of the objectives by conducting a needs analysis survey and by administering CRTs. The results they obtain can be informative in designing sound, defensible objectives for particular groups of students, deciding on textbooks, and teaching methods that should be used to achieve them. They can also learn to use NRTs and CRTs at appropriate occasions and for appropriate purposes.

Delimitations

CRTs are designed to test students' mastery of specific objectives for a particular group of students in a local program. In this study, two CRT forms were developed in the context of a particular reading and listening class in a required English program and administered to a limited number of students in order to test their mastery of the objectives. Thus, the findings in this study should be interpreted cautiously and not generalized too quickly to other educational contexts. In addition, the results should be generalized cautiously to groups of English language learners whose background and native language are not Japanese or whose English proficiency level is considerably higher or lower than the participants in this study.

The Organization of the Study

In Chapter 2, the literature on CRT development, empirical studies on criterion-referenced testing, and test theories are reviewed. The research questions are stated at the end of the chapter. In Chapter 3, the research context is described. In Chapter 4, the results of two pilot studies are reported to show how the CRT forms investigated in this study were developed. In Chapter 5, the participants, instrumentation, procedures, and analyses are described. In Chapter 6, the results of this study are reported. In Chapter 7, the research questions are discussed based on the results of this study. In Chapter 8, the limitations and implications of this study, future suggestions, and final conclusions are given.

CHAPTER 2

REVIEW OF THE LITERATURE

This chapter is made up of four sections. In the first section, the literature on designing and developing criterion-referenced tests is reviewed. In the second section, the results of empirical investigations of the dependability and validity of criterion-referenced tests are canvassed. In the third section, test theories such as classical test theory, generalizability theory, and item response theory are reviewed. In the fourth section, based on the literature review, problems with developing CRTs are identified, and in the fourth section, the purpose of the study and the research questions are presented.

Systematic Criterion-Referenced Test Development in Foreign Language Programs

Glaser (1963) is credited with distinguishing NRTs and CRTs in order to draw attention to the need for different families of tests for use in classroom settings. Popham and Husek (1969) defined NRTs as "those tests that are used to ascertain an individual's performance against the performances of other individuals using the same measuring device" (p. 2). This definition suggests that NRTs serve as a psychological measurement estimating examinees' scores by comparing their scores with those of other examinees in a distribution using statistical procedures such as *z*-scores. In other words, in NRTs, examinees' scores are determined on the basis of where their scores are located in the distribution. In this sense, it is a "norm" or group-of-examinees

referenced test because the interpretation of the test score is based on where a score lies in the normal distribution.

Popham and Husek (1969) defined the CRT as a psychological measurement device "used to ascertain an individual's status with respect to some criterion, i.e., performance standard. It is because the individual is compared with some established criterion, rather than other individuals, that these measures are described as criterion-referenced" (p. 2). With CRTs, decisions are made on the basis of students' test scores with reference to specific criteria. The term "criterion" has two connotations. It refers to a domain of a score or construct that is being measured using a test. It also refers to a set cut-point. Therefore, decisions are made based on the extent to which students master a domain, whether or not they exceed the set cut-point, or a combination of both.

Since Popham and Husek's (1969) study was published, criterion-referenced testing has become a debated topic in the field of educational measurement and variations of CRTs, such as domain-referenced tests, mastery-referenced tests, and objective-referenced tests, have been discussed in the literature (Brown & Hudson, 2002). In domain-referenced testing, items are a representative of a certain domain that is to be measured. In mastery-referenced testing, the results are used to make mastery decisions for specific instructional objectives that are related to certain domain. In objective-referenced testing, items are developed to assess specific objectives of a class. The other aspect of this debate concerns how to best analyze data gathered with CRTs. In the 1970s and 1980s, a number of studies were published concerning statistical procedures that can be used to analyze CRTs. This is because the conventional statistics used for examining NRTs are not entirely appropriate for analyzing CRTs (e.g., Berk, 1980).

In the field of language testing, the distinctions between NRTs and CRTs were first discussed by Cartier (1968). A number of authors have pointed out the differences between the two types of tests (Bachman, 1990, 2004; Brown, 1989, 1990a, 1996, 2005; Brown & Hudson, 2002; Genesee & Upshur, 1996; Lynch, 2003; Lynch & Davidson, 1997). Hudson and Lynch (1984) referred to Popham's (1978) definition and stated that the major difference between CRTs and NRTs is that CRTs are absolute and NRTs are relative. The term *absolute* is used to indicate that the examinees' test scores are used to ascertain whether they have surpassed a certain standard based on their performance demonstrated on a set of particular domains. On the other hand, *relative* refers to a decision that is made based on the examinees' test scores, which are compared to one another on the score distribution.

The scope of CRTs was further expanded by integrating criterion-referenced testing into curriculum development. Brown (1989, 1990a) pointed out the relationship between testing and curriculum development by discussing how testing can be used to help teachers and administrators make curricular decisions. NRTs and CRTs are two families of tests that are designed to arrive at program decisions and classroom decisions, respectively.

Test scores obtained from NRTs are used to make program decisions concerning examinees' entrance into a program and/or their placement in the program. For instance, some administrators are interested in gauging whether or not examinees are qualified or proficient enough to enter their programs on the basis of an international or nationwide general proficiency test like the TOEFL. They may also use examinees' test scores on a general proficiency test or an institutionalized test to place candidates into an optimal class level.

In contrast, classroom decisions are referred to as diagnostic and achievement decisions. CRTs are used to make classroom decisions in order to diagnose students' mastery or non-mastery of the objectives and to estimate their achievement in a particular class. Before actually teaching, classroom teachers should investigate the extent to which their students have already learned the targeted instructional objectives, using a diagnostic test to measure their degree of mastery of the course objectives. After instruction, teachers should use an achievement test to determine the students' achievement by measuring their mastery of the objectives as a result of the instruction.

Lynch and Davidson (1994) described five stages for criterion-referenced language test development (CRLTD): (a) selecting skills, (b) writing specifications, (c) writing item specifications, (d) assembling the test and piloting, and (e) finalizing operational measures. In the CRLTD process chart, the main focus is the use of test specifications as a means of linking curriculum developers, teachers, and language testers. Davidson and Lynch (2002) defined a test specification as "a generative blueprint from which test items or tasks can be produced" (p. 4). The purposes of developing a test specification are to: (a) provide the theoretical underpinning of CRTs, (b) set guidelines for developing test items, (c) provide a clear procedure for the test administration, (d) defend each item's content validity, and (e) set guidelines for developing parallel test forms. In designing a theoretical underpinning, teachers have to refer to the class syllabus in which the goals and objectives of the class are unambiguously identified. By referring to the syllabus, teachers can consider how adequately these objectives are measured with the CRT. The test specification can be a warrant for the content validity. To have high content validity, what is taught in class has to be closely related to what is stated and tested in the test specifications. Once a

well-designed test specification is developed, it can be used as a guideline to develop other parallel forms of the CRT.

The content of a test specification can be subsumed into the following components: (a) general description, (b) test design, (c) sample items, (d) prompt attributes, (e) response attributes, and (f) specification supplement (adapted from Brown & Hudson, 2002; Davidson & Lynch, 2002). The general description should include statements of the class goals and objectives and the purpose of testing. Additional information, such as the testing situation and students' background information, can also be added. Test designers should indicate the number of items, passages, dialogs, sections, and allocated time. Sample items and the directions for taking the item should also be provided. Prompt attributes include the details of the passages including the sources, readability, genre, topics, the number of words and paragraphs, and the types of item and the target construct. Response attributes define what students are expected to do on the test, how they are evaluated, the scoring procedures, and the scoring weights. The specification supplement should include additional information that can support item writers. Examples are vocabulary lists, the ways the items can be analyzed, and the types of feedback students receive. With all the information documented in a test specification, teachers can decide to adopt, adapt, or create CRTs that best suit their own purposes.

Bachman and Palmer (1996) divided test development into three phases: (a) design, (b) operationalization, and (c) administration. In the design phase, test designers define basic test elements such as the purpose of the test, the test-takers, and target constructs. In the operationalization phase, the designers develop a test specification or what they call a blueprint, which is the basis of test development. Test administration involves both test administration and analyses of the results. The

procedure outlined by Bachman and Palmer is a theoretical approach to test development in the sense that they emphasize the theoretical underpinnings supporting the design and operationalization of the test.

Wada (2005) divided test development into the following phases: (a) defining the purposes of the test, (b) developing a test specification, (c) developing instructional objectives, (d) developing test items, (e) administering a pilot test, (f) analyzing the items, (g) revising the test, (h) administering the test, (i) analyzing the test items, and (j) using the test results to make a decision. Wada's approach provides a complete picture of the test development procedure because phases (a)-(c) concern the theoretical underpinnings of test development, and phases (e)-(i) focus on test administration and item revision based on an analysis of item functioning. Test development must be based on a solid theoretical design, but necessary adjustments have to be made based on the results.

The Place of Criterion-Referenced Tests in Foreign Language Programs

In order to demonstrate how criterion-referenced test development is integrated into all the components in foreign language programs, Brown (1989, 1995) introduced the systematic curriculum development model (Figure 1). The purpose of this systematic curriculum model is to show how a curriculum can be systematically unified so that it is consistent, efficient, and effective in terms of maximizing students' learning.

According to this model, curriculum development is a continuous process that can be considered a set of phases: needs analysis, setting goals and objectives, test development, material development, teaching, and program evaluation. In the needs analysis phase, administrators conduct a needs analysis to obtain information on the

learning needs of the students enrolled in the program. After identifying the students' needs, administrators can set goals to satisfy their needs. Once the goals are set, administrators can decide on the types of tests that can be used to make proficiency and placement decisions. Administrators can then decide upon the materials that are suitable for their needs, proficiency levels, and the goals of the curriculum. Once the materials are designed, teaching can take place.

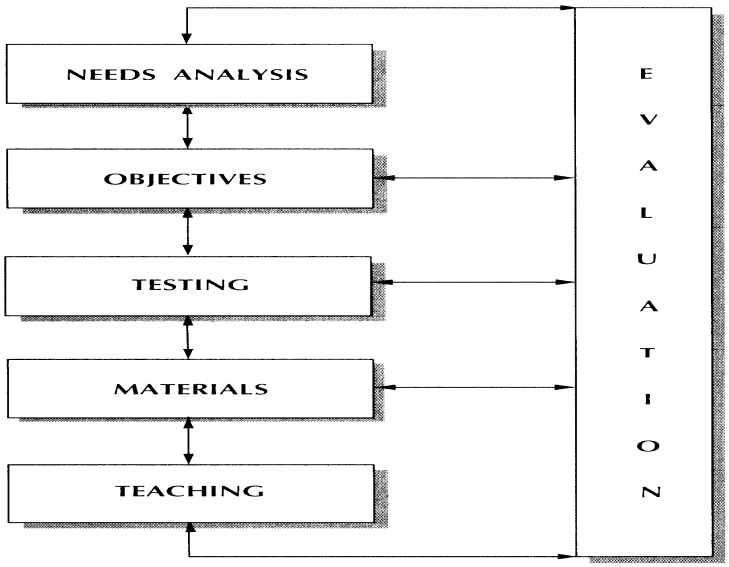

Figure 1. The systematic curriculum development model (Brown, 1995).

Likewise, teachers can conduct a needs analysis in order to obtain specific information about their students' learning needs and proficiency level. To best satisfy the students' needs, teachers can select objectives that are linked to the goals of the curriculum. Once the objectives are set, teachers can develop classroom tests and materials. Using previously developed tests and materials, they can administer the CRTs and teach the class based on the course objectives in order to maximize students' mastery of the objectives. Therefore, if there are no explicit goals or no

coordinated curriculum, teachers cannot develop classroom tests and make valid decisions regarding student achievement.

The testing phase in the systematic curriculum development model can be further divided into four phases: (a) CRT specification development, (b) CRT development, (c) CRT analysis, and (d) test revision. Once needs are identified in phase 1 and objectives are set and stated in the class syllabus in phase 2, teachers can develop a test specification in order to develop CRTs.

Based on a test specification developed for a particular testing purpose, there are three ways to develop CRTs: adopt, adapt, or create them. If the items in other tests match the purpose of a particular CRT, it is possible to adopt them. When items in other tests do not satisfy the purpose of the CRT, it may still be possible to adapt them. A number of tests for assessing language proficiency are available, but few proficiency tests can adequately estimate students' achievement in a particular language program because the test items are not designed to measure students' mastery of course objectives. As a result, in most cases, teachers have to create their own criterion-referenced items.

Once CRTs are developed in the testing phase in the model, materials can be developed. In the teaching phase in the model, CRTs are administered as a diagnostic and/or achievement test. Once the tests are administered in the teaching phase, teachers should conduct classical item analyses that include: (a) descriptive statistics such as the mean, standard deviation, minimum score, maximum score, and dependability; (b) item facility; (c) the *B*-index; and (d) the difference index. Descriptive statistics are necessary for understanding students' overall test scores and the number of students who achieve a certain cut-point. Brown (1990) proposed an easy-to-calculate formula to estimate the dependability of CRTs using a conventional

reliability index such as Kuder-Richardson coefficient 20 internal consistency formula (K-R20). Item facility and the *B*-index can be easily calculated to show item difficulty and how well the items differentiate mastery and non-mastery students. The difference index indicates how well students have learned the course objectives since the administration of the pretest, presumably as a result of instruction. Brown (2005) explained the procedures for criterion-referenced item analyses using EXCEL in a comprehensible way.

Conducting item analyses is necessary because the results allow test makers to improve the items, recycle items that function well, and improve their item-writing skills. Using the results of statistical analyses, criterion-referenced items can be improved by changing the wording of the correct choice and/or the distracters to make them more or less attractive. Items that work well should be recycled because it is extremely difficult to develop them. Because items that teachers think are effective may not function as anticipated, teachers need to refer to the results of the item analyses and consider possible reasons why the items do not perform as planned. It is also important that teachers carefully examine good items and develop items that are similar to them. By doing so, teachers can improve their item-writing skill.

None of the above components should be set in stone; instead, they should be implemented flexibly so that changes that improve the effectiveness of the program can be incorporated at any stage. It should be noted that in the systematic curriculum development model (Figure 1) the double headed arrows indicate the interrelatedness of the components. The purpose of the program evaluation is to collect information, examine each of the phases, and ensure that all the phases are functioning coherently and effectively (Brown, 1995, p. 24). Situations in which the phases are not functioning optimally should be targets for revision.

Studies of Criterion-Referenced Tests

In the field of language testing, Cziko (1982) carried out the first empirical study to examine how scoring procedures differ in the reliability of criterion-referenced integrative dictation tests. A total of 102 ESL university students took an audio-recorded 14-sentence dictation test. Three scoring procedures were adopted. In the first scoring procedure, each word was considered to be an item, and to score the item, the word had to be written with appropriate spelling. In the second procedure, each sentence was considered to be an item, and to score the item, the all the words in the sentence had to be written with appropriate spelling. The third procedure was similar to the second procedure except that all the words in the sentence had to be exactly correct. When the exact spelling procedure was adopted, the test was as reliable as the ones scored with the first and second procedures.

Hudson and Lynch (1984) conducted an empirical study concerning statistical procedures for analyzing CRTs. They stated that if students studied the materials, they would get a score of close to 100%; thus, a negatively skewed distribution is ideal on an achievement test. They then compared statistical procedures for NRTs and CRTs and introduced the difference index (DI) as a statistic that can be used to examine student achievement resulting from instruction. Instead of using the NRT concept of reliability, the notion of dependability was introduced for use with CRTs because the major concern is not the consistency of relative decisions but the consistency of absolute decisions. Hudson and Lynch demonstrated three types of statistics for estimating dependability: threshold loss agreement, squared-error loss agreement, and domain score dependability. Threshold loss agreement measures the consistency of the master/non-master classification using the agreement coefficient (p_0) and the kappa coefficient (κ). Squared-error loss agreement such as the phi lambda

dependability index ($\Phi(\lambda)$) indicates the distances of students from the cut-point for the master/non-master classification. Domain score dependability estimates the consistency of items purporting to measure a certain domain without reference to a cut-point.

Based on a test specification, they developed 84 listening, reading, and grammar items and administered them to two groups of ESL learners as an achievement test; 57 students were in the instructed group, and 42 students were in the uninstructed group. The DI values were all positive, indicating that the instructed group scored higher than the uninstructed one. They also applied generalizability theory (G theory) to conduct generalizability studies (G studies) with the design of persons crossed with items. No decision studies (D studies) were reported. The results showed that the largest amount of variance in the observed score resulted from interaction effects, indicating that a large number of unidentified errors contributed to the total test variance. The dependability indices ranged from low to moderate. Finally, in terms of validity, item-specification congruence and decision consistency were discussed. To build an argument for item-specification congruence, they had content experts rate how well item content matched the domains described in the test specification. The grammar sections had higher rating scores, but the reading sections had lower rating scores. Decision consistency, which was one argument for construct validity, was estimated by the proportion of examinees scoring above the cut-point in the instructed group plus the proportion of examinees scoring below the cut-point in the uninstructed group, divided by two. The values were moderate to high, ranging from .62 to .81. Thus, the construct validity of the test was supported. However, they argued that the CRTs were still in the development stage and required further refinements.

Brown (1989) pointed out the differences between NRTs and CRTs and argued that placement tests were NRTs that should spread examinees out along a continuum while being curriculum-specific. The purpose of his study was to develop a placement test that would match the curriculum for an ESL reading course. A placement test was administered to 194 ESL students and again given to 61 students after 15 weeks of instruction. He calculated both norm-referenced item statistics such as item facility and item discrimination as well as criterion-referenced item statistics such as DI in order to select 35 out of 60 reading items that were sensitive to curriculum content and that discriminated between higher and lower ability examinees. He reported the internal consistency K-R20 and found the test to be reliable at .89 with 60 items. The revised 35-item test was equally reliable at .85. To support the construct validity of the test, Brown reported the pretest/posttest gain score and found a statistically significant gain.

Brown (1990b) argued that conventional reliability formulas are not appropriate for use with achievement tests because low variance and negatively skewed distributions are expected. He then demonstrated how to calculate three types of criterion-referenced consistency estimates: threshold loss agreement, squared-error loss agreement, and domain score dependability. Agreement (p_0) and kappa coefficients (κ) (the threshold loss agreement approach), phi lambda dependability index ($\Phi(\lambda)$) (the squared-error loss agreement approach), and phi coefficient (Φ) (the domain score dependability approach) were calculated using data from a CRT administered in the intermediate ESL reading classes. A shortcut formula to estimate dependability was proposed:

$$\Phi = \frac{\frac{(nSDp^2)}{(n-1)}(K-R20)}{\frac{(nSDp^2)}{(n-1)}(K-R20) + \frac{(Mp(1-Mp)-SDp^2)}{(k-1)}}$$

Where n = number of test-takers

k = number of items

Mp = mean of proportion scores

SDp = the standard deviation of the proportion scores

K-R20 = K-R20 coefficient

The numerator is calculated by the number of test-takers times the standard deviation of the proportion scores squared, divided by the number of test-takers minus one, times the K-R20 coefficient or $(nSDp^2) / (n - 1)$ (K-R20). The denominator is calculated by the mean of proportion scores times one minus the mean of proportion scores, minus the standard deviation of the proportion scores squared, divided by the number of items minus one or $(Mp\ (1 - Mp) - SDp^2) / (k - 1)$. Further, what has been calculated as $(nSDp^2) / (n - 1)$ (K-R20) is added. Then, the numerator is divided by the denominator. The formula can be easily calculated with EXCEL, but it is only applicable to selected response items. The generalizability coefficient for absolute decisions with a *p X i* design and the short-cut dependability index should be nearly equivalent.

Hudson (1991) investigated the relationships among Rasch fit statistics, the two-parameter slope parameter, and point-biserial correlations based on data obtained from two CRT forms, the General Test of English Language Proficiency (GTELP), which was administered to 430 and 400 examinees. He reported the Cronbach alpha internal consistency reliability that was used for NRTs, and found all the subtests to be highly reliable, ranging from .84 to .95. Strong correlations were found among the point-biserial correlations, residual fit statistics, and slope parameter, so he argued

that the residual fit statistics could be used as a substitute for the slope parameter. He recommended using the two-parameter slope parameter over the Rasch fit statistics and point-biserial correlations. The main reason is that, although the steepness of the information curves in the Rasch model was the same, the curves in the two-parameter model indicated the item difficulty and discriminability near the cut-point. He also suggested that instead of including highly discriminating items, items with item difficulty estimates near the cut-point should be used on CRTs to arrive at more dependable pass/fail decisions because highly discriminating items with steep slopes only discriminate high and low scoring test-takers around a certain range of item difficulty; test-takers' responses to items near each cut-point are necessary to determine if they are above or below the standard.

Kunnan (1992) investigated the dependability and validity of CRTs that were used to make placement decisions with 390 non-native speakers of English in a language program at UCLA. The participants took a 100-item placement test that was comprised of 30 listening, 40 reading and vocabulary, and 30 grammar items. They also took a writing test. G theory, a principle components analysis (PCA), and a cluster analysis were used to estimate test dependability, provide arguments for construct validity, and identify the number of groups that could be formed on the basis of the cluster of the test scores. A wide range of variance was observed, which is a common feature of NRTs. Kunnan reported the norm-referenced K-R20 reliability coefficient to be .95. He also reported the reliability coefficients for four groups of examinees. The K-R20 reliability coefficient was high at .82 for the lower proficiency group (n = 71). He reported that the coefficient for the higher proficiency group comprising 73 examinees was -1.49 but the coefficient had to be rounded to .00. This indicated that the test was better able to discriminate the lower proficiency group.

Kunnan also reported that the threshold-loss agreement coefficients and the squared-error loss coefficients were high. He did not report the results of the G and D studies. It is not clear whether generalizability coefficients for relative or absolute decisions were reported, but the dependability index was high at .93. The results of the PCA showed that there were two underlying factors. The listening items loaded on one factor, and the reading and grammar items loaded on the other factor. Kunnan also reported the results of the PCA when different rotations were used and when the PCA was conducted with the four groups of test takers; however, it should be noted that the sample sizes for the three groups were less than 100. He concluded that because different results were obtained depending on the rotations and samples used in the PCA, the test was not unidimensional.

Hudson (1993) investigated surrogate indices for item information function (IIF) in criterion-referenced testing. He reported that the IIFs estimated with the two-parameter IRT model were the best indicators of criterion-referenced item discrimination indices when developing CRTs, but the parameter was problematic with small sample sizes and narrow distributions. He examined the relationship between item discrimination indices such as the phi-coefficient (Φ), the *B*-index, and the agreement statistic and the IIF, using the data obtained from three types of reading tests: a 50-item multiple-choice placement test given to 463 ESL students; a 34-item achievement test given to 216 ESL students; and a 25-item cloze test given to 206 ESP students. He concluded that the criterion-referenced item indices could be used as surrogates for the IIF.

Brown (1993) discussed the systematic curriculum development model (Figure 1) and explained how test development is part of curriculum development in a language program at the University of Hawai‘i. The teachers designed 14 CRT forms designed to test 294 students' mastery of the objectives for reading, writing, and listening classes and administered the CRTs at the beginning and end of the course using a counterbalanced design. The results of the norm-referenced statistics and IRT item difficulty estimates showed that most of the items were fairly easy for the students, even on the pretests. Based on the criterion-referenced item statistics, DI and differences in means showed that only slight score gains were observed. The *B*-index indicated that most items differentiated between the passing and failing students. Cronbach alpha internal consistency reliability coefficients ranged from .28 to .99, and the dependability indices for the CRTs used as pretests and posttests using his shortcut formula (Brown, 1990b) ranged from .24 to .71. The squared-error loss agreement coefficients were generally high. The content validity of the tests was supported because all the items were developed to match the course objectives. The construct validity of the test was also supported because score gains were observed in all the courses. Brown argued that developing CRTs to test course objectives enhances curriculum development.

In the Japanese context, only a few studies have been conducted on CRTs. Griffee (1995) administered an equivalent CRT form to 50 Japanese students in order to test the content of an assigned textbook at the beginning and end of the course. Out of the maximum score of 100, the students' scores increased from 52 points to 67 points. Based on the criterion-referenced item statistics, he concluded that it is difficult even for experienced teachers to write accurate, instruction-sensitive criterion-referenced items.

Brown (2001) developed four 45-item and four 55-item pilot CRT forms for a listening textbook series. Twenty-five teachers helped administer one of the eight forms to 1247 Japanese university students as a pretest and posttest. Only the results of students who took both the pretest and posttest were used to calculate both norm-referenced and criterion-referenced item statistics (N = 678). The dependability for the four 55-item forms given as a pretest and posttest ranged from .71 to .79. The mean differences between the pretest and posttest for three 45-item forms increased by about one or two points. However, the mean difference for one form decreased by one point. The dependability for the four 45-item forms given as a pretest and posttest ranged from .76 to .86. The mean differences between the pretest and posttest ranged from -3.00 to 8.02 points. The items that were not sensitive to the textbook instruction were discarded to make four 40-item CRT forms.

Kumazawa (2009) conducted norm-referenced and criterion-referenced item analyses such as item facility and the *B*-index for a 25-item paper-pencil multiple-choice vocabulary CRT used as a diagnostic pretest. The test was administered to 131 first-year Japanese university students who were intermediate EFL learners of English. He found that most of the items were working well as pretest items. In addition, he conducted a G study to investigate the percentages of persons, items, sections, and interaction variance components. Thus, the study design was persons crossed with items nested within sections balanced design or p X (i:s). Further, a D study was conducted to compute the G coefficients 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 at .30.

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:

$X = T + E$

where: X = observed score
T = true score
E = error

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 accounts for noise in the test. The underlying concept of reliability theory states that if the test succeeds in creating a large degree of variance in the examinees' test scores, the reliability coefficient is likely to be high. Therefore, the test can be said to estimate the test-takers' 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, 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:

$$r_{xx} = \frac{\sigma^2(t)}{\sigma^2(t) + \sigma^2(e)}$$

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

The most widely reported reliability coefficient is the Cronbach alpha internal consistency reliability formula:

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma \sigma^2(i)}{\sigma^2(t) + \sigma^2(e)}\right)$$

where: α = Cronbach alpha internal consistency reliability
k = number of items
$\Sigma \sigma^2(i)$ = sum of items variance
$\sigma^2(t) + \sigma^2(e)$ = observed score variance

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. If a test is said to have high test-retest consistency, observed scores on two administrations are parallel. For example, 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 and interpret:

$$SEM = SD_x\sqrt{(1 - r_{xx})}$$

where: *SEM* = standard error of measurement
SD_x = standard deviation of the test score
r_{xx} = reliability

For instance, if the *SEM* is 2.00 and a particular examinee's score is 50.00, the

SEM indicates that the examinees' test scores will fall between 48.00 and 52.00 about 68% of the time if the same test were administered repeatedly.

In short, reliability breaks down a set of observed test scores into true score and error. 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 specific sources of error using the traditional approach. In an actual testing situation, numerous sources of variance such as the number of tasks, passages, and raters can cause measurement error. Further, examinees can respond to such facets in complex ways. Therefore, when numerous facets are inherent in a testing situation, the various sources of measurement error should be investigated carefully.

Generalizability Theory

In this section I introduce the background of univariate G theory or G theory, which deals with a single construct only, and discuss its advantages over CTT. G theory, which was first 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, 2001a; 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 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, a G study. The multifaceted nature of testing can be broken down into relevant facets of variance, enabling researchers to determine the degree to which the

variance of each facet contributes to the total variance of the test scores. The facets to be examined depend on the testing situation. 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 CRTs. In CTT, the variability of the test scores is often highly related to the reliability of the test. Because the purpose of NRTs is to spread examinees' test scores out along a continuum, such variance is appropriate for determining the reliability of NRTs. In contrast, with CRTs, the variance can 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 usually relatively small because a limited number of students take classroom-level tests. In a language program in which placement tests are administered to create homogeneous classes in terms of proficiency level, that homogeneity is likely to suppress test score variance. 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 creates a negatively skewed distribution that suppresses score variance.

Because the purpose of criterion-referenced testing is not to spread students' test scores but to estimate students' achievement in a specific domain, CTT reliability is not appropriate for estimating criterion-referenced measurement consistency. G theory, however, can be applied to estimate dependability for CRTs because it

provides not only coefficients for relative decisions, but also coefficients for absolute decisions. 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, a D study. In CTT, the Spearman-Brown prophecy formula can be employed to estimate reliability with different numbers of items:

$$r_{kk} = \frac{k r_{xx}}{1 + (k - 1) r_{xx}}$$

where: r_{kk} = estimated reliability when the multiple of test items is set at k

k = number of items

r_{xx} = reliability

However, this formula cannot deal with multifaceted sources of error. 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 design should be carefully planned 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 five categories for assessing writing (e.g., 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 (e.g., 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):

$$E\rho^2(\delta) = \frac{\sigma^2(p)}{\sigma^2(p) + \frac{\sigma^2(pi)}{ni}}$$

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

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):

$$E\rho^2(\Delta) = \frac{\sigma^2(p)}{\sigma^2(p)+\left(\frac{\sigma^2(i)}{ni}\right)+\left(\frac{\sigma^2(pi)}{ni}\right)}$$

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

The difference between the equations for relative and absolute decisions lies in how error variance is defined. For relative decisions, as 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; 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, the teacher's aim is

to estimate the 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, as it provides an estimate of 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 latter 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 applications of G theory have been published, particularly for performance testing (Brennan, 2000; Brennan, Gao, & Colton, 1995; Cronbach, Linn, Brennan, & Haertel, 1997). With regard to language testing, only a few writers have referred to G theory (Bachman, 1990, 1997, 2004; Brown & Hudson, 2002). In the field of language testing, Bolus, Hinofotis, and Bailey (1982) were the first to suggest the usefulness of G theory. Brown (1982) applied G theory to examine the effects of numbers of items and passages on the dependability of an engineering English reading comprehension test. Since then, G theory has been used for analyzing performance tests (e.g., Brown, & Bailey, 1984; Schoonen, 2005; Stansfield, & Kenyon, 1992; Yamanishi, 2005) and paper-and-pencil

NRTs such as the TOEFL (Brown, 1999; Brown & Ross, 1996), and the TOEIC (Zhang, 2006).

Multivariate Generalizability Theory

Multivariate G theory is an extension of G theory and is viewed as a special case of G theory because multivariate G theory incorporates all the analyses that can be conducted with G theory and also provides additional results such as covariance components and multivariate G coefficients (Brennan, 2000). Cronbach, Gleser, Nanda, and Rajaratnam (1972) initially introduced multivariate G theory and discussed the computational procedures and interpretation of variance and covariance components. Multivariate G theory was reviewed in Shavelson, Webb, and Rowley (1989) and fully discussed in Brennan (2001a) and Webb, Shavelson, and Maddahian (1983).

Multivariate G theory is a powerful test theory because it allows researchers to: (a) estimate multivariate G coefficients for multiple measures, (b) estimate variance and covariance components accounted for by facets effects and error in a G study, and (c) simultaneously estimate both univariate and multivariate G coefficients in a D study.

CTT reliability coefficients are inappropriate when the reliability of multiple measures designed to measure multiple dimensions is estimated simultaneously, because it is likely that it lacks internal consistency. Most other reliability indices are only applicable for analyzing one dimension. Thus, language testers have to rigorously investigate the unidimensionality of measures and then conduct separate analyses for each dimension. However, the most powerful feature of multivariate G

theory is that the reliability for multiple measures can be simultaneously estimated as a multivariate G coefficient.

Reliability in CTT can be employed to decompose an observed test score into true score and error. However, in testing situations, a number of facets create test score variance. A G study can identify the amount of variance in each component and the error contributing to the total variance. The advantage of multivariate G theory is that the theory provides not only facets variance components but also attenuated correlation coefficients for multiple measures, and the coefficients for the universe score form a convergent/divergent validity argument (Campbell & Fiske, 1959). For instance, if a language tester is interested in estimating the dependability of listening, reading, and vocabulary tests, the attenuated correlation coefficient for the reading and vocabulary tests should be higher than the coefficient for listening and reading tests. Listening and reading skills are considered separate skills. That is, while test-takers process orthography to understand reading passages, they process phonology to understand listening passages. If the coefficients among related skills are reasonably high, it can be a convergent validity argument. If the coefficients among unrelated skills are reasonably low, it can be a divergent validity argument.

The Spearman-Brown prophecy formula can demonstrate change in reliability only when the number of items has changed. A univariate D study goes a step further by showing changes in reliability when the number of facets such as items, raters, and passages has changed. A multivariate D study goes even further by indicating changes in reliability when the number of facets for multiple measures has changed. For instance, multivariate G coefficients can simultaneously indicate the number of items that are necessary to design speaking and writing tests with a certain magnitude of

reliability. This result is informative when deciding which facets in a measure have to be changed in order to achieve a certain magnitude of reliability.

The procedure for estimating the variance and covariance components for multiple measures in a multivariate G study is first done by setting the design (Brennan, 2001a). Take the example of $p\cdot \times i^{\circ}$ for reading and listening tests. The filled (·) circle indicates that the persons are the same across the measures and the unfilled (°) circle denotes that the items are different across the measures. Once the variance and covariance components are computed, the variance-covariance matrices for persons ($\sum p$), items ($\sum i$), and interactions ($\sum pi$) can be displayed:

$$\sum p = \begin{pmatrix} \sigma^2 p(\mathrm{R}) & \rho RL(p) \\ \sigma p(\mathrm{RL}) & \sigma^2 p(\mathrm{L}) \end{pmatrix}$$

$$\sum i = \begin{pmatrix} \sigma^2 i(\mathrm{R}) & \\ & \sigma^2 i(\mathrm{L}) \end{pmatrix}$$

$$\sum pi = \begin{pmatrix} \sigma^2 pi(\mathrm{R}) & \\ & {}^2 pi(\mathrm{L}) \end{pmatrix}$$

where: $\sigma^2 p(\mathrm{R})$ = persons variance for the reading test
$\sigma^2 p(\mathrm{L})$ = persons variance for the listening test
$\sigma p(\mathrm{RL})$ = persons covariance for the reading and listening tests
$\rho RL(p)$ = disattenuated correlation for the reading and listening tests
$\sigma^2 i(\mathrm{R})$ = items variance for the reading test
$\sigma^2 i(\mathrm{L})$ = items variance for the listening test
$\sigma^2 pi(\mathrm{R})$ = persons-by-items interaction variance for the reading test
$\sigma^2 pi(\mathrm{L})$ = persons-by-items interaction variance for the listening test

The diagonal elements in the matrices, such as $\sigma^2 p(\mathrm{R})$ and $\sigma^2 p(\mathrm{L})$, are the variance components. The lower off-diagonal elements, such as $\sigma p(\mathrm{RL})$, are the covariance components. Although covariance between the person's variance

components for reading and listening (σp (RL)) shows the relatedness between two sets of variance components, disattenuated correlations can be estimated and shown in the upper off-diagonal elements (e.g., $\rho RL(p)$) for easy interpretation (Brennan, 2001a):

$$\rho RL(p) = \frac{\sigma p(\mathrm{RL})}{\sqrt{\sigma^2 p(\mathrm{R})\sigma^2 p(\mathrm{L})}}$$

where: σp(RL) = persons covariance for the reading and listening tests
$\sigma^2 p$(R) = persons variance for the reading test
$\sigma^2 p$(L) = persons variance for the listening test

A multivariate D study can be conducted by changing the number of items (*ni*) to estimate multivariate G coefficients based on the variance and covariance components estimated in the multivariate G study. To achieve this, $\sum p$ is not changed but is designated as $\sum \tau$ to show the variance and covariance for the universe scores. Then, the variance components in $\sum I$ and $\sum pI$ are divided by the number of items (*ni*). $\sum pI$ is the error variance component for relative error ($\sum \delta$). The variance component for absolute error ($\sum \Delta$) consists of the items variance component ($\sum I$) plus the error variance component ($\sum pI$). The variance and covariance matrices for the universe scores ($\sum \tau$), items ($\sum I$), and persons-by-items interaction ($\sum pI$) can be displayed as:

$$\sum \tau = \begin{pmatrix} \sigma^2 p(\mathrm{R}) & \\ \sigma p(\mathrm{RL}) & \sigma^2 p(\mathrm{L}) \end{pmatrix}$$

$$\sum I = \begin{pmatrix} \sigma^2 I(\mathrm{R})/ni & \\ & \sigma^2 I(\mathrm{L})/ni \end{pmatrix}$$

$$\sum pI = \begin{pmatrix} \sigma^2 pI(\mathrm{R})/ni & \\ & \sigma^2 pI(\mathrm{L})/ni \end{pmatrix}$$

where: $\sigma^2 p$(R) = persons variance for the reading test

$\sigma^2 p$(L) = persons variance for the listening test

σp (RL) = persons covariance for the reading and listening tests

$\sigma^2 I$(R) = items variance for the reading test

$\sigma^2 I$(L) = items variance for the listening test

$\sigma^2 pI$(R) = persons-by-items interaction variance for the reading test

$\sigma^2 pI$(L) = persons-by-items interaction variance for the listening test

ni = number of items

Using the variance and covariance components for universe score variance ($\sum \tau$), relative error variance ($\sum \delta$), absolute error variance ($\sum \Delta$), composite universe score variance ($\sigma^2 c(\tau)$), composite relative error variance ($\sigma^2 c(\delta)$), and composite absolute error variance ($\sigma^2 c(\Delta)$) are estimated (Brennan, 2001a):

$$\sigma^2 c(\tau) = \omega_v^2 \sigma^2 \tau(\mathrm{R}) + \omega_{v'}^2 \sigma^2 \tau(\mathrm{L}) + 2\omega_v \omega_{v'} \sigma \tau(\mathrm{RL})$$

where: ω_v = a priori weight for the reading test

$\omega_{v'}$ = a priori weight for the listening test

$\sigma^2 \tau$(R) = universe score variance for the reading test

$\sigma^2 \tau$(L) = universe score variance for the listening test

$\sigma \tau$(RL) = universe score covariance for the reading and listening tests

$$\sigma^2 c(\delta) = \omega_v^2 \sigma^2 \delta(\mathrm{R}) + \omega_{v'}^2 \sigma^2 \delta(\mathrm{L}) + 2\omega_v \omega_{v'} \sigma \delta(\mathrm{RL})$$

where: ω_v = a priori weight for the reading test

$\omega_{v'}$ = a priori weight for the listening test

$\sigma^2 \delta$(R) = relative error variance for the reading test

$\sigma^2 \delta$(L) = relative error variance for the listening test

$\sigma \delta$(RL) = relative error covariance for the reading and listening tests

$$\sigma^2 c(\Delta) = \omega_v^2 \sigma^2 \Delta(\mathrm{R}) + \omega_{v'}^2 \sigma^2 \Delta(\mathrm{L}) + 2\omega_v \omega_{v'} \sigma \Delta(\mathrm{RL})$$

where: ω_v = a priori weight for the reading test

$\omega_{v'}$ = a priori weight for the listening test

$\sigma^2 \Delta$(R) = absolute error variance for the reading test

$\sigma^2 \Delta$(L) = absolute error variance for the listening test

$\sigma \Delta$(RL) = absolute error covariance for the reading and listening tests

The composite universe score variance ($\sigma^2 c(\tau)$) is the weighted sum of all of the elements in the universe score matrix ($\sum \tau$) where the weights ($\omega \upsilon$) are defined a priori by the investigator or nominal weights, which show the investigator's judgment regarding the relative importance of the various levels of the fixed facet (Brennan, 2001a). Once composite universe score variance, composite relative error variance, and composite absolute error variance are estimated, the multivariate G coefficients can be estimated for relative decisions ($\hat{E}\rho^2(\delta)$) and absolute decisions ($\hat{E}\rho^2(\Delta)$):

$$\hat{E}\rho^2(\delta) = \frac{\sigma^2 c(\tau)}{\left(\sigma^2 c(\tau) + \sigma^2 c(\delta)\right)}$$

where: $\sigma^2 c(\tau)$ = composite universe score variance
$\sigma^2 c(\delta)$ = composite relative error variance

$$\hat{E}\rho^2(\Delta) = \frac{\sigma^2 c(\tau)}{\left(\sigma^2 c(\tau) + \sigma^2 c(\Delta)\right)}$$

where: $\sigma^2 c(\tau)$ = composite universe score variance
$\sigma^2 c(\Delta)$ = composite absolute error variance

The development of a computer software program called mGENOVA (Brennan, 2001c) led to more frequent use of multivariate G theory. This software can handle univariate/multivariate, balanced/unbalanced, and crossed/nested G and D study designs.

In the field of educational measurement, Brennan, Gao, and Colton (1995) utilized both univariate and multivariate generalizability theory to analyze listening and writing tests. Li and Brennan (2007) conducted a series of generalizability analyses on a large-scale standardized reading comprehension test given to 500 native English speakers and 500 non-native English speakers. The G study designs adopted were *p X i*, *p˙ X i°*, *p X (i:h)*, and *p X (i:h)* where *p*, *i*, and *h* are persons, items, and

passages. The amounts of the variance components estimated were slightly different between the two groups, and the passages facet had a greater effect on the non-native English speakers. Li and Brennan concluded that a single facet design such as *p X i* overestimated the coefficients, so it should not be used when there are multiple facets.

In the field of language testing, multivariate G theory has been applied to performance testing in order to estimate dependability for speaking tests (Lee, 2006; Sawaki, 2007; Xi, 2007) and writing tests (Lee & Kantor; 2007). In the Japanese context, multivariate G theory was applied to estimate the dependability of a writing test (Okubo, 2006), to rating students' interest, motivation, and attitude toward learning English (Yamamori, 2003), and to reading/summary and reading/free recall tasks (Sawaki, 2005).

The Rasch Models

In this section, I introduce the Rasch models and the merits of applying many-faceted Rasch measurement (MFRM). Lord (1953) first proposed item response theory (IRT), and seven years later, Rasch (1960) first proposed the Rasch measurement theory (RMT). Although the two theorists had different ideas about measurement, both theories can be defined as test theories that estimate the probability of a person's chance of getting an item correct or incorrect. In the dichotomous Rasch model, the probability is based on the response of a student with a particular ability level to an item with a particular level of difficulty. Thus, the Rasch model is a probabilistic model in which person ability is estimated without reference to the item characteristics (e.g., item difficulty) of any particular set of items (item-independent) and item difficulty is estimated without reference to the person characteristics (e.g., person ability) of any particular group of test takers

(sample-independent) (Bond & Fox, 2001).

Fundamentally, two assumptions have to be met for the application of the Rasch model: (a) unidimensionality and (b) local independence (Bond & Fox, 2001). The assumption of unidimensionality means that only one ability is measured by a set of items in a test. This issue has been contentious because it is unlikely that a set of test items measures only one latent trait or construct (Henning, 1988, 1992). For instance, in foreign language testing, considering that reading is closely interrelated with lexical knowledge, a reading test also taps students' vocabulary knowledge. The assumption of local independence means that items have to function independently of one another. Simply put, a response to one item on a test should not influence the probability of correctly answering other items.

The special features of the Rasch models are that they provide more precise measurement and additional information about test takers and items than CTT. First, the models convert raw scores into interval measures. Total test scores based on CTT result in an ordinal scale that provides information about the order of the test takers' ability along a continuum. However, the application of RMT makes it possible to transform the ordinal scale into an interval scale, which gives information about the interval distances on the measurement continuum. In other words, while ordinal scales can show that 2 is greater than 1, interval scales can show to what degree 2 is greater than 1. In most cases, Rasch item difficulty and person ability estimates are expressed on a logit scale that usually ranges from -3, indicating an easy item or an

unable person to +3, indicating a difficult item or an able person. By convention, the average item difficulty estimate is set at 0.

Second, because item difficulty and person ability are placed on the same interval scale, it is possible to estimate the probability that a particular person will respond to a particular item successfully. For instance, a person with a person ability estimate of 0 has a 50 percent chance of correctly answering an item with item difficulty estimate of 0. Likewise, a person with person ability estimate of 0 has a 75% chance of correctly answering an item with an item difficulty estimate of -1.

Third, *SEM* in CTT is estimated for a whole test and is given one value, but *SEM* in RMT can be computed for each item and test taker. This information concerning the local error for all the items and test takers is more detailed and shows which items or persons are the source of error (Embretson & Reise, 2000). If the *SEM* for an item or a person is larger than, say .50, special attention must be paid when interpreting the results.

The Rasch models, popularized by Benjamin Wright, are widely used in the field of educational measurement, psychology, and medical science. The dichotomous Rasch model takes account of two elements: person ability (*B*), or theta (θ), and item difficulty (*D*), *b* parameter or delta (δ). The probability of a person answering an item correctly is determined by the distance between person ability and item difficulty. The dichotomous Rasch model is expressed by the following equation:

$$Pni = \frac{exp(Bn - Di)}{(1 + exp(Bn - Di))}$$

where: *Pni* = the probability of getting item *i* correct for person *n*
Bn = the person ability for person *n* in logit score
Di = the item difficulty for item *i* in logit score
exp = the constant or natural log function of 2.7183

In order to estimate person ability and item difficulty, a parameter estimation procedure such as unconditional maximum likelihood (UCON) and approximation (PROX) is used. Because of the mathematical complexity of the procedure, it is not described here, but the step-by-step procedure of PROX is explained in Wright and Stone (1978) and Otomo (1996). Once person ability and item difficulty are estimated after the parameter estimation, probability can be estimated. Several extensions of the Rasch model have been proposed. Andrich (1978) developed the Rasch rating scale model, and Wright and Masters (1982) developed the partial credit model. These models are extensions of the dichotomous model that incorporate item threshold levels to the equation for the Rasch dichotomous model.

Another extension to the Rasch model is the many-faceted Rasch model (Linacre, 1989), which takes account of person ability, item difficulty, and additional facets inherent in testing "situations in which more than two facets interact to produce an observation" (Linacre, 1989, p. 1). Although both theories can incorporate several facets in the analyses, MFRM is different from G theory in that "MFRM considers each examinee as an individual, and attempts to liberate, as far as statistically possible, each examinee's measure from the distributional details of the other examinees, items, and raters that happen to be in the same analysis" (Linacre, 1996, p. 86). That is, although random sampling theory, such as CTT and G theory, shows how reliable observed scores are and how generalizable observed scores are in a defined universe,

MFRM shows how accurate an examinee's person ability is estimated without reference to other facet effects. Thus, the most powerful feature of MFRM is that a FACETS analysis can incorporate possible measurement facets and can estimate an individual examinee's person ability as precisely as statistically possible.

A powerful feature of MFRM is the flexibility of the design. The model can encompass all crossed or nested designs, and the many-faceted Rasch model is robust against missing data (Bond & Fox, 2001, p. 111) in that missing data do not drastically affect the precision of the estimates. Finally, by adding a certain facet, an ANOVA-like analysis can be conducted using the many-faceted Rasch model. By adding facets such as occasions and proficiency level, a *p*-value for the chi square and the Rasch separation index can signify whether one group is more or less able than other groups on the interval scale. For instance, if a test is given twice in a pretest/posttest design to the same group of students, a repeated-measures ANOVA shows whether the difference in the two scores is statistically significant. The many-faceted Rasch model can be used for the same purpose, but it provides additional information about the invariant-examinees' ability without reference to other facets.

The many-faceted Rasch model is often used in performance testing to take multiple rating facets, such as different tasks, different prompts, and rater severity into account. The mathematical equation for the model is expressed as:

$$Pnikj = \frac{exp(Bn - Di - Fk - Cj)}{1 + exp(Bn - Di - Fk - Cj)}$$

where: *Pnikj* = the probability of person *n* when rated on item *i* by rater *j* for category threshold *k*

Bn = the person ability for person *n* in logit score

Di = the item difficulty for item *i* in logit score

Fk = the category threshold difficulty for threshold *k* in logit score

Cj = the rater severity for rater *j* in logit score

exp = the constant or natural log function of 2.7183

As with the dichotomous Rasch model, all the estimates for each facet are calibrated by a parameter estimation procedure such as UCON; this allows the probability of person *n* rated by rater *j* on item *i* for category *k* to be estimated.

In the field of psychometrics and educational measurement, a number of books have been published on IRT (e.g., Embretson & Hershberger, 1999; Embretson, & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991) and on the Rasch models (Bond & Fox, 2001; Wilson, 2005). In language testing, several books referring to IRT (Bachman, 1990, 2004; Brown, & Hudson, 2002; Henning, 1987; Lynch, 2003) and the Rasch models (McNamara, 1996) have been published. Henning (1984) initially introduced IRT in language testing and listed a number of the advantages of IRT. Since then, a number of researchers have analyzed paper-and-pencil tests using the dichotomous Rasch, and the two and three parameter logistic models (e.g., Beglar, 2010; Beglar, & Hunt, 1999; Choi & Bachman, 1992; McNamara, 1990). In recent years researchers have used the many-faceted Rasch model more frequently due to the increasing number of studies on performance testing. A number of researchers have investigated rater performance (e.g., Kondo-Brown, 2002; Lumley & McNamara, 1995; Schaefer, 2008), while others have used MFRM for test development and test validation (Kenyon, 1998; Norris, Brown, Hudson, & Bonk, 2002; Tyndall & Kenyon,

1995; Upshur & Turner, 1999). G theory and the many-faceted Rasch model have been used in the same study (e.g., Bachman, Lynch, & Mason, 1995; Lynch & McNamara, 1998; Kozaki, 2004), and a number of researchers have used the many-faceted Rasch model for ANOVA-like analyses. For instance, when gender was incorporated in an analysis of a speaking test, no significant impact was found on test performance (O'Loughlin, 2002; Lumley & O'Sullivan, 2005). In addition, Monsaas and Engelhard (1996) administered an instrument with seven subscales two times in a pretest/posttest design and conducted a FACETS analysis to evaluate the effects of an intervention program by creating a time effect as a facet in the many-faceted Rasch model. The results showed that changes were observed.

Gaps in the Literature

The following gaps remain in the field of criterion-referenced testing research:

1. Few criterion-referenced studies have been conducted in Japan, even in the many years following the implementation of the absolute evaluation system, as described in the *Course of Study* (1989). No researchers have discussed writing, statistically analyzing, and revising criterion-referenced test items in Japan.
2. With the exception of Griffee (1995) and Kumazawa (2007b, 2009), no researchers have reported the results of classroom criterion-referenced tests created by teachers in the Japanese university context.
3. A number of researchers have investigated the dependability and validity of CRTs (e.g., Hudson, 1991; Kunnan, 1992), but pretest/posttest score gains in a counterbalanced design have not been thoroughly investigated. The only exceptions are studies by Brown (1993, 2001). Without the investigation of score gains, it is difficult to estimate students' mastery of course objectives.

4. There is little information about the relationships and interactions among the variance components of CRTs, as well as the extent to which the variance components contribute to the total variance. No researcher has applied multivariate G theory to investigate the extent to which facets in paper-and-pencil CRT forms contribute to the total variance and the relationships among the variance components.

Research Questions

The main purpose of this study is to report on CRT development as part of systematic curriculum development. The performance of two CRT forms is discussed in terms of their dependability and validity. To this end, the following research questions were posed:

1. To what extent do the criterion-referenced items function effectively for the students in this study?
2. To what extent do the facets of persons, classes, items, sections, and subtests contribute to the total score variation in two CRT forms?
3. To what extent are two CRT forms dependable when administered as pretests and posttests?
4. To what extent are two CRT forms valid when administered as pretests and posttests?

CHAPTER 3

GENERAL ENGLISH LANGUAGE PROGRAM

The purpose of this chapter is to describe the research context. Criterion-referenced tests are curriculum-specific; thus, it is important that the context be described in some detail.

Unified Curriculum

This study was conducted at a private liberal arts university in eastern Japan. This is a high-ranking university and is renowned for having a well-coordinated general English language program that is one part of the required general education program. In 2007 the university was made up of nine departments: literature, economics, business, science, social science, law, tourism, welfare, and modern psychology. The approximately 4,300 first-year students enrolled in the university had to earn eight credits in this program. Students have to earn 126 credits to graduate from the university.

The salient features of the program are as follows: (a) a unified English language curriculum, (b) two major English courses, (c) a placement test, and (d) a Reading and Listening Personal Computers unified examination. The curriculum is unified in the sense that classes taken in the first and second semesters are sequenced in order of difficulty, and a unified syllabus developed by the administrators is used to unify the class content. All the classes in the two courses are based on a syllabus that describes (a) the class goals, (b) materials, (c) teaching methods, and (d) evaluation methods. The section on class goals provides a general description of the class and the goals of the course. The section on materials specifies the assigned textbooks and the

content of the textbooks. Usually, teachers choose from two or three alternative textbooks. The section on teaching methods identifies activities that can be used in class and suggests ways to efficiently manage the class. The section on evaluation describes assessment methods, grading criteria for the class, and the percentage accounted for by each criterion. The unified curriculum document unifies what teachers teach, but teachers still have to design their class objectives, classroom tests, lesson plans, and teach to the objectives.

The general English program is made up of two major courses, the Language and Culture Course and the Communicative Course (Table 1). The common goal for both courses is to improve students' general English proficiency. The Language and Culture Course is focused on raising students' intercultural awareness, while the Communicative Course is focused on developing students' communication skills. Students can choose the

Table 1

General English Program Course Goals

Communicative Course (COC) Goals	Language and Culture Course (LCC) Goals
Interactive World English (IWE): Students read about social issues, exchange and present their ideas.	English for Cultural Understanding (ECU): Students read about cultural issues, exchange and present their ideas.
Media English (MED): Students read newspapers and magazines to obtain information from the media.	English through Video (ETV): Students listen to cultural topics.
Reading and Listening Regular (R&L): Improve students' reading and listening skills.	Reading and Listening Regular (R&L): Improve students' reading and listening skills.
Reading and Listening PC (R&L PC): Improve students' reading and listening from autonomous learning on a PC.	Reading and Listening PC (R&L PC): Improve students' reading and listening from autonomous learning on a PC.
Writing (WRT): Improve writing, e.g., being able to write accurate sentences and coherent, structured paragraphs. Students also learn techniques, rules, and manners in writing.	Writing (WRT): Improve writing, e.g., being able to write accurate sentences and coherent, structured paragraphs. Students also learn techniques, rules, and manners in writing.

course that they desire to take. Each course has a core class in which a teacher teaches the same group of students twice a week. The core Language and Culture Course class is English for Cultural Understanding, and the core Communicative Course class is Interactive World English. The following classes are designed to supplement skills that cannot be covered in the two core classes: Media English, Reading and Listening Regular, Reading and Listening Personal Computers, and Writing in the Communicative Course; and English through Videos, Reading and Listening Regular, Reading and Listening Personal Computers, and Writing in the Language and Cultural Course.

Under the unified curriculum, a placement test is administered at the beginning of each academic year. The in-house paper-and-pencil multiple-choice placement test consists of listening and reading sections that are designed to spread out examinees' test scores. This one-hour exam is administered in order to identify the top five percent of the students, who can choose to be exempted from the general English program, and place the remaining students into two groups based on their test scores. Those who wish to be exempted can automatically earn the eight credits and proceed to take other elective English classes offered for second-, third-, and fourth-year students. Based on the placement test scores, the students are placed into two groups, Group A and Group B (Figure 2). There are three levels for each course: level 1 (most difficult), level 2, and level 3 (least difficult). In the first semester, the students in Group A take level 2 classes,

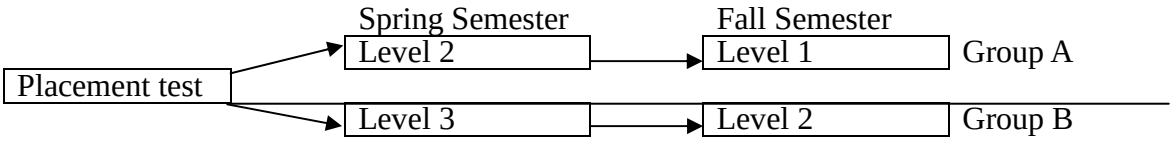

Figure 2. Systematic level sequence in the English language program.

and in the second semester, they proceed to the level 1 classes. The students in Group B take level 3 classes in the first semester and level 2 classes in the second semester. This system is designed to offer classes that match the students' proficiency levels.

Students taking the Communicative Course and Language and Culture Course classes are required to take the Reading and Listening Personal Computers unified examination at the end of the semester. In the Reading and Listening Personal Computers class, students autonomously study web-based materials on computers for the purpose of improving listening and reading skills. All the students taking the Reading and Listening Personal Computers class have to take the Reading and Listening Personal Computers unified examination, which is a NRT used to measure students' general English proficiency; thus, the test is not sensitive to the objectives of the Reading and Listening Personal Computers class, and is not linked to the curriculum content. The test has three eight-item reading sections, and two eight-item listening sections for a total of 40 questions. Each item is worth one point.

Reading and Listening Regular Class

Students taking the Communicative Course and Language and Culture Course classes are required to take the Reading and Listening Regular class. This text-based class meets once a week and is comprised of approximately 35 students and one teacher who is assigned to teach the class throughout the semester. The goals are to improve the students' listening skills, reading speed, and reading skills such as skimming, scanning, and identifying topic sentences. To achieve the class goals, students watch videos several times and study an assigned reading skills textbook.

The subgoal of this class is to increase the students' vocabulary size. The teaching method is generally formal and teacher-centered.

A textbook was developed for the Reading and Listening class to help students achieve the class goals. The book has five units, and each unit has two chapters, one for reading and one for listening. Each reading chapter consists of (a) warm-up questions designed to help students generate ideas about the topic, (b) speed reading 1, (c) four multiple-choice questions about the passage, (d) a brief introduction to a targeted skill and exercises, (e) speed reading 2, (f) speed reading 3, and (g) an analysis task that requires students to thoughtfully consider the topic. The listening chapters are made up of the following sections: (a) listening 1, (b) three true/false questions, (c) listening 2, (d) three questions asking students to provide reasons for their answers to the true/false questions, and (e) communication round-up, which involves students in a discussion.

I examined the assigned textbook from the following perspectives: (a) topics, (b) number of words, (c) readability, and (d) target skills covered in the chapters. The textbook covered a wide range of topics. It starts with general topics and shifts to more scientific and specific topics. As shown in Table 2, the number of words in the passages ranged from 254 to 547. Readability indices showed that the reading passages were generally more difficult than the listening passages. The listening passage in Chapter 4 was the easiest while the reading passage in Chapter 2 was the most difficult. Thus, the passages were not sequenced in order of difficulty.

With regard to the grading system, students' attendance accounts for 30% of their final grade, and participation accounts for 30%. Teachers have different ways of calculating student participation, but according to the university attendance policy, students have to attend more than 80% of the lessons to get credit for the class. Some

Table 2

Textbook Analysis for the Reading and Listening Regular Class

Unit: Theme	Chapter Title	# of words	Readability (FRE/FGL)	Target skill
1: University Life	1 Reading (Stepping into Adulthood)	408	67.28/5.7	Guessing
	1 Listening (Stepping into Adulthood)	254	84.67/3.6	
	2 Reading (Myths about American Universities)	418	62.34/6.3	Reading phrase by phrase
	2 Listening (Myths about American Universities)	545	78.85/4.6	
2: Relationships	3 Reading (Making Friends)	347	70.12/5.1	Longer phrase reading
	3 Listening (Making Friends)	315	90.73/2.1	
	4 Reading (Conflict!)	421	65.80/5.8	Focusing on stressed words
	4 Listening (Conflict!)	349	92.60/1.4	
3: Mental Health	5 Reading (Stress)	415	73.15/4.6	Inference 1
	5 Listening (Stress)	392	83.75/3.0	
	6 Reading (Depression)	432	62.80/6.2	Inference 2
	6 Listening (Depression)	474	75.71/4.2	
4: Travel	7 Reading (The Soul of Seoul)	393	81.66/3.6	Skimming
	7 Listening (The Soul of Seoul)	437	80.09/4.4	
	8 Reading (A Different Bali)	452	71.51/4.9	Scanning
	8 Listening (A Different Bali)	311	75.40/4.4	
5: Family	9 Reading (Women, Marriage, and Equality)	350	74.63/4.7	Identifying a topic sentence
	9 Listening (The American Family?)	547	83.95/4.9	
	10 Reading (Women, Marriage, and Equality)	386	64.57/6.0	Paragraph structure
	10 Listening (Women, Marriage, and Equality)	260	77.28/4.3	

Note. FRE = Flesch reading ease; FGL = Flesch grade level.

teachers give quizzes, collect student portfolios, or evaluate the students' attitudes toward learning in class. The final 40% is based on students' scores on the final examination, which is designed by the teachers. At least 30% of the test items are based on the content of the textbook. No guidelines are provided regarding how the test should be developed or administered in the unified syllabus. The final

examination, attendance, and participation points are added to determine the students' grades. Students with less than 60 total points receive a failing grade, D. Students with 60-69, 70-79, 80-89, and 90-100 points receive a C, B, A, and S, respectively. Students who receive a failing grade, D, cannot retake the class; they have to pay 3000 yen to take a credit approval test. If their score exceeds a certain cut-point, they can earn credit for the course.

CHAPTER 4

PILOT STUDIES

The purpose of conducting two pilot studies was to demonstrate how two CRT forms used in the main study were developed. In this chapter the methodology and results of the pilot studies are described.

Pilot Study 1

The purpose of Pilot Study 1 was to conduct an intervention construct validity study using a pretest/posttest design with two counterbalanced test forms. Two CRT forms were developed and administered to two groups of students both as pretests and posttests in a proficiency-based curriculum. To this end, the following research questions were formulated:

1. To what extent are the two CRT forms dependable in the two administrations?
2. To what extent do the students master the items on the two test forms?

Methods

Participants.

This study involved 88 first-year Japanese university students at the university where the main study was conducted. The students were divided into two groups: 45 first-year welfare majors were designated as Group A and 33 first-year literature majors were designated as Group B. Because the test involved a listening component and microphones were not used, it was important that students in the same room take the same test form. Based on the placement test results, the students in Group A were

placed in one of the lowest class levels and students in Group B were placed in one of the mid-ranking classes.

CRT development.

The three primary purposes in developing the two Reading and Listening CRT forms were to: (a) diagnose the students' non-mastery before instruction, (b) teach the items on the final examination, and (c) assess the students' achievement after the instruction. Developing the two forms before delivering the instruction allowed them to be used as diagnostic tests. Once the students' areas of non-mastery and mastery were identified, the teachers could focus on the course objectives that the students had not yet mastered and teach them the item content that would be tested on the final test. By administering the tests twice, it was possible to compare the results obtained on the pretest and posttest and examine the effectiveness of the instruction.

Two teachers set the objectives and designed the class syllabus by referring to the class goals that had been set by the administrators (See the class syllabus in Appendix E). Before the instruction and prior to developing the test specifications, I meticulously examined each chapter of the textbook (see Table 2). Once the textbook was analyzed, a test specification was developed. The two CRT forms were used to assess how well the students had learned the listening materials, the vocabulary items, and the reading materials that appeared in the assigned textbook. Table 3 shows the number of chapters, items per chapter, and items tested on each form of the CRT. In the listening test, five listening chapters were selected from the assigned textbook. Three multiple-choice comprehension items were developed to test how well the students learned the content of the dialog, and two multiple-choice items were written to test how well the students knew the receptive meaning of the vocabulary items that

appeared in the dialog. The listening section was made up of 25 items. The students had to watch a video to be able to answer the questions. A sample listening comprehension item is as follows:

How was the speaker's experience in high school?

a. He took many tests.
b. He had to go to prison.
c. He had to go to many places.
d. He was tired of high school.

The correct choice is d.

A sample vocabulary item on the listening test is as follows:

Which word has the closest meaning to "trapped" in the passage?

a. Tricked
b. Caught
c. Deceived
d. Fascinated

The word "trapped" appeared in the dialog and the students watched the video to choose the option that had the closest meaning to the word. The correct choice was b.

Twenty vocabulary items were taken from the four reading passages selected from the assigned textbook. Five items were nested within each chapter to test whether the students learned the receptive meaning of the vocabulary items in the textbook. A sample item is as follows:

The first step is the most crucial stage.

a. Essential

b. Testing

c. Commercial

d. Searching

The underlined word was the target word. The test sentence was modified from a sentence that included the word in the textbook. Students read the sentence and selected the option that had the closest meaning to the target word. The correct choice was a.

Four reading chapters were selected from the textbook, and five items were nested within each chapter for a total of 20 items. Six items were designed to estimate students' mastery of the reading strategies learned in class such as phrase reading, inferring vocabulary meanings, reading content words, and 12 items were developed to test the content of the passages. A sample reading strategy item was as follows:

Choose the LEAST important word(s) in the sentence.

a. Each step b. takes c. them farther along in their personal d. journey.

The sentence was taken from the textbook. This item was designed to test a reading strategy called focusing on reading content words. The textbook explained that focusing on reading content words was a key to reading faster. The underlined parts were the choices. The only function word in the choices was *them*, which was the least important word. Thus, the correct choice was c. A sample reading comprehension item was as follows:

What is one of the internal products?

a. The quality of experience

b. Test scores

c. The university one goes to

d. The job one had

The students read a passage taken from the textbook and chose the best option. The correct choice was a. Based on the test specification, I developed all the items and had native and non-native teachers carefully proofread the examinations. After considering their comments, the CRT forms were finalized.

Table 3

2005 Form A and Form B of the Reading and Listening CRT

	Listening	Vocabulary	Reading
# of chapters	5 chapters	4 chapters	4 chapters
# of items per chapters	3 comprehension 2 vocabulary	5 receptive	2 strategy 3 comprehension
# of items	25	20	20

Procedures.

In Brown's (1995) model, CRTs are developed in the testing phase and administered as diagnostic tests before instruction and as achievement tests after instruction to examine the effectiveness of the instruction. There are four main administration designs: (a) posttest only, (b) pretest/posttest with one form, (c) pretest/posttest with two forms, and (d) pretest/posttest with two counterbalanced forms (adapted from Popham, 2003).

All too often only a posttest is administered as an achievement test in order to calculate students' final grades. This design has two limitations: (a) no evidence of student score gain and (b) no teach-to-test instruction. In other words, it is not

possible to know whether or not the students learned anything from the course and whether or not the information that was taught matched the test content. If a diagnostic test is not administered, teachers have no information concerning what students can do before instruction. An achievement test should show the extent to which students understand the designated content by the end of semester. By comparing students' scores on pretests and posttests, teachers can understand student learning more clearly.

The pretest/posttest design using one form solves the problems arising from posttest-only designs. By comparing students' test scores before and after instruction, teachers can at least partly determine what students have learned in class. If teachers set a cut-point in a situation where they have to identify non-mastery and mastery, the *B*-index can be calculated to determine how much each item contributes to the pass/fail decisions that are often made with CRTs (Brown, 2003, p. 15). With this design, the most suitable indicator of sound criterion-referenced items, the difference index (DI), can be easily calculated in order to determine the extent to which students learned the item content as a result of instruction. However, this design suffers from a drawback known as pretest reactivity (Popham, 2003, p. 152). If students know that a test administered as a diagnostic test is also going to be used as an achievement test, they might only study the parts of the class content that are on the pretest. In addition, given that there is a limit to what can be tested with one form, teachers cannot test a wide range of class content.

Using a design that incorporates multiple pretest/posttest forms minimizes pretest reactivity effects, and teachers can test a wider range of class objectives. However, this design also entails a pitfall. If the difficulties of the two CRT forms

differ, then it is not possible to estimate students' achievement simply by subtracting their test scores on the diagnostic test from their test scores on the achievement test.

One feasible solution to this problem is to adopt a counterbalanced pretest/posttest design in which the class is divided into two groups and half of the students take one form each as a pretest, then the forms are switched so that the students are tested with a different form on the posttest. Although this design does not solve all the problems that have been mentioned above, it does minimize them. In addition, intervention construct validity studies can be carried out with this design. If the CRTs measure the construct well, the students' scores should increase significantly between the pretest and the posttest, and the score gains can be used as the basis for one argument for the construct validity of the CRTs (Brown & Hudson, 2002, p. 225). Another advantage of this design is that the results obtained from diagnostic tests can be used to revise the CRTs so that more refined CRTs can be administered as achievement tests. This design was adopted in two published studies by Brown (1993, 2001), who reported that only a slight increase was observed in terms of the DI value in one semester.

For the above reasons, the counterbalanced pretest/posttest design was adopted in this pilot study. Forms A and B were administrated to Groups A and B respectively as a pretest. The teacher informed the students that their pretest scores would have no effect on their final grades and explained the purpose of the diagnostic test. Although the test score sheets were returned to the students, the question sheets were collected to maintain test security. The students were asked to study the class material, such as the textbook and handouts, because they would be tested on the final examination day (See sample activity sheets in Appendix F). Twenty-six 90-minute classes were taught by two teachers twice a week for 13 weeks. At the end of the semester, the test form

that the students did not take at the beginning of the class was administered. Their test scores determined 40% of their final grades. Students were given 90 minutes to complete the test.

Analysis.

The responses of the students were dichotomously scored, and then processed in EXCEL [version 2007]. Missing blanks were treated as incorrect responses. Item facility (IF) and the difference index (DI) were calculated for all items. Dependability indices or generalizability coefficients for absolute decisions were calculated using GENOVA [version 2.1] (Crick & Brennan, 1983) with the *p X (I:S)* D study design. In this design, students take a test form in which the items are nested within chapters.

Results

Tables 4 and 5 display the descriptive statistics for the 45 students in Group A and the 31 students in Group B who took Forms A and B, respectively, as pretests. Group A was less proficient as indicated by their means on the listening, vocabulary, and reading subtests of 9.27, 9.29, and 7.24, respectively. Group B obtained slightly higher means of 10.68, 12.23, and 9.42 on the listening, vocabulary, and reading subtests. The Cronbach alpha internal consistency reliability coefficients, a norm-referenced reliability estimate yielded estimates of .36 (listening), .62 (vocabulary), and .37 (reading) for Form A and .20 (listening), .32 (vocabulary), and .23 (reading) for Form B. The dependability indices were slightly lower than the Cronbach alpha coefficients, ranging from .09 to .56. At the end of the semester, Group A (n = 45) took Form B as a posttest and obtained means of 11.66 (listening), 9.75 (vocabulary), and 9.50 (reading), as shown in Table 7.

Table 4

Descriptive Statistics for Form A Pretest in Pilot Study 1

	Listening	Vocabulary	Reading
k	25	20	20
N	45	45	45
M	9.27	9.29	7.24
SD	2.79	3.29	2.44
Minimum	4.00	3.00	3.00
Maximum	18.00	16.00	13.00
Alpha	.36	.62	.37
Dependability	.28	.56	.27

Table 5

Descriptive Statistics for Form B Pretest in Pilot Study 1

	Listening	Vocabulary	Reading
k	25	20	20
N	31	31	31
M	10.68	12.23	9.42
SD	2.40	2.43	1.98
Minimum	7.00	7.00	5.00
Maximum	15.00	16.00	12.00
Alpha	.20	.32	.23
Dependability	.09	.25	.14

The standard deviation for the vocabulary section on the Form B posttest was the largest at 3.96, indicating that some students studied for this final examination but others did not. Group B (n = 33) took Form A as a posttest. The students obtained means of 14.27, 14.67, and 12.67 for the listening, vocabulary, and reading tests, respectively, as shown in Table 6. The reliability coefficients varied widely from .01 to .77, as did the dependability indices, which ranged from .00, to .76. The data obtained from Form A and Form B used as pretests and as posttests were averaged, and the means for the pretests and the posttests were 9.69 and 12.09. Thus, a slight increase in the mean scores was observed.

Table 6

Descriptive Statistics for Form A Posttest in Pilot Study 1

	Listening	Vocabulary	Reading
k	25	20	20
N	33	33	33
M	14.27	14.67	12.67
SD	2.84	2.10	1.68
Minimum	8.00	9.00	10.00
Maximum	19.00	18.00	16.00
Alpha	.45	.40	.01
Dependability	.28	.28	.00

Table 7

Descriptive Statistics for Form B Posttest in Pilot Study 1

	Listening	Vocabulary	Reading
k	25	20	20
N	44	44	44
M	11.66	9.75	9.50
SD	3.02	3.96	2.78
Minimum	6.00	3.00	2.00
Maximum	18.00	17.00	16.00
Alpha	.47	.77	.53
Dependability	.46	.76	.50

Table 8 displays the item facility values and the difference index for both forms. As a rule of thumb, IF should be close to .00 when tests are used as pretests. Conversely, IF should be close to 1.00 when tests are used as posttests. The IF values for items 26, 32, and 56 in Form A and for items 9, 11, 16, 19, 20, 30, 35, 38, 42, 43, 45, 48, 55, 59, 60, 63, and 64 in Form B used as pretests, were above .70, indicating that most of the students could already answer the items before instruction. The IF values for items 8, 10, 18, 21, 24, 29, 50, 57, and 62 in Form A and for items 1, 4, 6, 10, 18, 22, 28, 40, 47, 49, 50, 51, and 60 in Form B used as posttests, were below .30, indicating that the students did not learn the items even after the instruction. Eight items on Form A and 37 items on Form B had negative DI values. These unwanted results were probably due to the difference in the proficiency levels between Group A and Group B.

Table 8

Criterion-Referenced Item Statistics for Form A and Form B

Item	IF (Post Form A)	IF (Pre Form A)	DI (Form A)	IF (Post Form B)	IF (Pre Form B)	DI (Form B)
1	.82	.24	.57	.20	.13	.07
2	.67	.18	.49	.37	.50	-.13
3	.55	.40	.15	.30	.23	.06
4	.85	.55	.30	.20	.23	-.03
5	.55	.51	.03	.39	.50	-.11
6	.52	.62	-.11	.07	.03	.04
7	.64	.60	.04	.43	.20	.23
8	.27	.13	.14	.50	.19	.31
9	.45	.32	.14	.68	.77	-.09
10	.18	.47	-.28	.11	.48	-.37
11	.91	.33	.58	.77	.81	-.03
12	.64	.42	.21	.41	.10	.31
13	.67	.25	.42	.59	.23	.37
14	.94	.69	.25	.36	.55	-.18
15	1.00	.52	.48	.59	.42	.17
16	.85	.60	.25	.58	.74	-.16
17	.39	.25	.14	.77	.40	.37
18	.24	.55	-.30	.25	.45	-.20
19	.42	.32	.11	.84	.84	.00
20	.27	.09	.18	.68	.87	-.19
21	.21	.19	.02	.59	.52	.07
22	.91	.40	.51	.25	.16	.09
23	.52	.20	.32	.27	.48	-.21
24	.24	.27	-.03	.73	.39	.34
25	.58	.25	.33	.73	.52	.21
26	.97	.73	.24	.66	.61	.05
27	1.00	.68	.32	.61	.45	.16
28	.88	.58	.30	.20	.45	-.25
29	.12	.53	-.41	.41	.65	-.24
30	.73	.47	.26	.47	.77	-.31
31	.48	.40	.08	.34	.42	-.08
32	1.00	.73	.27	.45	.65	-.19
33	.41	.18	.22	.55	.52	.03
34	.61	.30	.30	.32	.61	-.29
35	1.00	.56	.44	.70	.90	-.20
36	.58	.18	.40	.50	.58	-.08
37	.85	.38	.47	.70	.35	.35
38	.88	.59	.29	.44	.84	-.40
39	.97	.66	.31	.32	.61	-.29
40	.82	.59	.23	.23	.39	-.16
41	.55	.34	.20	.45	.32	.13
42	.85	.34	.51	.57	.71	-.14
43	.45	.27	.19	.73	.77	-.05
44	.58	.34	.23	.55	.68	-.13
45	.97	.56	.41	.57	.94	-.37
46	.15	.22	-.07	.36	.61	-.25
47	.79	.38	.41	.20	.65	-.44
48	.88	.67	.21	.73	.94	-.21
49	.36	.18	.19	.09	.10	-.01
50	.03	.09	-.06	.25	.52	-.27
51	.52	.22	.29	.07	.13	-.06
52	.48	.33	.15	.59	.23	.37
53	.94	.36	.58	.64	.68	-.04
54	.82	.27	.55	.32	.32	.00
55	.94	.60	.34	.70	.77	-.07
56	.97	.76	.21	.43	.26	.17
57	.18	.29	-.11	.43	.06	.37

(Table 8 continues)

(Table 8 continued)

Item	IF (Post Form A)	IF (Pre Form A)	DI (Form A)	IF (Post Form B)	IF (Pre Form B)	DI (Form B)
58	.85	.38	.47	.55	.00	.55
59	.88	.49	.39	.68	.71	-.03
60	.97	.67	.30	.25	.90	-.65
61	.48	.27	.22	.61	.45	.16
62	.24	.11	.13	.39	.07	.32
63	.79	.07	.72	.98	1.00	-.02
64	.91	.51	.40	.84	.77	.07
65	.48	.44	.04	.39	.33	.05

Discussion

Let us consider the results of this pilot study in terms of the research questions. Research question asked, "To what extent are the two CRT forms dependable in the two administrations?" The dependability indices for the Form B listening and reading subtests, which were taken by Group B as a pretest and the Form A reading subtest, which was taken by Group B as a posttest, were low because little variance was observed in the test scores. The other tests were found to have moderate dependability values ranging from .25 to .76. These results indicated that, prior to conducting Pilot Study 2, the tests with low dependability had to be revised. Although statistics can be one indicator for deciding the quality of items, teachers should also examine test content carefully in order to decide whether or not items are measuring the target objectives of the class, especially when the sample size and the number of items are both small.

Research question 2 asked, "To what extent do the students master the items on the two test forms?" The pretest/posttest design with two counterbalanced forms enables teachers to determine the effectiveness of their instruction to some degree by inspecting two indicators; DI and IF. To calculate DI, the same items have to be administered as both the pretest and posttest. Recall that DI for the items in Form B had negative values. Considering that the proficiency level of Group A and Group B

differed, this was unsurprising. For that reason, it is difficult to say how many items were learned based on DI. Ideally the DI statistic should be used when the proficiency levels of two groups are approximately equal.

The other indicator of student learning is score gain. This is a simple but useful statistic for indicating the effectiveness of instruction. For instance, recall that the means for the combined pretests and posttests for the vocabulary tests were 10.76 and 12.21, respectively. This suggests that a limited degree of learning occurred in the interval between the tests. This finding constitutes one of the arguments for the construct validity of the CRT because the objectives taught in the class were learned to some extent and this is reflected in the gain score.

Pilot Study 2

The purposes of Pilot Study 2 were to revise the two test forms based on the results obtained in Pilot Study 1 and to pilot the test forms again to the students in the class in which the main study was conducted. The two test forms described in the first pilot study were revised and administered to 34 students in one class that I was in charge of teaching. The following two research questions were posed:

1. To what extent are the two revised CRT forms dependable in the two administrations?
2. To what extent do the students learn the class materials tested on the two revised CRT forms?

Methods

Participants.

Thirty-four first-year literature students taking the Communicative Course

participated in this pilot study. This group of students, who were taking the Reading and Listening Regular class, was classified as a below-average proficiency group based on their scores on the in-house placement test. It was desirable to include more students in Pilot Study 2, but because I was assigned to teach only one class and was not able to ask other teachers to administer the CRT, the tests were administered only to one class.

Instrumentation.

Due to the low dependability indices obtained in Pilot Study 1, major changes were made to the designs of the listening and reading CRT. The original listening test had five dialogs with three comprehension and two vocabulary items nested within each dialog. Recall that the dependability indices for the listening section ranged from .09 to .46. The low values were mainly due to the performance of the vocabulary items in the listening section. To improve dependability, the vocabulary items that did not measure the students' listening achievement were excluded. In addition, because each of the dialogs lasted about two to five minutes and two minutes had to be allocated between the dialogs to allow students to answer questions, it was a rather time-consuming test. For the above reasons, the vocabulary items and one dialog were eliminated; thus, the revised test was made up of four dialogs and 20 comprehension questions.

The reading test originally had four passages with two strategy questions and three comprehension questions nested within each passage. The dependability indices were low, ranging from .00 to .50 except the one in Form B taken as a posttest. Because the best strategy for improving dependability is to add more items, one more strategy question and two more comprehension questions were added to each passage.

Thus, three strategy questions and five comprehension questions were nested within each passage. In total, there were 32 reading test items.

Table 9

2006 Form A and Form B of the Reading and Listening CRT

	Listening	Vocabulary	Reading
# of chapters	4 chapters	4 chapters	4 chapters
# of items per chapters	5 comprehension	5 receptive	3 strategy 5 comprehension
# of items	20	20	32

In addition to redesigning the test specification and the test forms, minor changes were made to the items based on the output produced by ITEMAN [version 3.6] (Assessment Systems Cooperation, 1998), which provided descriptive statistics and the results of a distracter efficiency analysis. The difference index is the best indicator to judge sound criterion-referenced items, but because several items displayed negative DI values due to the difference in the students' proficiency levels in the two groups, DI was not a useful source of information for revising the test forms. The items with excessively high or low IF values were revised by changing the wording of the correct choice to make it more attractive or the wording of the distracters to make them more or less attractive. For instance, the third item on the listening subtest was revised so that the correct choice would attract more students.

How was the speaker's experience in college?

a. He wanted to study harder.

b. He had a lot of free time.

c. He did not go to classes.

d. He was always late for classes.

The correct choice was a. The distractor efficiency analysis showed that 40%, 31%, 7%, and 21% of the students selected choice a, b, c, and d, respectively when the third item in Form A was given to the more proficient group of the students as a pretest. When the item was given to the less proficient group of the students as a posttest, 55%, 42%, 3%, and 0% of the students selected choices a, b, c, and d, respectively. The IF values were .40 and .55, and the DI value was .15. In terms of the DI value, the item was not problematic, but choice b was as attractive as choice a. To make the correct choice more attractive, it was changed to "He was treated as equal as his teachers." Also, to make choice b less attractive, it was changed to "He was treated equally with other students."

The following item was revised because one of the distractors did not function well.

How was the speaker's experience in high school?

a. He took many tests.
b. He had to go to prison.
c. He had to go to many places.
d. He was tired of high school.

The correct choice was choice d. The distractor efficiency analysis revealed that 36%, 29%, 11%, and 24% of the students selected choice a, b, c, and d, respectively when the item was given as the first question on the pretest to the less proficient group, and 9%, 9%, 0%, and 82% of the students selected the choices when the item was given as a posttest to the more proficient group. The item was the first question on the test, so it was designed to be easy. The IF values were .24 and .82, and the DI value was .58, indicating that the item was working well. However, choice c was

selected by many students on the pretest and no one selected it on the posttests. Thus, choice c was rewritten as “He had few friends.”

The following item provides an example of how the items were revised when one of the distractors was too attractive.

What does Elizabeth do to relax on the way back home?

a. She runs.

b. She enjoys the scenery of her neighborhood.

c. She listens to songs.

d. She takes a walk.

The correct choice was d; 4%, 31%, 2%, and 62% of the students chose choice a, b, c, and d, respectively when the sixth item in Form A was given as the pretest to the more proficient group of students. When the item was given to the less proficient group, 0%, 45%, 3%, and 52% selected the choices. The IF values were .62 and .52 and the DI value was -.11. The primary problem with this item was that choice b was too attractive. Thus, it was rewritten as “She works out a lot” to make it less attractive.

Procedures.

Form A and Form B were given as a pretest and posttest. The teacher announced that the pretest scores would be part of the participation grade and encouraged the students to do their best on the first day of class. Most of the students took 90 minutes to complete the test. The test scores and the means for each of the subtests were reported to the students in the third class. The students who scored lower than the mean were particularly encouraged to get higher scores than the pretest

means on the final examination because the students who scored lower than the means on the pretest or who were categorized as “unsatisfactory” might fail the class. Although the teacher would not make pass/fail decisions based on their test scores, it was considered important to set a cut-point for each of the subtests so that the students would clearly understand the teacher’s expectation and would be encouraged to study harder. For teachers, whether or not students scored higher than a cut-point could be a useful source of information used to determine whether the students would pass the class. The final examination accounted for 40% of the students’ final grade, and the remaining 60% was determined by their attendance and participation grades. The details of the final examination, including the number of items, the content, and the coverage of the textbook, were announced one month before the examination was given. Twelve 90-minute classes were taught once a week. The students took the 90-minute CRT on the last day of class.

Analysis.

The responses for all the items were dichotomously scored and missing responses were treated as incorrect. ITEMAN [version 3.6] (Assessment Systems Corporation, 1998) was used to calculate the descriptive statistics. mGENOVA [version 2.1] (Brennan, 2001c) was used to estimate multivariate coefficients for absolute decisions for the pretest and posttest using the $p^{\cdot} \, X \, I^{\circ}$ design. A multivariate decision study also provided univariate generalizability coefficients for absolute decisions for the listening, vocabulary, and reading pretests and posttests (Brennan, 2001a). Criterion-referenced item analysis was carried out using EXCEL [version 2007]. Because the pretest means were set as the cut-points, the *B*-index was calculated to determine how well each item differentiated satisfactory and

unsatisfactory groups of students. The difference index could not be computed because the same items were not administered twice as a pretest and posttest. Instead of using DI as an indicator of students' learning, their total test scores were used to determine the degree of their learning.

Results

Table 10 and Table 11 display the descriptive statistics for the Form A pretest and Form B posttest, respectively. Thirty-three students took Form A as a pretest. The means were around the midpoints, and the dependability indices were slightly lower than the alpha coefficients. The vocabulary test had the highest dependability index at .63, and the listening test had the lowest value at .25. Thirty-four students took Form B as a posttest. The means on the posttest were higher than the means on the pretest. The means for the vocabulary and reading tests increased more than the mean for the listening test. The dependability of the listening posttest was much higher than that of the pretest, but the dependability indices of the vocabulary and reading tests were slightly lower than the pretest indices. The multivariate generalizability coefficients for absolute decisions for the pretest and posttest were .63 and .70.

Table 12 shows the results of the criterion-referenced item analysis. Twenty-five items had an IF value above .70 on the pretest. Based on the *B*-index values, the pretest items differentiated the satisfactory and unsatisfactory groups of students except for nine items that had negative values. Only six posttest items had IF values below .30. Most of the items had moderate or high IF values. On the listening and vocabulary tests, four posttest items had negative *B*-index values, and 14 items with negative *B*-index values were identified on the reading posttest, especially the ones measuring the students' mastery of the reading strategy, such as items 41, 42, and 43.

However, the *B*-index values for the reading test had to be carefully interpreted because only one student scored below the cut-point and the index was excessively sensitive to that student's responses.

Table 13 displays the students' subtest, total, and gain scores. Although the contents and difficulties of the two forms differed, comparing the students' scores on the pretest and posttest provides an indication of the students' learning. Though the listening test scores of nine students did not improve, an average gain of 1.61 points was observed. Only three students' vocabulary test scores decreased; there was an average gain of 3.67. One student's reading test score decreased, but the mean for the reading test improved by 5.55 points. Two students did not improve their test scores, but overall, the students' total score increased by an average of 10.97 points.

Table 10

Descriptive Statistics for Form A Pretest in Pilot Study 2

	Listening	Vocabulary	Reading
k	20	20	32
N	33	33	33
M	10.42	11.70	16.27
SD	2.36	3.37	3.66
Minimum	6.00	6.00	8.00
Maximum	15.00	17.00	25.00
Alpha	.30	.67	.52
Dependability	.25	.63	.47

Table 11

Descriptive Statistics for Form B Posttest in Pilot Study 2

	Listening	Vocabulary	Reading
k	20	20	32
N	34	34	34
M	11.97	15.38	21.97
SD	3.03	2.57	3.06
Minimum	6.00	9.00	15.00
Maximum	18.00	20.00	27.00
Alpha	.59	.50	.45
Dependability	.54	.45	.39

Table 12

Criterion-Referenced Item Analysis for Form A and Form B in Pilot Study 2

Item	IF (pretest)	IF (posttest)	*B*-index (pretest)	*B*-index (posttest)
1	.06	.29	.10	.36
2	.61	.68	.21	.01
3	.76	.71	.03	.45
4	.61	.76	.49	.52
5	.94	.65	.04	.58
6	.55	.82	.26	.19
7	.64	.50	.12	.00
8	.24	.62	-.17	.35
9	.15	.41	.10	.50
10	.52	.32	.35	.19
11	.73	.97	.26	.17
12	.79	.91	.36	.50
13	.79	.24	.36	.08
14	.36	.88	.39	.46
15	.76	.29	.45	-.25
16	.58	.65	.30	.58
17	.39	.65	.06	.38
18	.42	.56	.06	-.13
19	.12	.68	-.23	.42
20	.42	.38	.34	.26
21	.73	.97	.52	.25
22	.82	.91	.30	.18
23	.64	.59	.22	.38
24	.12	.85	.10	.40
25	.39	.97	-.10	-.03
26	.45	.88	.20	.15
27	.85	.76	.36	.58
28	.30	.76	.12	.30
29	.61	.71	.42	.23
30	.85	.50	.36	.57
31	.55	.56	.31	.35
32	.61	.94	.29	.22
33	.79	.44	.50	.50
34	.88	.53	.03	.32
35	.58	1.00	.49	.00
36	.48	1.00	.33	.00
37	.39	.76	.54	.30
38	.48	.79	.33	.50
39	.36	.50	.16	.57
40	.82	.97	.30	-.03
41	.73	.94	.17	-.06
42	.73	.97	.43	-.03
43	.18	.79	.06	-.21
44	.91	1.00	-.03	.00
45	.61	.88	.35	.91
46	.76	.94	.23	.97
47	.70	.76	.00	.79
48	.70	.94	.51	-.06
49	.24	.79	.29	-.21
50	.19	.42	-.23	-.59
51	.32	.36	.04	.38
52	.56	.64	.21	.66

(Table 12 continues)

(Table 12 continued)

Item	IF (pretest)	IF (posttest)	*B*-index (pretest)	*B*-index (posttest)
53	.70	.76	.51	-.24
54	.70	.68	.43	.70
55	.21	.68	.06	.70
56	.48	.74	.14	-.27
57	.81	.24	.23	.25
58	.28	.47	-.07	.48
59	.33	.18	-.12	.18
60	.78	.71	.31	.73
61	.41	.76	-.13	.79
62	.25	.91	.37	.94
63	.44	.59	.14	.61
64	.72	.21	.34	.21
65	.75	.88	.26	-.12
66	.38	.56	-.10	-.45
67	.66	.88	.10	-.12
68	.31	.36	.20	.38
69	.41	.56	.22	-.45
70	.47	.85	.06	-.15
71	.41	.74	.14	-.27
72	.27	.68	.03	.70

Discussion

The answers to the research questions are discussed based on the results obtained from Pilot Study 1 and Pilot Study 2.

Research question 1 asked, “To what extent are the two revised CRT forms dependable in the two administrations?” In both pilot studies, generalizability theory was applied to estimate the dependability of the two CRT forms. Few changes were made to the vocabulary test because the dependability was higher than those of the other two tests. The magnitude of the dependability was maintained in Pilot Study 2. However, it was decided to choose another chapter from the textbook and add five more items to measure the students’ mastery of the vocabulary in that chapter and thereby improve the dependability even further in the main study.

Table 13

Section Scores, Total Scores, and Gain Scores for Pilot Study 2

ID	List (pre)	List (post)	List (gain)	Voc (pre)	Voc (post)	Voc (gain)	Read (pre)	Read (post)	Read (gain)	Tot (pre)	Tot (post)	Tot (gain)
1	14	15	1	17	9	-8	14	20	6	45	59	14
2	10	13	3	12	18	6	20	23	3	42	54	12
3	10	13	3	15	11	-4	13	20	7	38	52	14
4	8	7	-1	8	11	3	11	21	10	27	44	17
5	12	9	-3	12	14	2	18	21	3	42	51	9
6	12	9	-3	9	16	7	19	19	0	40	42	2
7	8	16	8	16	16	0	20	27	7	44	63	19
8	11	18	7	15	20	5	21	25	4	47	54	7
9	13	7	-6	15	15	0	16	17	1	44	48	4
10	7	12	5	8	16	8	14	19	5	29	39	10
11	15	6	-9	13	17	4	16	19	3	44	42	-2
12	13	13	0	11	18	7	18	25	7	42	57	15
13	13	11	-2	12	19	7	8	24	16	33	55	22
14	12	12	0	8	17	9	14	26	12	34	49	15
15	10	13	3	13	15	2	17	21	4	40	47	7
16	11	12	1	12	16	4	14	18	4	37	55	18
17	7	18	11	16	18	2	21	24	3	44	60	16
18	10	10	0	6	17	11	17	21	4	33	56	23
19	10	12	2	6	15	9	12	25	13	28	45	17
20	10	11	1	7	19	12	17	27	10	34	44	10
21	7	12	5	10	16	6	10	26	16	27	47	20
22	9	14	5	13	14	1	20	19	-1	42	47	5
23	8	6	-2	14	12	-2	14	15	1	36	44	8
24	8	10	2	14	14	0	19	21	2	41	54	13
25		10			16			21			42	
26	12	10	-2	14	14	0	17	18	1	43	46	3
27	12	10	-2	16	17	1	16	20	4	44	54	10
28	8	13	5	9	16	7	12	25	13	29	51	22
29	14	15	1	10	14	4	19	22	3	43	53	10
30	12	13	1	8	10	2	17	19	2	37	44	7
31	9	16	7	14	14	0	25	25	0	48	54	6
32	6	14	8	7	17	10	15	23	8	28	33	5
33	13	13	0	17	17	0	20	21	1	50	47	-3
34	10	14	4	9	15	6	13	24	11	32	39	7
M	10.42	11.97	1.61	11.70	15.38	3.67	16.27	21.79	5.55	38.39	49.15	10.97
SD	2.36	3.03	4.22	3.37	2.57	4.46	3.66	3.06	4.71	6.63	6.75	6.78
Max	15	18	11	17	20	12	25	27	16	50	63	23
Min	6	6	-9	6	9	-8	8	15	-1	27	33	-3

Note. List = Listening; Voc = Vocabulary; Read = Reading; Tot = Total; pre = pretest; post = posttest.

The listening tests in the two forms were not dependable in Pilot Study 1, so they were revised by deleting the vocabulary items that had not functioned well in Pilot Study 1. The listening test was more dependable in Pilot Study 2 even though

fewer items were included on the test. Adding more high quality items would have increased the dependability, but because adding items would lengthen the test, the length was maintained at 20 listening items.

In order to improve the dependability of the reading test, 12 strategy and comprehension questions were added in Pilot Study 2. Adding more items increased the dependability, but the test was still not highly dependable. Because the strategy items tended to have negative *B*-index values in Pilot Study 2, the 12 strategy items were deleted to make the test used in the main study more dependable.

Research question 2 asked, "To what extent do the students learn the class materials tested on the two revised CRT forms?" It was not possible to calculate DI with this design, but because the cut-points were set, the *B*-index was computed in order to determine how well the items differentiated the satisfactory and the unsatisfactory groups of students. The reading test given as a posttest had 14 items with negative values because the responses from only one student belonging to the unsatisfactory group greatly affected the values. The index was sensitive to the location of the cut-point.

All except two students were able to improve their test scores. Although very small gains were observed on the listening subtest, the students showed reasonably large gains on the vocabulary and reading subtests. The score could indicate that learning had taken place, and that the tests were sensitive enough to measure the students' learning. This result could be used as a construct validity argument.

CHAPTER 5

METHODS

This chapter consists of four sections. First, the participants in the main study are described. Second, the two CRT forms are explained in terms of the design and development procedures. Third, the procedures for conducting the study are described. Finally, the analyses used to answer the research questions are explained.

Participants

Three part-time teachers agreed to participate in the main study. They were in charge of developing their own syllabi and teaching materials based on a main unified syllabus developed by the program administrators, and teach 12 90-minute Reading and Listening Regular classes. The participants tested in the main study were 249 students who were taking seven Reading and Listening Regular classes. Their ages ranged from 18 to 19 years old. All the students were selected from Group B, which was made up of students who had average or below average scores on the in-house placement test (Figure 2). The students were attending one of the universities in Tokyo, but their English proficiency levels ranged from low intermediate to high intermediate, depending on how they had entered the university and what department they were in. Their major fields of study included literature, science, political science, tourism, and economics.

Instrumentation

The following steps were taken to develop the two forms of the Reading and Listening CRT: (a) identifying students' needs, (b) setting objectives, (c) developing a

test specification, (d) developing the CRT, (e) analyzing the CRT, and (f) revising the CRT. I thoroughly read the main syllabus that the administrators had developed. Bearing in mind my students' proficiency and my teaching capacity, I adapted the objectives from the goals that were stated in the main syllabus and customized them to better meet my students' needs by matching the difficulty of the tasks and the students' proficiency so that the teaching effect could be maximized. Once the syllabus and class objectives were developed, I revised the test specification that had been developed in the pilot studies (Appendix I). Based on the results of the pilot studies, the CRT forms were revised as described in Chapter 5 and finalized for the present study.

The chapters tested in the main study are shown in Table 14 and the test structure is displayed in Table 15. Each test form was made up of 70 items, which were divided into vocabulary (k = 25), listening (k = 20), and reading (k = 25) subtests. The multiple-choice item type was chosen because listening and reading are receptive skills. In order to select the vocabulary items, 25 words that were classified in "range 3" or words for the lower level learners, and "range 4" or words for the next level

Table 14

Chapters Tested in Form A and Form B

Form A			Form B		
Vocabulary	Listening	Reading	Vocabulary	Listening	Reading
Chapter 2	Chapter 1	Chapter 1	Chapter 1	Chapter 2	Chapter 2
Chapter 4	Chapter 3	Chapter 3*	Chapter 3	Chapter 4	Chapter 3*
Chapter 5	Chapter 6*	Chapter 5	Chapter 6*	Chapter 6*	Chapter 4
Chapter 6*	Chapter 7	Chapter 7	Chapter 7	Chapter 8	Chapter 6
Chapter 8*		Chapter 8*	Chapter 8*		Chapter 8*

Note. * = A set of five items appears on both test forms.

Table 15

Reading and Listening CRT Structure

	Vocabulary	Listening	Reading
# of chapters	5 chapters	4 chapters	5 chapters
# of items per chapter	5 receptive	5 comprehension	5 comprehension
# of items	25	20	25
# of linking items	2 sets of 5 items	1 set of 5 items	2 sets of 5 items

learners based on the frequency count estimated with WordClassifier [version 2.5] (Denies, 2002) software were selected from five chapters in the textbook. In the vocabulary section, the students read a sentence with a target word underlined and choose the word that is closest in meaning to the target word from four choices. The following is a sample vocabulary item:

Schools keep track of students' GPA.

a. evaluate

b. control

c. record

d. delete

The students had to read the sentence and underlined words to answer the question. The correct choice was c, and the remaining three choices were distractors. The same ten items in the vocabulary section appeared in both of the CRT forms.

The listening section had four parts. Chapters 1, 3, 6, and 7 were tested on Form A and chapters 2, 4, 6, and 8 were tested on Form B, and five items were written for each chapter for a total of 20 items. Students watched the video and answered five comprehension questions in each part. The following is a sample listening comprehension item:

How was the speaker’s experience in high school?

a. He had to take many tests.

b. He had to go to prison.

c. He had few friends.

d. He was tired of high school.

The correct choice was d and the remaining choices were distractors. Five items associated with one part of the listening section on Chapter 6 appeared in both test forms.

The reading section had five parts. Within each part, the students were required to read a passage that was taken from the textbook and answer five questions designed to measure their comprehension of the passages. The following is a sample reading comprehension item:

What do you have to do to grow?

a. Travel around the world

b. Meet many people

c. Take chances

d. Make many mistakes

The students had to read the passage and choose an option that was true about the content of the passage. The correct option was c. The same ten items taken from the passages in Chapters 3 and 8 appeared on both test forms (see Appendix A and Appendix B for the finalized Reading and Listening CRT forms).

Procedures

Two groups, Group A (n = 143) and Group B (n = 106), were formed. Group A was made up of four classes (n = 28, 43, 37, 35) and Group B was made up of three classes (n = 35, 37, 34). Test forms A and B were administered to Groups A and B, respectively, as a diagnostic pretest at the beginning of the semester. The two Group A teachers informed the students that their pretest scores would have no effect on their grades and that the pretest was given for diagnostic and research purposes only. The Group B teacher (the researcher) announced to the students in the three classes that if they did well on the pretest, it would improve their participation grades. He thought that this message would prevent them from not taking the test seriously and that it would be of interest to see if it affected the test results. The diagnostic tests were administered as described in the proctor manual in Appendix C. Thirteen 90-minute lessons were held during the semester, in which the teachers were assigned to teach eight units from the assigned textbook; however, although the overall teaching methods were the same, the pace of the lessons varied, depending on how feasible the activities were for their students.

At the end of the semester, the test form that the students did not take at the beginning was administered as a final examination to assess their mastery of the materials. Thus, it was a counterbalanced design with two forms. The teachers proctored the 90-minute final exam following the proctor manual shown in Appendix C. Their posttest scores accounted for 40% of their final class grades, while attendance and participation grades accounted for 30% each. The attendance policy of the Group B teacher was that if the students were on time to start a lesson, they were not penalized, but if they arrived after attendance had been taken, they were considered late. Arriving late three times was counted as one absence. To assess the

students' participation, all the activity sheets that the students worked on over the semester were collected as a portfolio and evaluated using a holistic scale (Appendix G). The final examination score, attendance, and participation points were added to calculate the total class points. If the students' total class points were close to or below 60, I examined their overall performance by looking holistically at whether or not they had score gains and had passed the cut-points on the posttest, their eagerness to attend the class, and their participation in class activities. If they did not perform well based on these criteria, I gave a failing grade, D. Students with total points from 60 to 69, from 70 to 79, from 80 to 89, and from 90 to 100 received a C, B, A, and S, respectively and earned one credit for the class.

Analysis

In order to answer research question 1, descriptive statistics were calculated using SPSS [version 17.0]. Both norm-referenced and criterion-referenced item analyses were conducted to determine item functioning. Two norm-referenced item statistics, item facility and item discrimination, and two criterion-referenced item statistics, the *B*-index, and difference index, were calculated.

To answer research question 2, 12 univariate generalizability G studies were conducted to examine the extent to which the following facets contributed to the total variance: persons, items, and sections. This balanced design was persons crossed with items nested within sections: *p X (i:s)*, treating persons as a random facet, and items and sections as fixed facets. GENOVA [version 2.1] (Crick & Brennan, 1983) was used for these G studies. To investigate sources of variance for the persons and classes random facets and the items and sections fixed facets, 12 G studies with an unbalanced *(p:c) X (i:s)* design was conducted with urGENOVA [version 2.1]

(Brennan, 2001b). Also, by treating the three subtests that were designed to gauge students' mastery of vocabulary, listening, and reading as a facet rather than as separate constructs, four G studies were conducted with an unbalanced *p X (i:s)* design with mGENOVA [version 2.1] (Brennan, 2001c). Then, multivariate G theory was applied to simultaneously estimate the variance components and relatedness of all the facets in the vocabulary, listening, and reading subtests using the mGENOVA program. Four multivariate generalizability studies with a $p^{\cdot} X i^{\circ}$ design were conducted. In this design, the same persons took the vocabulary, listening, and reading subtests. This design included three dimensions: vocabulary, listening, and reading. Persons and items were treated as random and fixed facets, respectively. Two multivariate generalizability studies with a $p^{\cdot} X (i^{\circ}: s^{\cdot})$ design were conducted. The same persons took the same vocabulary, listening, and reading subtests twice as a pretest and posttest, but each subtest was made up of a different set of items. The random facets are persons and items, and the fixed facet is subtests.

Four FACETS analyses were conducted with three facets: persons, items, and subtests. Each form given as a pretest and posttest was analyzed separately. Misfitting persons and items were identified using an infit mean-square statistic criterion of .70 to 1.30.

Research question 3 was answered by calculating norm-referenced reliability coefficients and criterion-referenced dependability indices. First, a Cronbach alpha reliability estimate was calculated for each subtest on each form. Second, Brown's (1990b) short-cut dependability formula was used to estimate the dependability for the Form A and Form B vocabulary, listening, and reading subtests. Third, phi lamda dependability ($\Phi(\lambda)$) indices were calculated to estimate decision consistency when the cut-points were set at 70%, 60%, and 50% for Form A and Form B as pretests and

posttests. Fourth, the confidence interval (*CI*) was calculated to estimate a band of scores that examinees would get if the tests were hypothetically administered repeatedly. The *CI* is similar to *SEM* in NRT reliability (Brown & Hudson, 2002, p. 184). Fifth, 16 univariate D studies were conducted with the persons, items, and sections facets on the Form A and Form B vocabulary, listening, and reading pretests and posttests. Sixth, four univariate D studies were conducted with the persons, items, and subtests facets on the Form A and Form B pretests and posttests. Seventh, four multivariate D studies and two multivariate D studies were conducted with $p^{\cdot} \times I^{\circ}$ and $p^{\cdot} \times (I^{\circ}: S^{\cdot})$ designs, respectively, to estimate univariate and multivariate generalizability coefficients for absolute decisions.

Research question 4 was answered by conducting two repeated-measures MANOVAs. Prior to the analysis, the following assumptions were checked: missing data, univariate and multivariate outliers, normality, and milticollinearity. Missing data were treated as incorrect responses. Univariate and multivariate outliers were deleted from the analysis. The skewed distributions were not transformed because that would change the means. Multicollinearity was analyzed by examining the correlation matrix. A repeated-measures MANOVA was used to compare students' score gains on the listening, vocabulary, and reading subtests twice for Group A and Group B. The within-subjects variables were time (pretest and posttest) and subtests (vocabulary, listening, and reading). The *p*-value was set at .05. In addition, two ANOVA-like FACETS analyses were conducted with the three facets: persons, items, and occasions. Form A pretest and Form B posttest taken by Group A, and Form B pretest and Form A posttest taken by Group B were run separately. Within each subtest, five items that appeared in both of the test forms were used as common items in order to equate the two test forms. Thus, it is a common item design (Wolfe, 2000). Table 16 provides a

summary of the analyses and shows the statistics or arguments used to answer the research questions posed in this study.

Table 16

Analyses of the Reading and Listening CRT

Research question	Analysis
1: Item analyses	Descriptive statistics; norm-referenced item analyses such as item facility and item discrimination; criterion-referenced item analyses such as the *B*-index, and difference index
2: Facets variance	Univariate and multivariate G studies with the *p X (i:s), (p:c) X (i:s), p X (i:s), p˙ X i˚*, and *p˙ X (i˚: s˙)* designs; FACETS analyses
3: Dependability	Brown's short-cut dependability, phi lamda dependability, confidence interval, univariate and multivariate D studies with the *p X (i:s), p X (i:s), p˙ X I˚, p˙ X (I˚: S˙)* designs
4: Validity	Content validity; Construct validity (difference index, divergent/convergent validity, repeated-measures MANOVA, ANOVA-like FACETS analyses)

CHAPTER 6
RESULTS

The results of the main study are reported in the order of the research questions posed in this study.

Research Question 1: Item Effectiveness

Research Question 1 asked the extent to which the criterion-referenced items function effectively in this study. To answer the research question, the descriptive statistics are first displayed to show the data set used to conduct item analyses. Then, norm-referenced and criterion-referenced items analyses are presented to show the soundness of the items.

Descriptive Statistics for the Criterion-Referenced Test Forms

Tables 17, 18, 19, and 20 show the descriptive statistics for the Form A and Form B pretests and posttests taken by Group A and Group B. The pretest means were lower than the midpoints, indicating that the students had not mastered the objectives before taking the Reading and Listening Regular class. The posttest means were higher than the midpoints, indicating that the students mastered the objectives to a greater extent. Overall, their posttest scores were higher than the pretest scores. Group A took Form A as a pretest and Form B as a posttest, and improved their scores by 5.23, 2.54, 5.55, and 13.32 for the vocabulary, listening, reading, and total scores, respectively. Group B took Form B as a pretest and Form A as a posttest, and improved their scores by 5.09, 5.02, 5.74, and 15.85 for the vocabulary, listening, reading, and total scores.

The minimum and maximum total scores on the pretests ranged widely from 3 to 48 on Form A and from 13 to 51 on Form B. The larger sample size of 144 in Group A possibly led to an even greater range of proficiency levels as compared to the one in Group B. The greatest variance occurred on the Form A pretest, which had a standard deviation of 9.18.

The minimum and maximum total scores on the posttests ranged from 24 to 62 on Form A and 16 to 62 on Form B. The standard deviations were fairly large and even greater than the ones on the pretests, possibly because of the amount of effort the students put into studying for the posttest.

Table 17

Descriptive Statistics for the Form A Pretest

	Vocabulary	Listening	Reading	Total
k	25	20	25	70
N	144	144	144	144
M	9.97	9.38	9.67	29.02
SD	4.39	3.22	4.78	9.18
Min	0	0	0	3
Max	20	17	20	48
Skew	.14	-.33	-.22	-.45
SES	.20	.20	.20	.20
Kurt	-.64	.22	-.42	.00
SEK	.40	.40	.40	.40

Table 18

Descriptive Statistics for the Form B Pretest

	Vocabulary	Listening	Reading	Total
k	25	20	25	70
N	103	103	103	103
M	10.53	9.02	11.17	30.72
SD	3.67	2.75	3.80	8.25
Min	3	2	2	13
Max	22	16	21	51
Skew	.15	.00	.14	.05
SES	.24	.24	.24	.24
Kurt	-.11	-.14	-.28	-.53
SEK	.47	.47	.47	.47

Table 19

Descriptive Statistics for the Form A Posttest

	Vocabulary	Listening	Reading	Total
k	25	20	25	70
N	106	106	106	106
M	15.62	14.04	16.91	46.57
SD	4.63	2.78	4.14	9.48
Min	5	5	7	24
Max	24	19	24	63
Skew	-.36	-.70	-.55	-.42
SES	.24	.24	.24	.24
Kurt	-.91	.93	-.40	-.63
SEK	.47	.47	.47	.47

Table 20

Descriptive Statistics for the Form B Posttest

	Vocabulary	Listening	Reading	Total
k	25	20	25	70
N	143	143	143	143
M	15.20	11.92	15.22	42.34
SD	4.62	3.20	5.51	12.03
Min	3	3	3	16
Max	24	19	24	62
Skew	-.33	-.27	-.31	-.22
SES	.20	.20	.20	.20
Kurt	-.72	-.10	-.92	-1.08
SEK	.40	.40	.40	.40

Norm-Referenced and Criterion-Referenced Item Analyses

Tables 21 and 22 display the values for item facility (IF), item discrimination (ID), and the difference index (DI) for the Form A pretest and posttest, and the Form B pretest and posttest. Although IF and ID are norm-referenced statistics, they provide valuable information for criterion-referenced items. For diagnostic tests, IF values should be low enough to enable students to perform well on an achievement test. For instance, the IF values for items 8, 48, 50 on the Form A pretest, and items 13, 32, 36, and 37 on the Form B pretest were higher than .70, indicating that most students had already learned the items before instruction. IF values should not be low for achievement tests because they indicate that the items have not been learned even

after instruction. For instance, one item had an IF value of .11 because it was excessively difficult, was not properly taught, or did not function properly. Because the purpose of CRTs is not to discriminate students' mastery, ID is not an adequate indicator of criterion-referenced items. There were five items with negative ID values on Forms A and B. DI is the most crucial statistic for criterion-referenced items because it shows the degree of the students' mastery of the items. Only six items had negative DI values on Forms A and B. Notice that item 16 on the Form A pretest had a ID value of .02. From a norm-referenced perspective, it is a poorly performing item because it does not discriminate students' ability. However, from a criterion-referenced perspective, it is a sound item because the DI value of .52 shows that students have learned the item. Conversely, item 37 on Form B is an excellent norm-referenced item but a poorly performing criterion-referenced item because the DI value indicated that the students were unable to learn it. Overall, most of the items worked effectively for criterion-referenced purposes because the students learned most of them.

Tables 23 and 24 show the *B*-indices for the Form A and Form B Pretests and the Form A and Form B Posttests. Depending on the location of the cut-point, the *B*-index values differed, but 13 items displayed negative values on the pretests, and four items had negative values on the posttests. Thus, the items on the posttests were more likely to differentiate students who passed and failed the CRT. Recall that items 1 and 32 on Form A and items 1, 13, 33, and 37 on Form B had negative DI values. Notice the *B*-index values for those items. There was a tendency that the items with low DI values also had weak power to differentiate students who passed and failed the tests. For instance, the DI value for item 33 on Form B was -.05, and the *B*-index

Table 21

Item Facility, Item Discrimination, and Difference Index for the Form A Pretest and Posttest

	Pretest (Form A)		Posttest (Form A)		
Item	IF	ID	IF	ID	DI
1	.44	.19	.42	.13	-.02
2	.42	.27	.80	.36	.38
3	.33	.24	.75	.42	.41
4	.50	.53	.80	.43	.30
5	.41	.45	.78	.32	.37
6	.59	.49	.74	.55	.15
7	.40	.43	.63	.38	.24
8	.75	.45	.87	.51	.12
9	.37	.25	.59	.37	.23
10	.24	.33	.35	.35	.11
11	.61	.45	.83	.44	.22
12	.47	.35	.66	.43	.20
13	.38	.29	.43	.24	.05
14	.40	.31	.41	.09	.01
15	.34	.48	.72	.57	.38
16	.10	.02	.62	.55	.52
17	.26	.34	.42	.32	.16
18	.51	.56	.66	.39	.15
19	.44	.30	.53	.41	.08
20	.43	.49	.58	.54	.14
21	.16	.04	.59	.42	.43
22	.15	.15	.31	.43	.16
23	.44	.61	.65	.57	.21
24	.51	.63	.83	.49	.32
25	.31	.47	.64	.49	.33
26	.46	.36	.75	.36	.30
27	.59	.47	.93	.22	.34
28	.48	.36	.71	.32	.24
29	.55	.44	.88	.34	.33
30	.11	.01	.71	.42	.60
31	.52	.40	.75	.45	.22
32	.45	.29	.30	.24	-.15
33	.30	.06	.58	.39	.28
34	.35	.25	.66	.45	.31
35	.44	.34	.60	.31	.17
36	.67	.46	.90	.31	.22
37	.69	.52	.70	.23	.01
38	.41	.38	.71	.44	.30
39	.34	.35	.59	.34	.25
40	.42	.26	.49	.35	.07
41	.62	.44	.88	.24	.26
42	.23	.06	.44	.24	.21
43	.67	.36	.92	.26	.26
44	.64	.53	.94	.36	.30
45	.46	.27	.60	.28	.15
46	.33	.32	.38	.14	.04
47	.62	.45	.77	.35	.16
48	.72	.47	.90	.34	.17
49	.67	.45	.79	.58	.12
50	.71	.59	.89	.33	.18
51	.26	.22	.51	.52	.26
52	.46	.41	.82	.32	.36
53	.31	.15	.63	.38	.33
54	.37	.52	.66	.55	.29
55	.51	.44	.82	.37	.31
56	.40	.54	.71	.43	.30

(Table 21 continues)

(Table 21 continued)

	Pretest (Form A)		Posttest (Form A)		
Item	IF	ID	IF	ID	DI
56	.40	.54	.71	.43	.30
57	.59	.66	.82	.44	.23
58	.53	.54	.87	.48	.33
59	.38	.49	.64	.44	.27
60	.39	.32	.57	.28	.18
61	.35	.59	.88	.46	.53
62	.38	.55	.80	.46	.43
63	.27	.49	.57	.21	.30
64	.27	.41	.67	.45	.40
65	.17	.41	.33	.18	.16
66	.26	.29	.48	.21	.22
67	.24	.29	.69	.52	.45
68	.15	.19	.54	.47	.39
69	.11	.12	.58	.35	.47
70	.22	.36	.60	.25	.38

Table 22

Item Facility, Item Discrimination, and Difference Index for the Form B Pretest and Posttest

	Pretest(Form B)		Posttest(Form B)		
Item	IF	ID	IF	ID	DI
1	.44	.26	.39	.06	-.05
2	.56	.50	.83	.37	.27
3	.14	.27	.45	.46	.31
4	.45	.29	.87	.37	.42
5	.35	.36	.49	.34	.14
6	.19	.40	.63	.54	.44
7	.56	.32	.61	.40	.04
8	.36	.26	.60	.33	.24
9	.85	.29	.92	.52	.06
10	.49	.33	.81	.53	.32
11	.26	.22	.74	.53	.48
12	.35	.33	.50	.34	.15
13	.71	.36	.70	.25	-.01
14	.51	.50	.70	.38	.18
15	.53	.48	.66	.28	.13
16	.37	.21	.79	.42	.42
17	.58	.20	.82	.45	.23
18	.02	-.04	.65	.58	.63
19	.42	.33	.74	.28	.32
20	.33	.32	.42	.16	.09
21	.25	.18	.70	.46	.44
22	.21	-.06	.34	.41	.13
23	.50	.50	.65	.46	.15
24	.66	.48	.82	.49	.16
25	.44	.44	.76	.60	.32
26	.19	.17	.34	.30	.15
27	.38	.40	.56	.52	.18
28	.42	.30	.70	.44	.28
29	.71	.31	.73	.34	.03
30	.46	.38	.54	.24	.08
31	.83	.26	.98	.34	.15
32	.65	.35	.73	.45	.08
33	.17	.06	.11	.00	-.05
34	.35	.18	.79	.43	.44

(Table 22 continues)

(Table 22 continued)

	Pretest(Form B)		Posttest(Form B)		
Item	IF	ID	IF	ID	DI
35	.12	.19	.62	.11	.51
36	.71	.39	.92	.47	.21
37	.81	.30	.79	.37	-.02
38	.56	.36	.69	.45	.12
39	.33	.36	.66	.50	.33
40	.39	.24	.51	.41	.13
41	.18	.22	.53	.03	.35
42	.43	.24	.67	.40	.24
43	.40	.37	.72	.25	.33
44	.55	.46	.68	.35	.13
45	.39	.41	.52	.52	.13
46	.69	.42	.90	.41	.21
47	.57	.55	.77	.46	.20
48	.59	.49	.66	.39	.07
49	.19	.10	.41	.39	.22
50	.30	.15	.61	.40	.31
51	.34	.26	.42	.37	.08
52	.48	.55	.84	.60	.37
53	.47	.32	.75	.59	.29
54	.37	.32	.71	.57	.34
55	.50	.39	.72	.52	.22
56	.65	.38	.83	.44	.18
57	.34	.32	.78	.48	.44
58	.40	.28	.60	.38	.20
59	.44	.35	.78	.49	.34
60	.71	.29	.82	.40	.11
61	.38	.49	.58	.55	.20
62	.55	.38	.69	.60	.13
63	.50	.35	.71	.54	.22
64	.46	.26	.67	.47	.21
65	.47	.43	.69	.58	.23
66	.34	.07	.57	.24	.23
67	.41	.18	.91	.38	.50
68	.36	.24	.67	.55	.31
69	.23	.27	.56	.42	.33
70	.45	.24	.52	.20	.07

value was -.17 at the cut-point of 70% on the pretest and .03 at the cut-point of 70% on the posttest. Overall, most of the items had adequate differential power based on the *B*-index value.

Research Question 2: The Contribution of the Test Facets

Research Question 2 asked the extent to which the facets of persons, items, sections, classes, and subtests contribute to the total score variation in the two CRT

Table 23

B-indices for the Form A and Form B Pretests

	Form A			Form B		
Item	*B*-index (.7)	*B*-index (.6)	*B*-index (.5)	*B*-index (.7)	*B*-index (.6)	*B*-index (.5)
1	.06	.10	.18	.57	.11	.18
2	.26	.22	.28	.45	.51	.45
3	.18	.17	.14	.37	.23	.21
4	.53	.50	.47	.56	.34	.17
5	.27	.45	.40	.15	.37	.26
6	.43	.39	.38	.31	.48	.26
7	.63	.46	.28	.45	.28	.27
8	.27	.25	.33	.65	.28	.20
9	.66	.39	.19	.15	.09	.17
10	.44	.38	.26	.52	.21	.20
11	.41	.47	.35	.75	.16	.16
12	.56	.48	.36	.15	.14	.31
13	.47	.27	.24	.30	.26	.29
14	-.06	.10	.25	.50	.41	.48
15	.34	.43	.46	.48	.47	.40
16	-.11	.03	-.01	.13	.19	.14
17	.42	.31	.27	.43	.25	.15
18	.52	.44	.50	-.02	-.02	-.03
19	.58	.35	.25	.59	.14	.39
20	.42	.47	.57	.17	.39	.29
21	.18	-.03	.01	.76	.25	.18
22	.19	.03	.09	-.22	.06	-.04
23	.59	.52	.56	.51	.36	.46
24	.52	.59	.56	.35	.32	.31
25	.72	.46	.30	.57	.35	.27
26	.43	.26	.36	.01	.15	.12
27	.45	.37	.41	.65	.36	.34
28	.49	.38	.29	.40	.20	.28
29	.50	.39	.31	.10	.23	.28
30	-.04	-.04	.04	.57	.55	.42
31	.53	.39	.34	.17	.17	.11
32	.10	.31	.29	.37	.33	.22
33	.10	.07	-.02	-.17	.10	.05
34	.21	.08	.23	.05	.14	.11
35	.28	.21	.35	.30	.10	.09
36	.36	.37	.29	.10	.35	.32
37	.35	.39	.43	.20	.19	.20
38	.48	.32	.30	.25	.26	.24
39	.30	.27	.25	.70	.33	.30
40	.22	.21	.20	.01	.23	.21
41	.34	.34	.30	.23	.15	.06
42	-.08	-.02	.04	.18	.34	.10
43	.29	.16	.25	.63	.51	.27
44	.32	.38	.45	.47	.36	.38
45	.26	.40	.25	.43	.41	.45
46	.40	.16	.20	.12	.31	.32
47	.41	.32	.36	.45	.45	.50
48	.15	.26	.22	.43	.36	.38
49	.20	.20	.18	.22	.03	.08

(Table 23 continues)

(Table 23 continued)

	Form A			Form B		
Item	*B*-index (.7)	*B*-index (.6)	*B*-index (.5)	*B*-index (.7)	*B*-index (.6)	*B*-index (.5)
50	.31	.34	.38	-.11	-.16	-.09
51	.48	.19	.03	.27	.26	.27
52	.12	.38	.23	.55	.57	.57
53	-.02	-.05	.00	.56	.31	.24
54	.52	.55	.39	.45	.16	.27
55	.36	.26	.30	.32	.28	.32
56	.48	.57	.48	.37	.29	.29
57	.43	.48	.47	.69	.40	.27
58	.34	.36	.34	.42	.53	.22
59	.51	.48	.42	.59	.35	.12
60	.49	.28	.30	.31	.08	.20
61	.54	.69	.59	.44	.48	.46
62	.66	.60	.56	.26	.41	.32
63	.62	.42	.42	.53	.34	.37
64	.77	.54	.35	.15	.25	.34
65	.57	.41	.38	.56	.38	.24
66	.18	.37	.27	.06	.06	.02
67	.20	.34	.33	.41	.31	.04
68	.44	.13	.14	.25	.24	.16
69	.03	-.01	.09	.18	.12	.23
70	.37	.35	.28	-.26	.20	.15

Table 24

B-indices for the Form A and Form B Posttest

	Form A			Form B		
Item	*B*-index	*B*-index (.6)	*B*-index	*B*-index (.7)	*B*-index	*B*-index (.5)
1	.12	.15	.05	.03	.02	.09
2	.24	.22	.22	.27	.26	.27
3	.30	.33	.28	.41	.41	.32
4	.28	.26	.36	.20	.21	.19
5	.28	.27	.19	.35	.26	.30
6	.36	.39	.40	.56	.45	.41
7	.28	.34	.39	.30	.40	.44
8	.20	.29	.36	.32	.29	.25
9	.31	.31	.25	.21	.26	.31
10	.33	.22	.22	.44	.43	.44
11	.27	.35	.35	.42	.47	.42
12	.46	.42	.30	.29	.29	.30
13	.21	.21	.20	.17	.07	.16
14	.15	.04	-.02	.41	.30	.37
15	.36	.52	.56	.15	.25	.38
16	.49	.52	.52	.32	.36	.36
17	.23	.23	.19	.22	.34	.36
18	.30	.31	.43	.54	.56	.54
19	.27	.40	.43	.30	.20	.24
20	.38	.44	.54	.21	.05	.09
21	.31	.39	.34	.46	.46	.35
22	.43	.32	.35	.33	.28	.31

(Table 24 continues)

(Table 24 continued)

	Form A			Form B		
Item	*B*-index (.7)	*B*-index (.6)	*B*-index (.5)	*B*-index (.7)	*B*-index (.6)	*B*-index (.5)
23	.44	.53	.51	.38	.43	.47
24	.27	.35	.35	.31	.41	.40
25	.49	.43	.50	.49	.49	.60
26	.20	.16	.62	.34	.21	.23
27	.05	.19	.28	.49	.49	.40
28	.28	.19	.57	.29	.36	.48
29	.20	.30	.22	.23	.14	.30
30	.32	.46	.40	.16	.19	.25
31	.38	.50	.44	.07	.08	.15
32	.17	.08	-.03	.34	.36	.36
33	.26	.31	.26	.03	-.04	-.01
34	.40	.58	.35	.29	.23	.40
35	.27	.07	.29	-.02	.11	.18
36	.07	.32	.60	.23	.25	.40
37	.02	.18	.56	.26	.21	.39
38	.32	.46	.57	.37	.38	.51
39	.37	.33	.45	.50	.40	.41
40	.32	.38	.34	.43	.38	.32
41	.12	.12	.22	-.01	.04	-.02
42	.23	.23	.28	.35	.32	.27
43	.12	.26	.27	.12	.21	.09
44	.11	.29	.47	.29	.32	.24
45	.19	.43	.29	.46	.48	.43
46	.13	.03	.17	.25	.21	.31
47	.29	.32	.29	.40	.35	.32
48	.17	.20	.30	.35	.28	.28
49	.33	.49	.59	.32	.35	.30
50	.22	.14	.22	.36	.26	.36
51	.51	.45	.40	.32	.32	.27
52	.13	.24	.28	.42	.45	.54
53	.23	.33	.33	.50	.55	.62
54	.46	.54	.57	.45	.54	.55
55	.20	.29	.35	.40	.48	.44
56	.39	.28	.35	.31	.28	.37
57	.24	.39	.35	.44	.37	.41
58	.22	.35	.54	.29	.32	.27
59	.36	.39	.27	.37	.44	.43
60	.21	.19	.18	.31	.28	.22
61	.20	.32	.41	.57	.49	.45
62	.28	.36	.46	.51	.54	.54
63	.17	.04	.18	.44	.52	.52
64	.35	.42	.52	.39	.43	.45
65	.14	.16	.04	.42	.60	.52
66	.19	.17	.22	.23	.15	.15
67	.38	.50	.54	.29	.24	.28
68	.42	.49	.43	.48	.51	.51
69	.32	.31	.35	.40	.34	.34
70	.14	.19	.16	.20	.14	.14

forms. To answer the question, univariate generalizability studies with the *p X (i:s)*, *(p:c) X (i:s)*, and *p X (i:s)* designs were conducted followed by multivariate generalizability studies with the $p^{\cdot}$ *X* i°, and $p^{\cdot}$ *X* $(i^{\circ}: s^{\cdot})$ designs. The studies revealed the percentages of the persons, items, sections, and subtests variance contributing to the total variance. In addition, four FACETS analyses with the person, item, and subtest facets were conducted to examine the separation of the facets.

Univariate Generalizability Studies

Tables 25, 26, 27, and 28 show the variance components and percentages of variance components for the vocabulary, listening, and reading Form A and Form B pretests and posttests with the G study design of *p* X (*i*:*s*). The components for the persons variance were small across all the subtests. Because the purpose of criterion-referenced testing is not to spread out the examinees' scores, a small proportion of persons variance is desirable. Table 2 showed that the readability of the passages, dialogs, and the number of words differed to some extent. However, these differences did not create variance in the test score, as there were no sections variance components across the subtests. The items variance components were large on the pretests and were even greater on the posttests, suggesting that the item difficulty varied a great deal. The persons-by-sections interactions had almost no effects across the subtests. The greatest variability was due to the interaction effects on the pretests. This indicates that the students responded to the items in rather different ways.

Table 25

Three G Studies for the Vocabulary, Listening, and Reading Form A Pretest with the p X (i:s) Design

Source	Vocabulary		Listening		Reading	
persons	.005635	3%	.002619	1%	.010918	6%
sections	.000000	0%	.000000	0%	.000000	0%
items:s	.073306	39%	.065471	36%	.078298	41%
ps	.000000	0%	.000946	1%	.000000	0%
pi:s	.107522	58%	.110392	62%	.099771	53%
Total	.186463	100%	.179428	100%	.188987	100%

Table 26

Three G studies for the Vocabulary, Listening, and Reading Form B Pretest with the p X (i:s) Design

Source	Vocabulary		Listening		Reading	
persons	.003295	2%	.0000793	0%	.004484	2%
sections	.000000	0%	.0000000	0%	.000000	0%
items:s	.066421	39%	.0779364	45%	.075127	41%
ps	.000000	0%	.0023621	1%	.000000	0%
pi:s	.099598	59%	.0946364	54%	.105844	57%
Total	.169314	100%	.1750142	100%	.185455	100%

Table 27

Three G studies for the Vocabulary, Listening, and Reading Form A Posttest with the p X (i:s) Design

Source	Vocabulary		Listening		Reading	
persons	.005053	2%	.004343	2%	.005369	2%
sections	.000000	0%	.000000	0%	.000000	0%
items:s	.152568	62%	.153747	62%	.162487	65%
ps	.000000	0%	.000000	0%	.000000	0%
pi:s	.087998	36%	.091301	37%	.080909	33%
Total	.245619	100%	.249391	100%	.248765	100%

Table 28

Three G studies for the Vocabulary, Listening, and Reading Form B Posttest with the p X (i:s) Design

Source	Vocabulary		Listening		Reading	
persons	.005622	3%	.002366	1%	.009580	4%
sections	.000000	0%	.000000	0%	.000000	0%
items:s	.120261	54%	.124290	55%	.127296	55%
ps	.000000	0%	.000000	0%	.000000	0%
pi:s	.095683	43%	.098263	44%	.096060	41%
Total	.221566	100%	.224919	100%	.232936	100%

Tables 29, 30, 31 and 32 display the variance components and percentages of variance components for the vocabulary, listening, and reading pretest with the G study design of *(p:c) X (i:s)*. Group A, which consisted of four classes (*n* = 28, 43, 39, 35), took the Form A pretest and the Form B posttest. Group B consisted of three classes (*n* = 35, 37, 34), took the Form B pretest and the Form A posttest. The classes variance components on the pretests were small, ranging from 1% to 6%. This shows that there was almost no classes effect that led to the variability in the test scores. The variance components for classes on the Form A posttest were small, but the ones on the Form B posttest were large, ranging from 6% to 14%. This occurred because Group A was a larger sample than Group B, and some students studied for the posttest but others did not. The persons variance components were small, ranging from 2% to 9%. No sections effects were presented in the results of the *p X (i:s)* design. However, the Form A reading subtest on the pretest accounted for 9% of the sections variance. This occurred because the teacher reported that sufficient administration time was not given to her students in one class so they could not finish all the sections. Compared with the results of the *p X (i:s)* design, the items variance components were much smaller, ranging from 4% to 13%. There were almost no classes-by-sections, classes-by-items nested within sections, and persons-by-sections nested within classes interactions. However, the persons-by-items nested within classes-by-sections

Table 29

Three G Studies for the Vocabulary, Listening, and Reading Form A Pretest with the (p:c) X (i:s) Design

Source	Vocabulary		Listening		Reading	
classes	.00259	1%	.00767	3%	.01458	6%
persons:c	.02173	9%	.00985	4%	.01515	6%
sections	.00000	0%	.00000	0%	.02118	9%
items:s	.01949	8%	.02275	9%	.00945	4%
cs	.00037	0%	.00000	0%	.00608	2%
ci:s	.00149	1%	.00558	2%	.00234	1%
ps:c	.00037	0%	.00568	2%	.02027	8%
pi:cs	.19502	81%	.20306	80%	.15634	64%
Total	.24106	100%	.25459	100%	.24539	100%

Table 30

Three G Studies for the Vocabulary, Listening, and Reading Form B Pretest with the (p:c) X (i:s) Design

Source	Vocabulary		Listening		Reading	
classes	.00510	2%	.00202	1%	.00737	3%
persons:c	.01001	4%	.00781	3%	.00862	3%
sections	.00000	0%	.00000	0%	.00012	0%
items:s	.03617	14%	.04573	18%	.01326	5%
cs	.00000	0%	.00128	0%	.00000	0%
ci:s	.00085	0%	.00000	0%	.00599	2%
ps:c	.00120	0%	.00000	0%	.00570	2%
pi:cs	.19715	79%	.19929	78%	.21046	84%
Total	.25048	100%	.25613	100%	.25152	100%

Table 31

Three G Studies for the Vocabulary, Listening, and Reading Form A Posttest with the (p:c) X (i:s) Design

Source	Vocabulary		Listening		Reading	
classes	.00543	2%	.00376	2%	.00398	2%
persons:c	.02186	9%	.00624	3%	.01730	8%
sections	.00000	0%	.00347	0%	.00000	0%
items:s	.02516	10%	.02653	12%	.02363	11%
cs	.00149	1%	.00043	0%	.00010	0%
ci:s	.00388	2%	.00032	0%	.00273	1%
ps:c	.00782	3%	.00937	4%	.00226	1%
pi:cs	.17454	73%	.16276	76%	.17232	78%
Total	.24018	100%	.21288	100%	.22232	100%

Table 32

Three G Studies for the Vocabulary, Listening, and Reading Form B Posttest with the (p:c) X (i:s) Design

Source	Vocabulary		Listening		Reading	
classes	.02037	8%	.01475	6%	.03555	14%
persons:c	.01069	4%	.00395	2%	.01195	5%
sections	.00000	0%	.00000	0%	.00000	0%
items:s	.02804	11%	.03240	13%	.01355	5%
cs	.00056	0%	.00218	1%	.00120	0%
ci:s	.00242	1%	.00549	2%	.00148	1%
ps:c	.00328	1%	.00204	1%	.01388	6%
pi:cs	.18353	74%	.18921	76%	.17154	69%
Total	.24889	100%	.25002	100%	.24915	100%

interactions had the largest variance components, ranging from 64% to 84% of the total variance.

Tables 33 and 34 show the results of four G studies for the pretests and posttests with the *p X (i:s)* design. A small amount of the persons variance components was found in the pretests and posttests, ranging from 4% to 11%. The variance components due to subtests on the Form A pretest and posttest were .00064 and .00029, indicating slight variability in the difficulty of the subtests. The items variance components accounted for 9% to 12% of the total variance while the persons-by-subtests interaction effects accounted for 1% to 5% of the total variance.

Table 33

Two G Studies for the Form A and Form B Pretests with the p X (i:s) Design

Source	Form A		Form B	
Persons	.00999	4%	.01014	4%
Subtests	.00064	0%	.00000	0%
items:s	.02411	10%	.02928	12%
ps	.01308	5%	.00241	1%
pi:s	.19557	80%	.20571	83%
Total	.24339	100%	.24754	100%

Table 34

Two G Studies for the Form A and Form B Posttests with the p X (i:s) Design

Source	Form A		Form B	
persons	.01370	6%	.02584	11%
Subtests	.00029	0%	.00000	0%
items:s	.02552	11%	.02202	9%
ps	.00627	3%	.00292	1%
pi:s	.17755	80%	.18941	79%
Total	.22333	100%	.24019	100%

What contributed to the total variance most was the persons-by-items -nested-within-subtests interaction effects, which accounted for 79% to 83% of the total variance.

Multivariate Generalizability Studies

Tables 35, 36, 37, and 38 display the variance and covariance components of the four G multivariate studies with the *p˙ X i°* design for the Form A pretest, Form B pretest, Form A posttest, and Form B posttest. The largest variance components were due to the persons-by-items interaction effects for all the subtests. The persons and items effects accounted for almost the same percentages of the total variance, ranging from 4% to 17%. The results of these multivariate G studies were more similar to the ones obtained from the univariate G studies with the *(p:c) X (i:s)* or *p X (i:s)* design in which the greatest variance was due to the interaction effect. The upper diagonals in the persons variance shows attenuated correlations. The lowest correlation was between the listening and vocabulary subtests. The reading and listening, and reading and vocabulary subtests generally correlated highly, but the vocabulary and listening subtests did not correlate as highly.

Table 35

Variance and Covariance Components for the Form A Pretest Taken by Group A with the G Study Design of p˙ X i°

	Vocabulary		Listening		Reading	
	.02264	9%	.29623		.43579	
∑persons =	.00547		.01507	6%	.56214	
	.01131		.01190		.02974	12%
	.01987	8%				
∑items =			.02269	9%		
					.02958	12%
	.19822	82%				
∑pi =			.21261	85%		
					.17973	75%
Total	.24073	100%	.25037	100%	.23905	100%

Table 36

Variance and Covariance Components for the Form B Pretest Taken by Group B with the G Study Design of p˙ X i°

	Vocabulary		Listening		Reading	
	.01360	6%	.67328		.90158	
∑persons =	.00747		.00905	4%	.81940	
	.01261		.00935		.01438	6%
	.03317	14%				
∑items =			.04240	17%		
					.01501	6%
	.19858	81%				
∑pi =			.19845	79%		
					.21858	88%
Total	.24535	100%	.24990	100%	.24797	100%

Table 37

Variance and Covariance Components for the Form A Posttest Taken by Group B with the G Study Design of p˙ X i°

	Vocabulary		Listening		Reading	
∑persons =	.02645	11%	.56774		.83372	
	.00930		.01015	5%	.66628	
	.01886		.00934		.01935	9%
∑items =	.02565	11%				
			.03027	14%		
					.02437	11%
∑pi =	.18492	78%				
			.17126	81%		
					.17639	80%
Total	.23702	100%	.21168	100%	.22011	100%

Table 38

Variance and Covariance Components for the Form B Posttest Taken by Group A with the G Study Design of p˙ X i°

	Vocabulary		Listening		Reading	
∑persons =	.02663	11%	.92632		.93953	
	.01902		.01582	7%	.97198	
	.03111		.02481		.04117	17%
∑items =	.02453	10%				
			.03038	13%		
					.01290	5%
∑pi =	.18843	79%				
			.19625	81%		
					.18498	77%
Total	.23959	100%	.24245	100%	.23905	100%

Tables 39 and 40 display the variance and covariance components of the two multivariate G studies with the *p˙ X (i°:s˙)* design for the Form A pretest and Form B posttest taken by Group A, and Form B pretest and Form A posttest taken by Group B. Only 3% to 4% of the total variance on the pretests was due to persons. However, on the posttests, the persons variance components accounted for 11% and 6% of the total variance. The three subtests were treated as a facet. No variance was accounted for

due to the subtests; thus, there were no differences in the difficulty of the subtests. The items variance components accounted for 9% to 12% of the total variance, indicating that there was variability in item difficulty. The persons-by-subtests interaction effects were fairly small, ranging from 1% to 6%. Most of the variance was due to the persons-by-items nested within subtests interaction, which accounted

Table 39

Variance and Covariance Components for the Form A Pretest and Form B Posttest Taken by Group A with the G Study Design of p˙ X (i°:s˙)

	Form A pretest		Form B posttest	
∑persons =	.00792	3%	.92866	
	.01338		.02622	11%
∑subtests =	.00036	0%		
	-.00024		-.00098	0%
∑items:s =	.02628	11%		
			.02150	9%
∑p:s=	.01362	6%		
	.00107		.00330	1%
∑pi:s =	.19679	80%		
			.18980	79%
Total	.24497	100%	.24082	100%

Table 40

Variance and Covariance Components for the Form B pretest and Form A Posttest Taken by Group B with the G Study Design of p˙ X (i°:s˙)

	Form B pretest		Form A posttest	
∑persons =	.01016	4%	.79345	
	.00921		.01327	6%
∑subtests =	.00000	0%		
	.00069		.00021	0%
∑items:s =	.02894	12%		
			.02515	11%
∑p:s=	.00247	1%		
	.00227		.00611	3%
∑pi:s =	.20629	84%		
			.17624	80%
Total	.24786	100%	.22098	100%

for about 80% of the total variance. The correlation coefficients between the Form A pretest and Form B posttest, and between the Form B pretest and Form A posttest were high at .93 and .80, indicating that the two forms were plausibly measuring the same domain.

FACETS Analyses

In order to examine the extent to which the examinees' ability, item difficulty, and subtest difficulty differ, four FACETS analyses on the Form A pretest, Form B pretest, Form A posttest, and Form B posttest were conducted. Figures 3 and 4 display the FACETS maps for the Form A and Form B pretests with three facets: persons, items, and subtests. These maps provide a summary of the results. The person ability estimates ranged from -1.64 to 1.13 logits on the Form A pretest, and from -1.70 to 1.13 on the Form B pretest. The person separation index, which indicates the separation of person ability estimates on a logit scale, for the Form A pretest and for the Form B pretest were 1.92 and 1.90, respectively. The Rasch person reliability estimates were .79 and .78. The χ^2 for the fixed effect was significant at .00.

Tables 41 and 42 show that the item facility values ranged from .10 and .79 on the Form A pretest and from .02 and .87 on the Form B pretest. The item difficulty estimates ranged from -1.52 to 2.04 on the Form A pretest and from -2.22 to 3.72 on the Form B pretest. Thus, the Form B pretest contained a wider range of items in terms of difficulty. The item separation index of 3.58 and 3.78, and the Rasch item reliability estimates of .93 and .93 indicated that the items were separated widely along the logit scale in relation to their standard errors. No misfitting items outside the range of .70 - 1.30 were identified. Item facility values and item difficulty

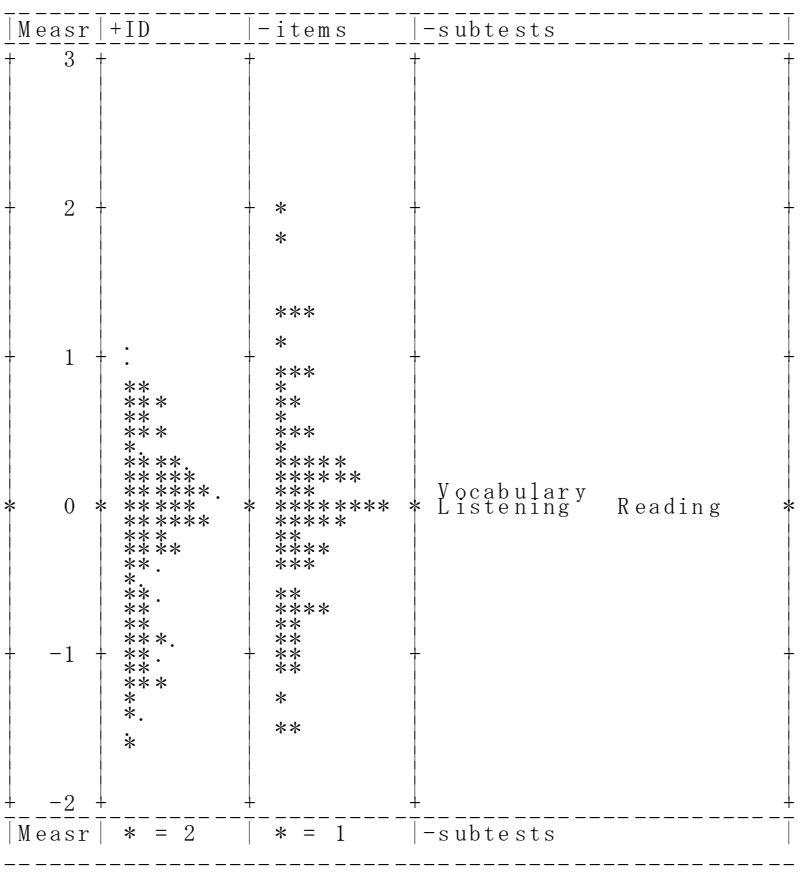

Figure 3. FACETS map for the Form A pretest taken by Group A with three facets: persons, items, and subtests.

estimates correlate highly but the difference is that the latter are converted to an interval scale and less affected by examinees' ability.

Tables 43 and 44 show that the Form A vocabulary subtest was the most difficult with a subtest difficulty estimate of .09. The subtest separation index was 1.18 and the Rasch subtest reliability estimate was .58. The χ^2 at the significance level of .02 also indicated that the subtests were significantly different in terms of difficulty. On the Form B pretest, based on the subtest separation index of .00, the Rasch subtest

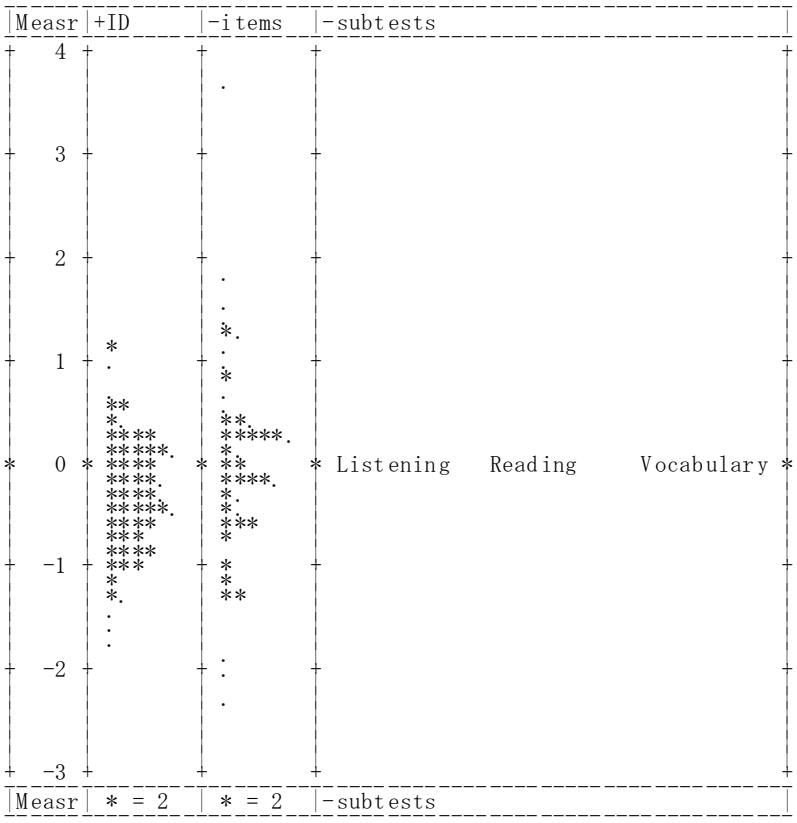

Figure 4. FACETS map for the Form B pretest taken by Group B with three facets: persons, items, and subtests.

reliability of .00, and χ^2 at the significance level of .24, no significant difference in the difficulty was observed.

Figures 5 and 6 show the FACETS maps for the Form A and Form B posttests with three facets: persons, items, and subtests. The person ability estimates ranged from -.76 to 2.52 for the Form A posttest taken by Group B. The person separation index of 2.35 and the Rasch person reliability of .85 indicated significant separation of the person ability estimates. Group A took the Form B posttest. The person ability

Table 41

Rasch Item Statistics for the Form A Pretest Taken by Group A

Item	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ
1	.46	-.06	.18	1.10	1.10
2	.44	.03	.18	1.00	1.00
3	.35	.42	.18	1.00	1.00
4	.53	-.31	.17	.90	.90
5	.43	.09	.18	1.00	.90
6	.63	-.73	.18	1.00	.90
7	.41	.15	.18	1.00	1.00
8	.79	-1.52	.20	.90	.90
9	.38	.27	.18	1.10	1.10
10	.25	.89	.20	1.00	1.10
11	.67	-.89	.18	1.00	1.00
12	.51	-.25	.18	1.00	1.00
13	.42	.10	.18	1.10	1.10
14	.44	.02	.18	1.00	1.00
15	.38	.28	.19	.90	.90
16	.11	1.85	.28	1.10	1.50
17	.30	.66	.20	1.00	1.10
18	.59	-.58	.18	.90	.90
19	.52	-.30	.18	1.10	1.10
20	.51	-.25	.18	.90	.90
21	.18	1.30	.24	1.10	1.40
22	.18	1.31	.24	1.10	1.30
23	.54	-.36	.19	.90	.90
24	.63	-.72	.19	.80	.80
25	.39	.25	.20	1.00	1.00
26	.45	.00	.18	1.00	1.00
27	.61	-.64	.18	1.00	1.00
28	.47	-.11	.18	1.10	1.10
29	.55	-.43	.18	1.00	1.00
30	.10	2.04	.27	1.10	1.60
31	.52	-.30	.18	1.10	1.10
32	.44	.03	.18	1.10	1.10
33	.28	.75	.19	1.20	1.30
34	.33	.51	.18	1.10	1.10
35	.44	.04	.18	1.10	1.10
36	.69	-1.02	.19	.90	.90
37	.70	-1.07	.19	.90	.90
38	.40	.22	.18	1.00	1.00
39	.32	.54	.18	1.00	1.10
40	.41	.17	.18	1.10	1.10
41	.63	-.73	.18	1.00	1.00
42	.21	1.14	.20	1.10	1.20
43	.68	-.97	.19	1.00	1.00
44	.65	-.83	.18	.90	.90
45	.45	.00	.18	1.10	1.10
46	.33	.52	.19	1.00	1.00
47	.65	-.84	.19	1.00	1.00
48	.78	-1.49	.22	.90	.80
49	.72	-1.14	.20	.80	.70
50	.76	-1.35	.21	.90	.90

(Table 41 continues)

(Table 41 continued)

Item	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ
51	.25	.90	.20	1.10	1.20
52	.48	-.12	.18	1.00	1.00
53	.30	.63	.19	1.00	1.00
54	.38	.29	.19	.90	.80
55	.56	-.44	.19	.90	.90
56	.44	.03	.19	.90	.80
57	.66	-.87	.20	.80	.80
58	.63	-.73	.20	.90	.90
59	.44	.03	.19	.90	.90
60	.47	-.07	.19	1.10	1.10
61	.47	-.08	.20	.90	.90
62	.52	-.30	.20	.90	.90
63	.39	.24	.21	1.00	.90
64	.40	.21	.21	1.00	1.00
65	.25	.89	.24	1.00	1.00
66	.42	.11	.22	1.10	1.10
67	.40	.22	.23	1.10	1.10
68	.26	.84	.25	1.10	1.30
69	.19	1.27	.28	1.10	1.20
70	.39	.25	.23	.90	.90

Note. Item numbers in bold appear on Form B.

Table 42

Rasch Item Statistics for the Form B Pretest Taken by Group B

Item	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ
1	.45	-.08	.21	1.00	1.00
2	.58	-.61	.21	.90	.90
3	.13	1.63	.29	1.00	.90
4	.45	-.11	.21	1.10	1.10
5	.35	.33	.21	1.00	1.00
6	.19	1.17	.26	1.00	1.00
7	.58	-.61	.21	1.00	1.00
8	.36	.28	.21	1.10	1.10
9	.87	-2.22	.29	.90	.80
10	.50	-.28	.20	1.00	1.00
11	.26	.76	.23	1.00	1.10
12	.35	.30	.22	.90	.90
13	.73	-1.29	.22	.90	.90
14	.53	-.40	.20	.90	.90
15	.55	-.49	.21	.90	.90
16	.37	.24	.21	1.10	1.10
17	.61	-.74	.21	1.10	1.10
18	.02	3.72	.72	1.00	1.70
19	.44	-.04	.21	1.00	1.00
20	.33	.39	.22	1.10	1.10
21	.25	.80	.23	1.10	1.10
22	.21	1.03	.25	1.20	1.40
23	.52	-.37	.21	.90	.90
24	.70	-1.13	.22	.90	.90

(Table 42 continues)

(Table 42 continued)

Item	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ
25	.47	-.16	.21	.90	.90
26	.18	1.25	.26	1.00	1.00
27	.36	.26	.21	1.00	1.00
28	.41	.08	.21	1.00	1.00
29	.72	-1.23	.22	1.00	1.00
30	.45	-.09	.21	1.10	1.10
31	.85	-2.00	.27	1.00	1.10
32	.66	-.94	.21	1.10	1.10
33	.15	1.45	.27	1.10	1.30
34	.33	.39	.21	1.10	1.10
35	.10	1.87	.31	1.00	1.00
36	.72	-1.23	.22	1.00	.90
37	.82	-1.79	.26	1.00	1.00
38	.56	-.55	.21	.90	.90
39	.31	.49	.22	1.10	1.00
40	.38	.21	.21	1.10	1.10
41	.17	1.31	.26	1.10	1.10
42	.42	.04	.21	1.10	1.10
43	.39	.17	.21	1.00	1.00
44	.55	-.51	.21	.90	.90
45	.38	.21	.21	.90	.90
46	.70	-1.13	.22	.90	.80
47	.57	-.59	.21	.90	.80
48	.59	-.68	.21	.90	.90
49	.18	1.23	.26	1.10	1.40
50	.28	.63	.22	1.20	1.40
51	.32	.44	.22	1.00	1.10
52	.47	-.17	.20	.90	.90
53	.46	-.13	.21	1.00	1.00
54	.35	.30	.21	1.00	.90
55	.49	-.25	.20	1.00	.90
56	.66	-.94	.21	1.00	.90
57	.32	.44	.22	.90	.90
58	.39	.17	.21	1.00	1.00
59	.43	.00	.21	1.00	1.00
60	.72	-1.23	.22	1.00	1.00
61	.37	.25	.21	.90	.90
62	.56	-.52	.21	1.00	.90
63	.50	-.29	.21	1.00	1.00
64	.46	-.12	.21	1.10	1.10
65	.47	-.17	.21	1.00	1.00
66	.35	.32	.22	1.10	1.20
67	.43	.01	.21	1.10	1.10
68	.37	.22	.22	1.10	1.10
69	.23	.89	.24	1.00	1.00
70	.47	-.19	.21	1.00	1.00

Note. Item number in bold appear on Form A.

Table 43

Rasch Subtest Statistics for the Form A Pretest Taken by Group A

	Vocabulary	Listening	Reading
Subtest difficulty	.09	-.04	-.05
SE	.04	.04	.04

Table 44

Rasch Subtest Statistics for the Form B Pretest Taken by Group B

	Vocabulary	Listening	Reading
Subtest difficulty	.04	-.02	-.02
SE	.04	.05	.05

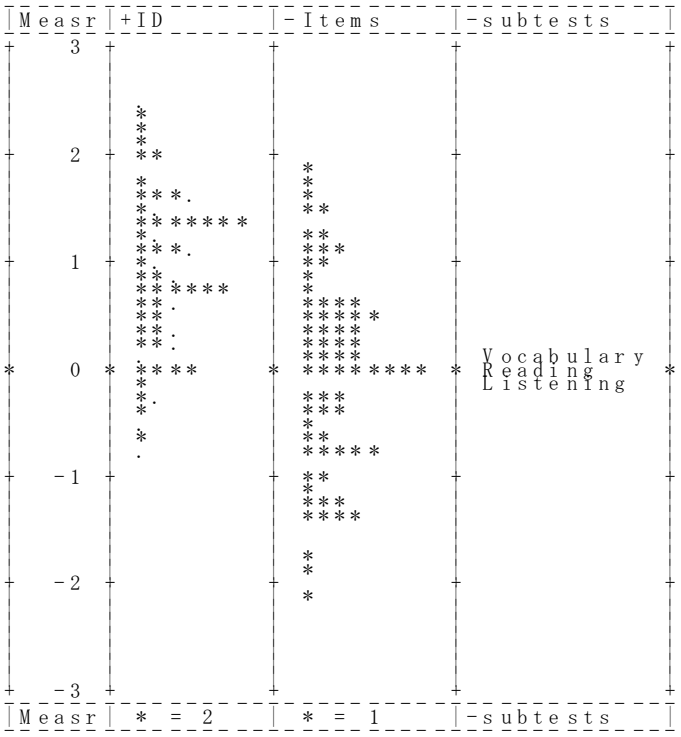

Figure 5. FACETS map for the Form A posttest taken by Group B with three facets: persons, items, and subtests.

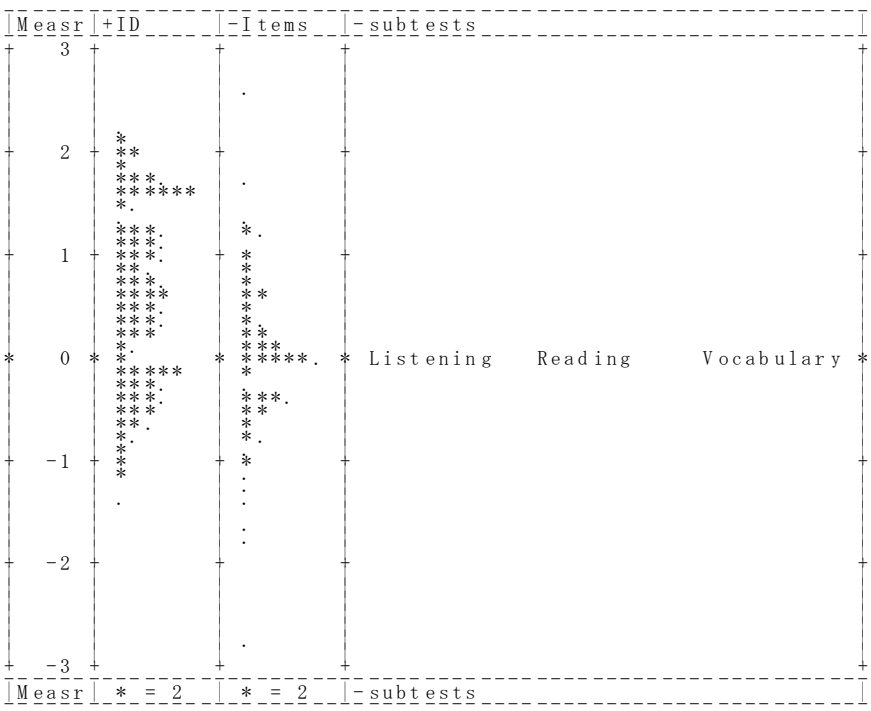

Figure 6. FACETS map for the Form B posttest taken by Group A with three facets: persons, items, and subtests.

estimates ranged from -1.36 to 2.29. The person separation index of 2.91 and the Rasch person reliability of .89 indicated that the person ability estimates were separated to a considerable degree.

Tables 45 and 46 show that the item facility values ranged from .26 to .95 on the Form A posttest and from .11 to .96 on the Form B posttest. The item difficulty estimates ranged from -2.08 to 1.92 on Form A and from -2.75 to 2.69 on Form B. The item separation indices of 2.35 and 2.91, and the Rasch item reliability coefficients of .85 and .89 for the two forms showed significant separation of the item difficulty estimates. Item 41 on Form B misfit the model (Infit MNSQ = 1.40; Outfit MNSQ = 1.50).

Table 45

Rasch Item Statistics for the Form A Posttest Taken by Group B

Item	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ
1	.45	1.11	.21	1.20	1.30
2	.84	-.79	.25	1.00	.90
3	.79	-.44	.23	1.00	.90
4	.84	-.79	.25	.90	.80
5	.83	-.67	.25	1.00	.90
6	.78	-.38	.23	0.90	.90
7	.68	.16	.21	1.00	1.00
8	.90	-1.31	.30	.80	.60
9	.64	.34	.21	1.00	1.00
10	.36	1.47	.22	1.00	1.00
11	.87	-1.00	.27	.90	.70
12	.71	.02	.22	1.00	1.00
13	.46	1.07	.21	1.10	1.20
14	.42	1.20	.21	1.20	1.30
15	.76	-.28	.23	.90	.80
16	.67	.20	.21	1.00	.90
17	.45	1.11	.21	1.10	1.10
18	.71	.02	.22	1.00	1.00
19	.56	.64	.21	1.00	1.00
20	.61	.42	.21	.90	.80
21	.64	.34	.21	.90	.90
22	.32	1.66	.22	.90	.90
23	.70	.07	.22	.80	.80
24	.87	-1.00	.27	.90	.70
25	.69	.11	.21	.90	.80
26	.76	-.28	.24	1.00	1.00
27	.94	-1.91	.40	1.00	.90
28	.72	-.04	.23	1.20	1.20
29	.89	-1.19	.31	1.00	1.00
30	.71	-.01	.23	1.10	1.10
31	.75	-.22	.23	.90	.90
32	.26	1.92	.22	1.20	1.40
33	.56	.64	.21	1.20	1.20
34	.66	.23	.22	1.10	1.10
35	.60	.50	.21	1.00	1.00
36	.91	-1.39	.33	.90	.80
37	.70	.04	.22	1.00	1.10
38	.71	-.01	.23	1.00	.90
39	.59	.55	.21	.90	.90
40	.47	1.02	.21	1.20	1.20
41	.89	-1.19	.31	1.00	1.00
42	.41	1.27	.21	1.20	1.20
43	.93	-1.76	.38	1.00	1.00
44	.95	-2.08	.43	1.00	.70
45	.60	.50	.21	1.20	1.20
46	.35	1.50	.21	1.20	1.30
47	.79	-.43	.24	.90	.90
48	.91	-1.43	.33	1.00	.90
49	.81	-.56	.25	.80	.70
50	.90	-1.33	.32	1.00	.80
51	.51	.87	.21	.90	.80

(Table 45 continues)

(Table 45 continued)

Item	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ
52	.84	-.75	.26	1.00	1.00
53	.64	.33	.21	1.00	.90
54	.66	.22	.22	.80	.80
55	.84	-.75	.26	.90	.90
56	.72	-.05	.23	1.00	.90
57	.84	-.75	.26	.90	.80
58	.88	-1.14	.30	.90	.70
59	.65	.28	.21	1.00	.90
60	.56	.64	.21	1.00	1.00
61	.89	-1.23	.31	.90	.70
62	.82	-.62	.25	.90	.80
63	.56	.64	.21	1.20	1.20
64	.68	.14	.22	.90	1.00
65	.30	1.73	.22	1.20	1.30
66	.47	1.02	.21	1.10	1.20
67	.70	.05	.22	.80	.80
68	.54	.74	.21	.90	.90
69	.58	.55	.21	1.00	1.10
70	.61	.46	.21	1.20	1.20

Note. Item number in bold appear on Form B.

Table 46

Rasch Item Statistics for the Form B Posttest Taken by Group A

Item	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ
1	.34	1.25	.19	1.30	1.50
2	.81	-.89	.21	1.00	1.10
3	.38	1.04	.18	1.00	1.00
4	.90	-1.64	.26	1.00	.80
5	.40	.97	.18	1.10	1.10
6	.52	.50	.18	.90	.80
7	.54	.40	.18	1.10	1.10
8	.58	.24	.18	1.10	1.10
9	.91	-1.73	.27	.90	.50
10	.74	-.47	.20	.90	.80
11	.75	-.51	.20	.90	.80
12	.43	.87	.18	1.10	1.10
13	.71	-.32	.19	1.20	1.20
14	.65	-.07	.19	1.00	1.00
15	.64	-.02	.19	1.10	1.10
16	.73	-.43	.20	.90	.90
17	.82	-.98	.22	.90	.80
18	.57	.27	.18	.90	.80
19	.74	-.50	.20	1.10	1.20
20	.42	.90	.18	1.20	1.30
21	.64	.00	.19	.90	.90
22	.24	1.69	.20	1.00	1.00
23	.63	.05	.19	1.00	1.00
24	.82	-.98	.22	.90	.90
25	.71	-.32	.19	.80	.80

(Table 46 continues)

(Table 46 continued)

Item	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ
26	.30	1.42	.19	1.00	1.10
27	.49	.60	.18	1.00	1.00
28	.65	-.04	.19	1.00	1.00
29	.72	-.37	.19	1.00	1.10
30	.52	.50	.18	1.20	1.20
31	.96	-2.75	.39	.90	.70
32	.73	-.44	.20	1.00	1.00
33	.11	2.69	.26	1.20	2.00
34	.77	-.64	.20	1.00	1.00
35	.61	.10	.18	1.30	1.40
36	.88	-1.38	.24	.90	0.70
37	.79	-.76	.21	1.00	1.10
38	.64	.00	.19	.90	.90
39	.58	.23	.18	.90	.90
40	.46	.73	.18	1.00	1.00
41	.55	.37	.18	1.40	1.50
42	.66	-.08	.19	1.10	1.00
43	.72	-.40	.19	1.20	1.30
44	.62	.07	.19	1.10	1.10
45	.48	.63	.18	.90	.90
46	.86	-1.24	.23	.90	.80
47	.71	-.35	.19	.90	.80
48	.63	.01	.19	1.00	1.00
49	.33	1.26	.19	1.00	1.00
50	.60	.15	.18	1.00	1.00
51	.34	1.25	.19	1.10	1.10
52	.76	-.61	.20	.80	.70
53	.61	.10	.19	.80	.80
54	.61	.14	.19	.80	.80
55	.64	.00	.19	.90	.80
56	.79	-.77	.21	.90	1.00
57	.70	-.29	.19	.90	.80
58	.57	.29	.19	1.00	1.00
59	.72	-.35	.20	.90	.80
60	.79	-.77	.21	1.00	.90
61	.49	.62	.18	.90	.90
62	.63	.05	.19	.80	.80
63	.64	.01	.19	.90	.90
64	.65	-.06	.19	1.00	1.00
65	.61	.12	.19	.90	.80
66	.55	.35	.19	1.20	1.20
67	.84	-1.13	.23	.90	.80
68	.63	.02	.19	.90	.80
69	.46	.74	.19	1.00	1.00
70	.48	.66	.19	1.20	1.20

Note. Item number in bold appear in Form A.

Tables 47 and 48 show the subtest measure results for the Form A posttest taken by Group B and the Form B posttest taken by Group A. The subtest separation index of 1.60 and the Rasch subtest reliability estimates of .72 showed that the three

subtests were separated to a moderate degree in terms of their overall difficulty. The vocabulary subtest contained more difficult items than those of the other subtests. On the Form B posttest taken by Group A, the subtest separation index of .00 and the Rasch subtest reliability of .00 revealed that there was almost no separation for the subtest difficulty estimates, which ranged from -.03 to .02. This showed that the subtests were almost equal in difficulty.

Table 47

Rasch Subtest Statistics for the Form A Posttest Taken by Group B

	Vocabulary	Listening	Reading
Subtest difficulty	.13	-.08	-.04
SE	.04	.05	.05

Table 48

Rasch Subtest Statistics for the Form B Posttest Taken by Group A

	Vocabulary	Listening	Reading
Subtest difficulty	.01	.02	-.03
SE	.04	.04	.04

Research Question 3: Dependability

Research Question 3 asked about the extent to which the two CRT forms are dependable when administered as pretests and posttests. First, the norm-referenced reliability coefficients were estimated with Cronbach alpha and K-R20. Then, the dependability indices of the two CRT forms were checked using the phi lamda dependability formula, and confidence intervals. Finally, univariate decision studies with the *p X (I:S)* and *p X (I:S)* and multivariate decision studies with the $p^{\cdot} X I^{\circ}$, and $p^{\cdot} X (I^{\circ}: S^{\cdot})$ designs were conducted to estimate the dependability.

Dependability of the Criterion-Referenced Forms

As shown in Tables 49 and 50, the Cronbach internal consistency reliability and K-R20 coefficients were almost equivalent, but the latter were slightly higher, ranging from .48 to .81. The K-R20 coefficients for the Form A pretest were higher due to the large standard deviations. Because norm-referenced reliability formulas such as K-R20 are not appropriate for interpreting criterion-referenced reliability (Brown, 1990), Brown's short-cut dependability formula (Φ) was used. The dependability indices, which ranged from .43 to .78, were lower than the norm-referenced reliability coefficients. Phi lamda dependability (Φ(λ)) indicates decision consistency. When the cut-point was set at 70%, the indices were higher. Generally, the indices were high, ranging from .43 to .84.

In Tables 51 and 52, the Cronbach alpha and K-R20 coefficients, which ranged from .56 to .84, were generally higher on the posttests than on the pretests. Once again the dependability indices, which ranged from .52 to .84, were slightly lower than the norm-referenced reliability coefficients. Because the listening subtests had a smaller number of items, the dependability indices were moderate, while the indices for the vocabulary and reading subtests were fairly high, as there were 25 items on each of these subtests.

The Phi lamda dependability (Φ(λ)) indices were moderate to high, ranging from .43 to .94. The index was sensitive to the location of the cut-point and the dependability index. When a cut-point was set apart from the mean and the dependability index was high, the Phi lamda dependability index tended to be high. The largest value for the *CI* was .1097, indicating that the students' test scores would hypothetically be in the range of plus or minus 11 points 68% of the time if they took

the test again. The *CIs* were smaller when the dependability indices were greater.

Overall, it was possible to make dependable decisions based on these results.

Table 49

Reliability and Dependability of the Form A Pretest

	Vocabulary	Listening	Reading	Total
Alpha	.74	.60	.80	.83
K-R20	.75	.61	.81	.84
Φ	.73	.59	.78	.83
Φ(.7)	.93	.86	.94	.97
Φ(.6)	.88	.74	.90	.94
Φ(.5)	.79	.59	.84	.87
CI	.0930	.1080	.0912	.0570

Table 50

Reliability and Dependability of the Form B Pretest

	Vocabulary	Listening	Reading	Total
Alpha	.63	.48	.62	.79
K-R20	.63	.48	.62	.79
Φ	.59	.43	.61	.76
Φ(.7)	.91	.85	.89	.96
Φ(.6)	.83	.71	.80	.92
Φ(.5)	.66	.43	.64	.81
CI	.0963	.1097	.0967	.0581

Table 51

Reliability and Dependability of the Form A Posttest

	Vocabulary	Listening	Reading	Total
Alpha	.79	.56	.74	.86
K-R20	.79	.56	.74	.86
Φ	.76	.52	.72	.84
Φ(.7)	.79	.48	.71	.85
Φ(.6)	.76	.66	.76	.87
Φ(.5)	.83	.83	.86	.93
CI	.0914	.1000	.0894	.0544

Table 52

Reliability and Dependability of the Form B Posttest

	Vocabulary	Listening	Reading	Total
Alpha	.78	.62	.85	.91
K-R20	.78	.62	.85	.92
Φ	.76	.58	.84	.90
Φ(.7)	.80	.69	.86	.92
Φ(.6)	.75	.56	.84	.90
Φ(.5)	.81	.67	.87	.92
CI	.0923	.1064	.0890	.0551

Univariate Decision Studies

Table 53 shows the results of the D studies on the pretest Form A and Form B vocabulary, listening, and reading subtests with the *p X (I:S)* design. Figure 7 shows a graphic representation of the D studies on the pretest Form A and Form B vocabulary, listening, and reading subtests with the *p X (I:S)* design. As can be seen, the reading subtest on the Form A pretest had the highest dependability, and the listening subtest on the Form B pretest had the lowest dependability. The test forms had five sections with five items per section on the vocabulary and reading subtests. The dependability indices were .44 and .61 on the Form A vocabulary and reading subtests, and .33 and .38 on the Form B vocabulary and reading subtests. The listening subtest had four sections with five items per section. The dependability indices were .27 on Form A and .61 on Form B. When the test forms had 40 items, divided into eight sections with five items per section and 10 sections with four items per section, the dependability indices were both at the same degree. Because no sections variance components were observed, adding sections had no impact on the dependability. However, adding more items increased the dependability indices. The only exception was the Form B listening subtest, which did not change because the index was too low to be affected by adding more sections and items. The dependability indices of the listening subtests were generally lower than those of the vocabulary and reading subtests.

Table 53

D Studies on the Pretest Form A and Form B Vocabulary, Listening, and Reading Subtests with the p X (I:S) Design

				Form A			Form B		
D study	Sect.	Item	Total	Vocabulary	Listening	Reading	Vocabulary	Listening	Reading
1	1	3	3	.09	.04	.16	.06	.00	.07
2	2	3	6	.16	.08	.27	.11	.00	.13
3	3	3	9	.22	.12	.36	.15	.00	.18
4	4	3	12	.27	.15	.42	.19	.01	.23
5	5	3	15	.32	.18	.48	.23	.01	.27
6	6	3	18	.36	.21	.52	.26	.01	.31
7	7	3	21	.40	.24	.56	.29	.01	.34
8	8	3	24	.43	.26	.60	.32	.01	.37
9	9	3	27	.46	.28	.62	.35	.01	.40
10	10	3	30	.48	.31	.65	.37	.01	.43
11	1	4	4	.11	.06	.20	.07	.00	.09
12	2	4	8	.20	.10	.33	.14	.00	.17
13	3	4	12	.27	.15	.42	.19	.01	.23
14	4	4	16	.33	.19	.50	.24	.01	.28
15	5	4	20	.38	.23	.55	.28	.01	.33
16	6	4	24	.43	.26	.60	.32	.01	.37
17	7	4	28	.47	.29	.63	.36	.01	.41
18	8	4	32	.50	.32	.66	.39	.01	.44
19	9	4	36	.53	.34	.69	.42	.02	.47
20	10	4	40	.55	.37	.71	.44	.02	.50
21	1	5	5	.13	.07	.23	.09	.00	.11
22	2	5	10	.24	.13	.38	.17	.00	.20
23	3	5	15	.32	.18	.48	.23	.01	.27
24	4	5	20	.38	.22	.55	.28	.01	.33
25	5	5	25	.44	.27	.61	.33	.01	.38
26	6	5	30	.48	.30	.65	.37	.01	.43
27	7	5	35	.52	.34	.68	.41	.01	.46
28	8	5	40	.55	.37	.71	.44	.02	.50
29	9	5	45	.58	.39	.73	.47	.02	.53

Note. Sect. = number of sections; Item = number of items; Vocabulary = generalizability coefficient for absolute decisions of the vocabulary subtest; Listening = generalizability coefficient for absolute decisions of the listening subtest; Reading = generalizability coefficient for absolute decisions of the reading subtest.

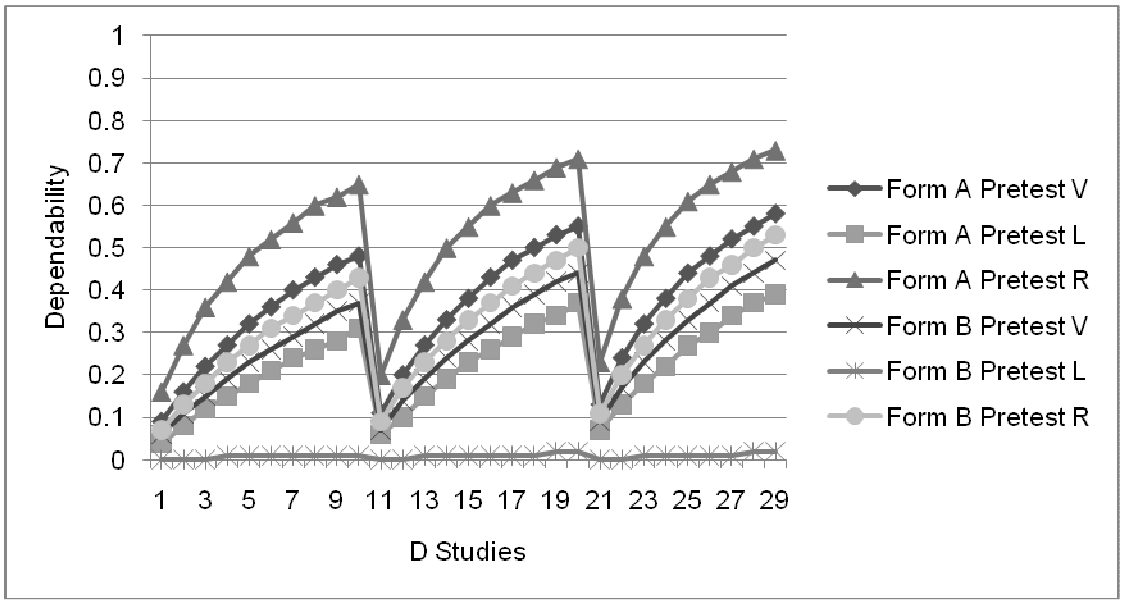

Figure 7. Graphic representation of the D studies on the pretest Form A and Form B vocabulary, listening, and reading subtests with the *p X (I:S)* design.

Table 54 shows the results of the D studies on the posttest vocabulary, listening, and reading subtests on Form A and Form B with the *p X (I:S)* design. Figure 8 shows a graphic representation of the D studies on the posttest Form A and Form B vocabulary, listening, and reading subtests with the *p X (I:S)* design. As can be seen, the reading subtest on the Form B posttest had the highest dependability and the listening subtest on the Form B posttest had the lowest dependability. When there were five sections, containing five items per section on the vocabulary and reading subtests the dependability indices ranged from .34 to .52. Compared to the indices estimated by Brown's short-cut formula (Tables 51 and 52), the indices estimated with this D study design were lower and unsatisfactory. As shown in Table 54, the indices of the listening subtest, containing four sections and five items per section were extremely low at .26 and .18 on Form A and Form B, respectively. Because no sections variance was observed in the G study, adding more sections did not have an impact on the dependability but adding more items did. However, adding 10

Table 54

D Studies on the Posttest Form A and Form B Vocabulary, Listening, and Reading Subtests with the p X (I:S) Design

				Form A			Form B		
D study	Sect.	Item	Total	Vocabulary	Listening	Reading	Vocabulary	Listening	Reading
1	1	3	3	.06	.05	.06	.07	.03	.11
2	2	3	6	.11	.10	.12	.14	.06	.20
3	3	3	9	.16	.14	.17	.19	.09	.28
4	4	3	12	.20	.18	.21	.24	.11	.34
5	5	3	15	.24	.21	.25	.28	.14	.39
6	6	3	18	.27	.24	.28	.32	.16	.44
7	7	3	21	.31	.27	.32	.35	.18	.47
8	8	3	24	.34	.30	.35	.38	.20	.51
9	9	3	27	.36	.32	.37	.41	.22	.54
10	10	3	30	.39	.35	.40	.44	.24	.56
11	1	4	4	.08	.07	.08	.09	.04	.15
12	2	4	8	.14	.12	.15	.17	.08	.26
13	3	4	12	.20	.18	.21	.24	.11	.34
14	4	4	16	.25	.22	.26	.29	.15	.41
15	5	4	20	.30	.26	.31	.34	.18	.46
16	6	4	24	.34	.30	.35	.38	.20	.51
17	7	4	28	.37	.33	.38	.42	.23	.55
18	8	4	32	.40	.36	.41	.45	.25	.58
19	9	4	36	.43	.39	.44	.48	.28	.61
20	10	4	40	.46	.41	.47	.51	.30	.63
21	1	5	5	.10	.08	.10	.12	.05	.18
22	2	5	10	.17	.15	.18	.21	.10	.30
23	3	5	15	.24	.21	.25	.28	.14	.39
24	4	5	20	.30	.26	.31	.34	.18	.46
25	5	5	25	.34	.31	.36	.39	.21	.52
26	6	5	30	.39	.35	.40	.44	.24	.56
27	7	5	35	.42	.38	.44	.48	.27	.60
28	8	5	40	.46	.41	.47	.51	.30	.63
29	9	5	45	.49	.44	.50	.54	.32	.66

Note. Sect. = number of sections; Item = number of items; Vocabulary = generalizability coefficient for absolute decisions of the vocabulary subtest; Listening = generalizability coefficient for absolute decisions of the listening subtest; Reading = generalizability coefficient for absolute decisions of the reading subtest.

adding 10 more items on the subtests would only change the dependability by .08. As Figure 8 shows, the dependability of the Form B reading subtest on the posttest had the highest dependability index and adding more items would increase the index.

Because the dependability index of the Form B listening subtest on the posttest was extremely low, adding items did not have an impact on the index.

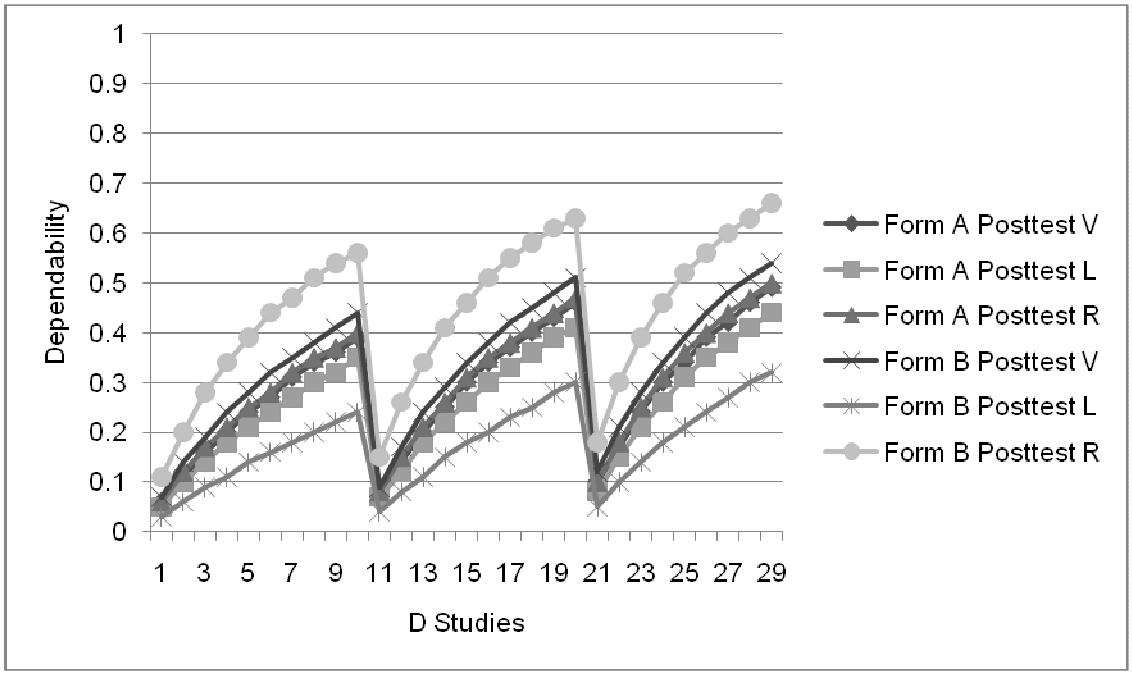

Figure 8. Graphic representation of the D studies on the posttest Form A and Form B vocabulary, listening, and reading subtests with the *p X (I:S)* design.

Table 55 shows the results of four D studies on the Form A and Form B pretests and posttests with the *p X (I:S)* design. Figure 9 shows a graphic representation of the results of the four D studies on the Form A and Form B pretests and posttests with the *p X (I:S)* design. As can be seen, the Form B posttest had the highest dependability and the Form A pretest had the lowest dependability. The indices for the 70-item form, containing 25-20-25 items per section, were .56, .71, .73, and .87 for the Form A pretest, Form B pretest, Form A posttest, and Form B posttest, respectively. Compared to the indices estimated with Brown's short-cut formula (Tables 49, 50, 51, and 52), these indices with this D study design underestimated the dependability. For example,

the dependability for the Form A pretest was .83 in Table 49, but as shown in Table 55 the dependability with this design was .56.

Adding more items only had a slight impact on the indices. For example, when the form contained 70 items and 90 items, the indices were .56 and .59 on the Form A pretest. Because this D study design was unbalanced in terms of the different number of items on each subtest, adding or reducing five items on the subtests did not change the dependability of the form. For instance, in D studies 5, 6, and 7, different numbers of items were set, but because the total number of items remained at 70, the indices were the same degree at.56 on the Form A pretest. Figure 9 shows the graphic representation of the D studies. Fifteen items were added to the Form each time D studies 1 to 4 were conducted. Up to D studies 1 to 3, the curves increased steeply, but they increased more gradually after that.

Table 55

Four D Studies on the Form A and Form B Pretests and Posttests with the p X (I:S) Design

D study	Voc	Lis	Read	Total	Form A Pretest ($E\rho^2(\Delta)$)	Form B Pretest ($E\rho^2(\Delta)$)	Form A Posttest ($E\rho^2(\Delta)$)	Form B Posttest ($E\rho^2(\Delta)$)
1	5	5	5	15	.34202	.38113	.46565	.63162
2	10	10	10	30	.45644	.54008	.60475	.76310
3	15	15	15	45	.51374	.62729	.67163	.82000
4	20	20	20	60	.54814	.68238	.71094	.85176
5	20	25	25	70	.56285	.70866	.72836	.86584
6	25	20	25	70	.56285	.70866	.72836	.86584
7	25	25	20	70	.56285	.70866	.72836	.86584
8	25	25	25	75	.57109	.72034	.73681	.87202
9	30	30	30	90	.58748	.74808	.75513	.88607

Note. Voc = number of items on the vocabulary subtest; Lis = number of items on the listening subtest; Read = number of items on the reading subtest.

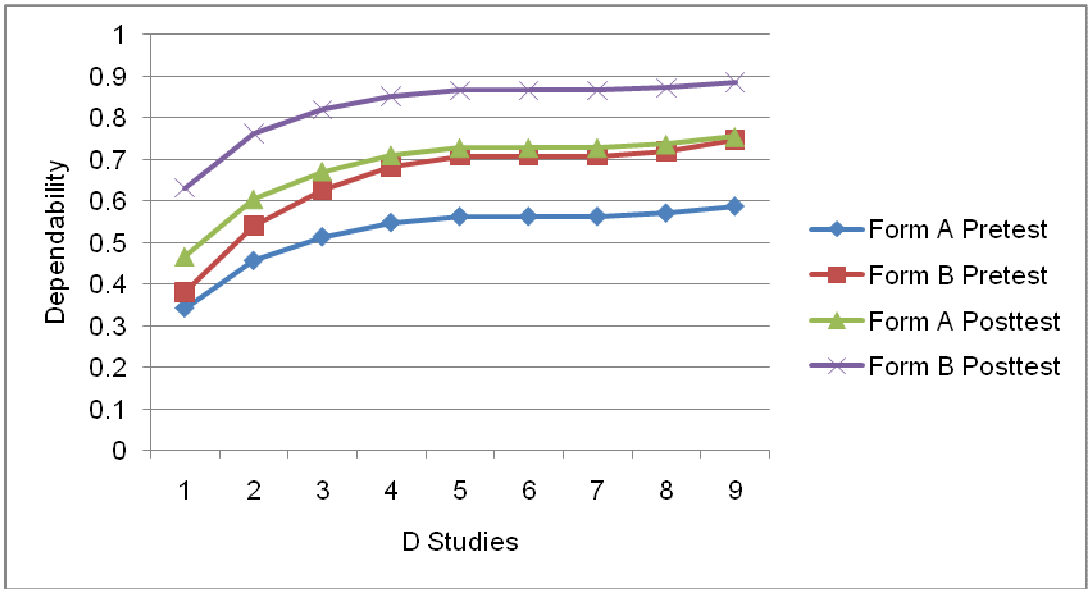

Figure 9. Graphic representation of the D studies on the Form A pretest, Form B pretest, Form A posttest, and Form B posttest with the *p X (I:S)* design.

Multivariate Decision Studies

Tables 56, 57, 58, and 59 show the dependability indices ($E\rho^2(\Delta)$) and composite dependability indices ($\hat{E}\rho^2(\Delta)$) for the four D studies for the Form A and Form B pretests and posttests with the $p^{\cdot} X I^{\circ}$ design. Figure 10 shows a graphic representation of D studies on the Form A pretest, Form B pretest, Form A posttest, and Form B posttest with the $p^{\cdot} X I^{\circ}$ design. The graph shows that the Form B posttest and Form B pretest had the highest and lowest composite dependability. The dependability indices estimated using Brown's short-cut formula were almost equivalent (see Tables 49, 50, 51, and 52). For instance, the dependability index for the Form A vocabulary pretest was .73 (Table 49) and .72 (Table 56) when the subtest had 25 items. The univariate D study for the Form A vocabulary pretest shows that if five more items measuring the same domain were added to the Form A vocabulary pretest, the dependability index would increase by only .04 (Table 56). Adding ten items would only result in an increase of .06. Generally, adding five more items to the

vocabulary, and reading subtests resulted in only slight increases in the univariate generalizability coefficients. Especially, for the listening subtest, adding more items increased the index by only .05 to .08. However, adding more items to each of the subtests would not lead to significant increases in the multivariate generalizability coefficient for absolute decisions. For instance, adding ten items to each of the subtests on the 70-item Form A pretest would only cause a .05 change in the index. Overall, the multivariate generalizability coefficients for absolute decisions for all the forms were fairly high, ranging from .77 to .90 when there were 25 vocabulary, 20 listening, and 25 reading items.

Table 56

Dependability Indices and Composite Dependability Indices for the Form A Pretest Taken by Group A with the D Study Design of p˙ X I°

D study	Vocabulary	$E\rho^2(\Delta)$	Listening	$E\rho^2(\Delta)$	Reading	$E\rho^2(\Delta)$	Total	$E\rho^2(\Delta)$
1	15	.61	10	.39	15	.68	40	.73
2	20	.68	15	.49	20	.74	55	.78
3	25	.72	20	.56	15	.68	60	.80
4	25	.72	10	.39	25	.78	60	.81
5	15	.61	20	.56	25	.78	60	.78
6	25	.72	20	.56	20	.74	65	.81
7	25	.72	15	.49	25	.78	65	.82
8	20	.68	20	.56	25	.78	65	.80
9	25	.72	20	.56	25	.78	70	.82
10	30	.76	20	.56	25	.78	75	.83
11	25	.72	25	.62	25	.78	75	.82
12	25	.72	20	.56	30	.81	75	.84
13	35	.78	20	.56	25	.78	80	.84
14	25	.72	30	.66	25	.78	80	.83
15	25	.72	20	.56	35	.83	80	.85
16	30	.76	25	.62	30	.81	85	.85
17	35	.78	30	.66	35	.83	100	.87

Note. Vocabulary = number of items on the vocabulary subtest; Listening = number of items on the listening subtest; Reading = number of items on the reading subtest.

Table 57

Dependability Indices and Composite Dependability Indices for the Form B Pretest Taken by Group B with the D Study Design of p˙ X I°

D study	Vocabulary	$E\rho^2(\Delta)$	Listening	$E\rho^2(\Delta)$	Reading	$E\rho^2(\Delta)$	Total	$E\rho^2(\Delta)$
1	15	.46	10	.27	15	.48	40	.66
2	20	.54	15	.36	20	.55	55	.72
3	25	.59	20	.43	15	.48	60	.73
4	25	.59	10	.27	25	.61	60	.75
5	15	.46	20	.43	25	.61	60	.73
6	25	.59	20	.43	20	.55	65	.75
7	25	.59	15	.36	25	.61	65	.76
8	20	.54	20	.43	25	.61	65	.75
9	25	.59	20	.43	25	.61	70	.77
10	30	.64	20	.43	25	.61	75	.78
11	25	.59	25	.48	25	.61	75	.77
12	25	.59	20	.43	30	.65	75	.78
13	35	.67	20	.43	25	.61	80	.79
14	25	.59	30	.53	25	.61	80	.78
15	25	.59	20	.43	35	.68	80	.79
16	30	.64	25	.48	30	.65	85	.80
17	35	.67	30	.53	35	.68	100	.82

Note. Vocabulary = number of items on the vocabulary subtest; Listening = number of items on the listening subtest; Reading = number of items on the reading subtest.

Table 60 displays the dependability indices ($E\rho^2(\Delta)$) and composite dependability indices ($\hat{E}\rho^2(\Delta)$) in the D study for the Form A pretest and Form B posttest with the *p˙ X (I°:S˙)* design. The univariate generalizability coefficients for absolute decisions for the Form A pretest and for the Form B posttest were .50 and .86, respectively when there were 25-20-25 items per form. Recall that the multivariate generalizability coefficients for absolute decisions of the Form A pretest was .82 when the vocabulary, listening, and reading subtests were treated as separate measures. However, treating the three subtests as one test and a single facet reduced the magnitude of the dependability. Adding five items to each of the subtests would not drastically change the univariate generalizability coefficients for absolute decisions for the Form A pretest. Likewise, adding ten items to each of the subtests would only

Table 58

Dependability Indices and Composite Dependability Indices for the Form A Posttest Taken by Group B with the D Study Design of p˙ X I°

D study	Vocabulary	$E\rho^2(\Delta)$	Listening	$E\rho^2(\Delta)$	Reading	$E\rho^2(\Delta)$	Total	$E\rho^2(\Delta)$
1	15	.65	10	.34	15	.59	40	.76
2	20	.72	15	.43	20	.66	55	.81
3	25	.76	20	.50	15	.59	60	.81
4	25	.76	10	.34	25	.71	60	.84
5	15	.65	20	.50	25	.71	60	.81
6	25	.76	20	.50	20	.66	65	.83
7	25	.76	15	.43	25	.71	65	.84
8	20	.72	20	.50	25	.71	65	.82
9	25	.76	20	.50	25	.71	70	.84
10	30	.79	20	.50	25	.71	75	.85
11	25	.76	25	.56	25	.71	75	.84
12	25	.76	20	.50	30	.74	75	.85
13	35	.81	20	.50	25	.71	80	.86
14	25	.76	30	.60	25	.71	80	.85
15	25	.76	20	.50	35	.77	80	.86
16	30	.79	25	.56	30	.74	85	.86
17	35	.81	30	.60	35	.77	100	.88

Note. Vocabulary = number of items on the vocabulary subtest; Listening = number of items on the listening subtest; Reading = number of items on the reading subtest.

result in an increase of .03 in the univariate generalizability coefficients for absolute decisions. Adding 45 more items to the two forms would result in an increase of .03 in the univariate generalizability coefficients for absolute decisions.

Table 61 shows the dependability indices ($E\rho^2(\Delta)$) and composite dependability indices ($\hat{E}\rho^2(\Delta)$) in the D studies for the Form B pretest and Form A posttest with the *p˙ X (I°:S˙)* design. The results of the multivariate D study were similar to those shown in Table 60. That is, adding more items would not change the dependability appreciably. The multivariate generalizability coefficient for absolute decisions of .79 (Table 61) was lower when there were 25-20-25 items on Form A and Form B than the coefficient of .83 shown in Table 60, but it was, nonetheless, fairly high.

Table 59

Dependability Indices and Composite Dependability Indices for the Form B Posttest Taken by Group A with the D Study Design of p˙ X I°

D study	Vocabulary	$E\rho^2(\Delta)$	Listening	$E\rho^2(\Delta)$	Reading	$E\rho^2(\Delta)$	Total	$E\rho^2(\Delta)$
1	15	.65	10	.41	15	.76	40	.84
2	20	.71	15	.51	20	.81	55	.88
3	25	.76	20	.58	15	.76	60	.87
4	25	.76	10	.41	25	.84	60	.89
5	15	.65	20	.58	25	.84	60	.89
6	25	.76	20	.58	20	.81	65	.89
7	25	.76	15	.51	25	.84	65	.90
8	20	.71	20	.58	25	.84	65	.89
9	25	.76	20	.58	25	.84	70	.90
10	30	.79	20	.58	25	.84	75	.91
11	25	.76	25	.64	25	.84	75	.90
12	25	.76	20	.58	30	.86	75	.91
13	35	.81	20	.58	25	.84	80	.91
14	25	.76	30	.68	25	.84	80	.90
15	25	.76	20	.58	35	.88	80	.92
16	30	.79	25	.64	30	.86	85	.91
17	35	.81	30	.68	35	.88	100	.93

Note. Vocabulary = number of items on the vocabulary subtest; Listening = number of items on the listening subtest; Reading = number of items on the reading subtest.

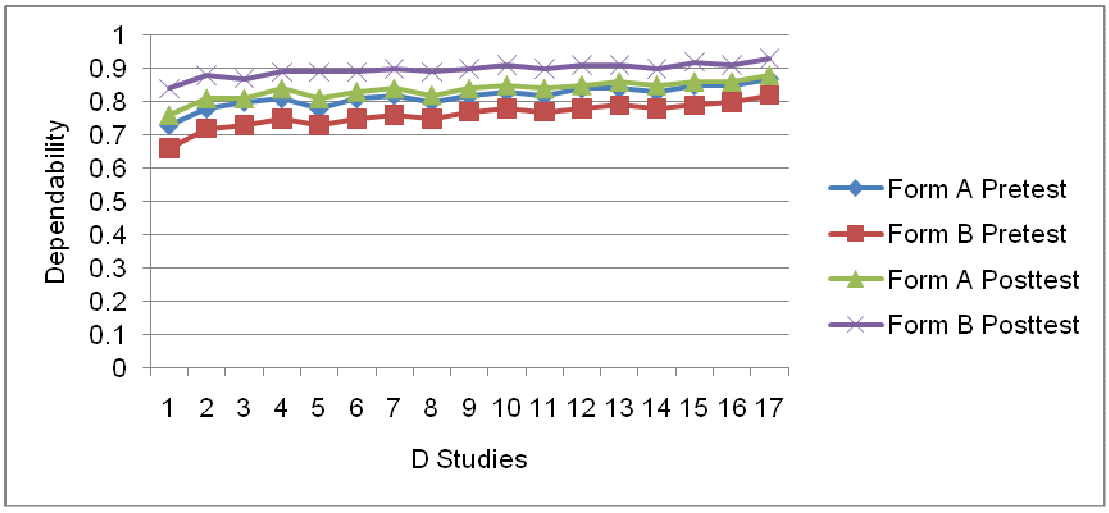

Figure 10. Graphic representation of D studies on the Form A pretest, Form B pretest, Form A posttest, and Form B posttest with the *p˙ X I°* design.

Table 60

Dependability Indices and Composite Dependability Indices for the Form A Pretest and Form B Posttest Taken by Group A with the D Study Design of p˙ X (I°:S˙)

	Form A Pretest					Form B Posttest						
D study	Voc	List	Read	Total	$E\rho^2(\Delta)$	Voc	List	Read	Total	$E\rho^2(\Delta)$	Total	$E\rho^2(\Delta)$
1	25	20	25	70	.50	20	15	20	55	.84	125	.80
2	25	20	25	70	.50	20	20	25	65	.86	135	.82
3	25	20	25	70	.50	25	15	25	65	.86	135	.82
4	25	20	25	70	.50	25	20	20	65	.86	135	.82
5	25	20	25	70	.50	30	15	20	65	.86	135	.82
6	25	20	25	70	.50	20	25	20	65	.85	135	.82
7	25	20	25	70	.50	20	15	30	65	.86	135	.82
8	20	15	20	55	.48	25	20	25	70	.86	125	.84
9	20	20	25	65	.49	25	20	25	70	.86	135	.83
10	25	15	25	65	.49	25	20	25	70	.86	135	.83
11	25	20	20	65	.49	25	20	25	70	.86	135	.83
12	30	15	20	65	.48	25	20	25	70	.86	135	.83
13	20	25	20	65	.49	25	20	25	70	.86	135	.83
14	20	15	30	65	.48	25	20	25	70	.86	135	.83
15	25	20	25	70	.50	25	20	25	70	.86	140	.83
16	30	20	25	75	.51	25	20	25	70	.86	145	.82
17	25	25	25	75	.51	25	20	25	70	.86	145	.83
18	25	20	30	75	.51	25	20	25	70	.86	145	.82
19	20	25	30	75	.51	25	20	25	70	.86	145	.82
20	30	15	30	75	.50	25	20	25	70	.86	145	.82
21	30	25	20	75	.51	25	20	25	70	.86	145	.82
22	30	25	30	85	.52	25	20	25	70	.86	155	.82
23	25	20	25	70	.50	30	20	25	75	.87	145	.84
24	25	20	25	70	.50	25	25	25	75	.87	145	.84
25	25	20	25	70	.50	25	20	30	75	.87	145	.84
26	25	20	25	70	.50	20	25	30	75	.87	145	.84
27	25	20	25	70	.50	30	15	30	75	.87	145	.84
28	25	20	25	70	.50	30	25	20	75	.87	145	.84
29	25	20	25	70	.50	30	25	30	85	.88	155	.85
30	30	25	30	85	.52	35	30	35	100	.89	185	.86

Note. Voc = number of items on the vocabulary subtest; List = number of items on the listening subtest; Read = number of items on the reading subtest.

Table 61

Dependability Indices and Composite Dependability Indices for the Form B Pretest and Form A Posttest Taken by Group B with the D Study Design of p˙ X (I°:S˙)

	Form B Pretest					Form A Posttest						
D study	Voc	List	Read	Total	$E\rho^2(\Delta)$	Voc	List	Read	Total	$E\rho^2(\Delta)$	Total	$E\rho^2(\Delta)$
1	25	20	25	70	.71	20	15	20	55	.70	125	.78
2	20	15	20	55	.67	25	20	25	70	.73	125	.78
3	25	20	25	70	.71	20	20	25	65	.72	135	.79
4	25	20	25	70	.71	25	15	25	65	.71	135	.78
5	25	20	25	70	.71	25	20	20	65	.72	135	.79
6	25	20	25	70	.71	30	15	20	65	.71	135	.78
7	25	20	25	70	.71	20	25	20	65	.72	135	.79
8	25	20	25	70	.71	20	15	30	65	.71	135	.78
9	20	20	25	65	.70	25	20	25	70	.73	135	.79
10	25	15	25	65	.69	25	20	25	70	.73	135	.78
11	25	20	20	65	.70	25	20	25	70	.73	135	.79
12	30	15	20	65	.69	25	20	25	70	.73	135	.78
13	20	25	20	65	.70	25	20	25	70	.73	135	.79
14	20	15	30	65	.69	25	20	25	70	.73	135	.78
15	25	20	25	70	.71	25	20	25	70	.73	140	.79
16	30	20	25	75	.72	25	20	25	70	.73	145	.79
17	25	25	25	75	.72	25	20	25	70	.73	145	.79
18	25	20	30	75	.72	25	20	25	70	.73	145	.79
19	20	25	30	75	.72	25	20	25	70	.73	145	.79
20	30	15	30	75	.72	25	20	25	70	.73	145	.79
21	30	25	20	75	.72	25	20	25	70	.73	145	.79
22	25	20	25	70	.71	30	20	25	75	.73	145	.79
23	25	20	25	70	.71	25	25	25	75	.73	145	.79
24	25	20	25	70	.71	25	20	30	75	.73	145	.79
25	25	20	25	70	.71	20	25	30	75	.73	145	.79
26	25	20	25	70	.71	30	15	30	75	.73	145	.79
27	25	20	25	70	.71	30	25	20	75	.73	145	.79
28	30	25	30	85	.74	25	20	25	70	.73	155	.80
29	25	20	25	70	.71	30	25	30	85	.75	155	.80
30	30	25	30	85	.74	35	30	35	100	.76	185	.81

Note. Voc = number of items on the vocabulary subtest; List = number of items on the listening subtest; Read = number of items on the reading subtest.

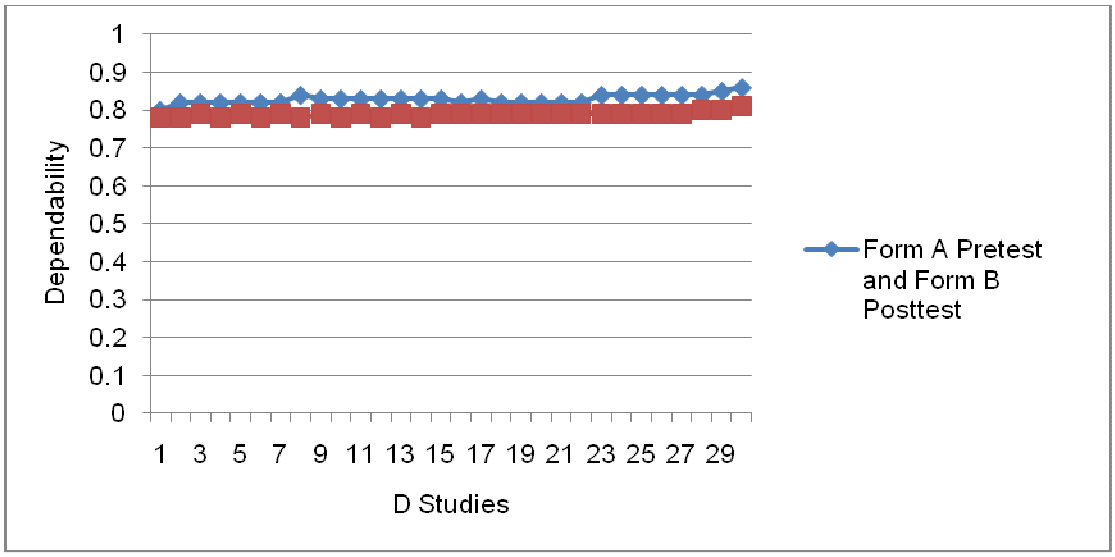

Figure 11. Graphic representation of the D studies on the Form A pretest and Form B posttest, and the Form B pretest and Form A posttest with the $p^{\cdot} X (I^{\circ}:S^{\cdot})$ design.

Research Question 4: Validity

Research Question 4 asked the extent to which the two CRT forms are valid when administered as pretests and posttests. To answer the question, two repeated-measures MANOVAs were conducted to investigate the test score gain in Group A and Group B. Prior to the analysis, the assumptions were checked. In addition, two ANOVA-like FACETS analyses with the person, item, class, occasion, and subtest facets were conducted to estimate the item difficulty, class ability, occasion difficulty, and subtest difficulty.

Repeated-Measures Multivariate Analysis of Variance

Prior to conducting the repeated-measures MANOVA, the data were screened to meet the required assumptions. Only 138 and 99 students in Groups A and B who took both the pretest and posttest were used in this analysis. First, univariate outliers were checked by examining standardized scores. No outliers were found in Group A.

However, one student in Group B was found to be an outlier, so that person was eliminated. Second, Mahalanobis distance was examined to identify multivariate outliers. A Chi square with 5 degrees of freedom at $p < .001$ was used to identify multivariate outliers. No multivariate outliers were found in Groups A and B. Third, the descriptive statistics were calculated without the outliers in order to check normality. The distributions for the listening and reading subtests on the Form A posttest were skewed (Table 63), and the distributions for the reading subtest and total score on the Form B posttest displayed negative kurtosis (Table 62). Although the non-normality violates the assumption, no transformation of the data was undertaken because it would change the means, and the analysis was considered to be robust to modest violations of normality as long as the violations were not due to outliers (Tabachnick & Fidell, 2001, p. 329). Fourth, the scatterplots were examined to check the assumption of linearity, and then correlation coefficients among the vocabulary, listening, and reading subtests were examined to check for multicollinearity. The correlation coefficients ranged from .14 and .77, showing the absence of multicollinearity.

Table 64 displays the results of the repeated-measures MANOVA for Group A. The analysis was used to test the effects of occasions (pretest and posttest) and subtests (vocabulary, listening, and reading subtests) on the Group A students' CRT scores. Notice that the occasion effect was significant, indicating that the mean differences on the pretest and posttest did not happen by chance alone. The subtest effect was also significant, indicating that the subtest means were significantly different. The interaction effect was also significant. In Figure 12, the three non-parallel lines showed that there was an interaction effect. The eta square of .64

Table 62

Descriptive Statistics for the Group A Students' Pretest and Posttest Scores

Pre/post	Form	Subtest	*N*	Min	Max	*M*	*SD*	Skew	*SES*	Kurt	*SEK*
Pretest	Form A	Vocabulary	138	0	20	10.13	4.39	.09	.21	-.64	.41
		Listening	138	0	17	9.57	3.03	-.11	.21	-.16	.41
		Reading	138	0	20	9.96	4.62	-.23	.21	-.33	.41
		Total	138	6	48	29.65	8.64	-.34	.21	-.20	.41
Posttest	Form B	Vocabulary	138	3	24	15.09	4.62	-.31	.21	-.72	.41
		Listening	138	3	19	11.88	3.22	-.24	.21	-.10	.41
		Reading	138	3	24	15.04	5.51	-.27	.21	-.93	.41
		Total	138	16	62	42.01	12.057	-.183	.206	-1.099	.410

Note. Statistics are based on raw scores.

Table 63

Descriptive Statistics for the Group B Students' Pretest and Posttest Scores

Pre/post	Form	Subtest	*N*	Min	Max	*M*	*SD*	Skew	*SES*	Kurt	*SEK*
Pretest	Form B	Vocabulary	98	3	22	10.60	3.733	.111	.244	-.184	.483
		Listening	98	2	16	9.11	2.788	-.077	.244	-.144	.483
		Reading	98	2	21	11.37	3.767	.083	.244	-.229	.483
		Total	98	13	51	31.08	8.288	-.048	.244	-.481	.483
Posttest	Form A	Vocabulary	98	5	24	15.87	4.531	-.389	.244	-.898	.483
		Listening	98	6	19	14.24	2.593	-.511	.244	.549	.483
		Reading	98	7	24	17.16	4.076	-.576	.244	-.406	.483
		Total	98	26	63	47.28	9.067	-.384	.244	-.700	.483

Note. Statistics are based on raw scores.

for the occasion effect indicated that 64% of the variability in the test score could be explained by that independent variable. The eta square for the subtest effect was .18. The observed power values of 1.00 in all the effects indicated that a Type II error was unlikely.

Table 64

Repeated-Measures MANOVA Between-Subjects Effects for Group A

Effect	*F*	*df*	*p*	η^2	Power
Occasions	246.46	1	.00	.64	1.00
Subtests	30.05	2	.00	.18	1.00
Occasions * Subtests	18.91	2	.00	.12	1.00

Note. Multivariate *F* ratios were generated using Greenhouse-Geisser statistic.

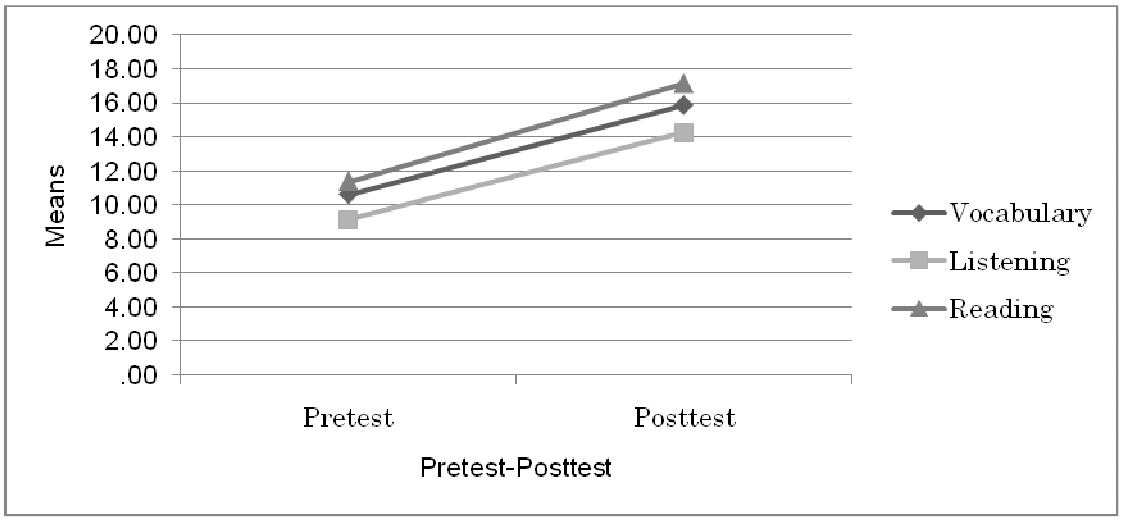

Figure 12. Means for the subtests on the pretest and posttest taken by Group A.

Table 65 shows the results of the repeated-measures MANOVA between-subjects effects for Group B. The pretest-posttest mean difference was significant, indicating that the Group B students' test scores improved within this interval. The eta square of .84 showed that this effect explains 84% of the total variance. The observed power of 1.00 showed that the analysis had enough power to detect significant effects. The subtest effect was significant, indicating that there were significant differences among the subtest means. The eta squared of .28 showed that the effect explained a considerable amount of the total variance. No interaction effect was observed.

Table 65

Repeated-Measures MANOVA Between-Subjects Effects for Group B

Effect	*F*	*df*	*p*	η^2	Power
Occasions	495.27	1	.00	.84	1.00
Subtests	37.74	2	.00	.28	1.00
Occasions * Subtests	1.02	2	.32	.01	.23

Note. Multivariate *F* ratios were generated from Greenhouse-Geisser.

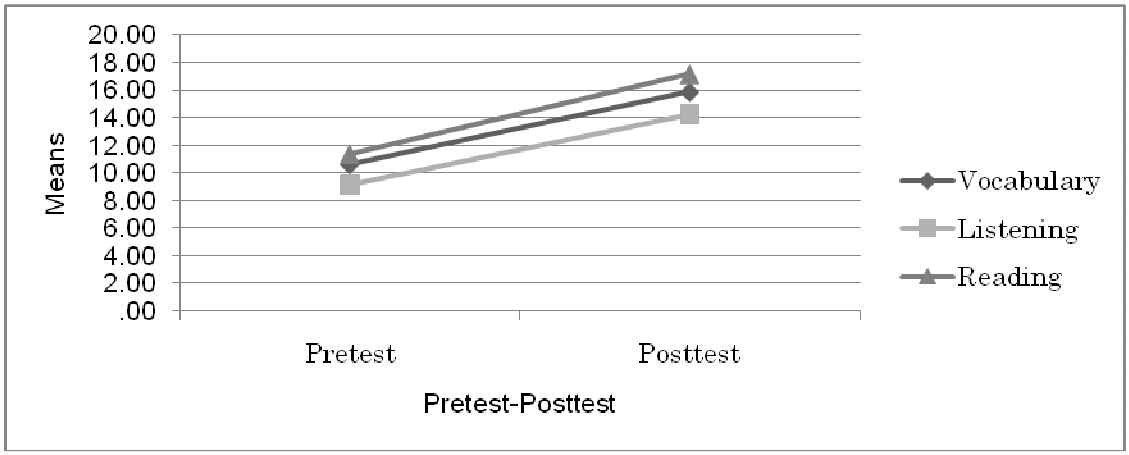

Figure 13. Means for the subtests on the pretest and posttest taken by Group B.

ANOVA-Like FACETS Analyses

Figures 14 and 15 display the FACETS maps for Group A, who took the Form A pretest and the Form B posttest, and for Group B, who took the Form B pretest and the Form A posttest. A FACETS control file used for this FACETS analysis for Group A is shown in Appendix J. Because there were more students in Group A and the degree to which they studied for the posttest differed greatly, the person ability estimates ranged from -5.62 to 2.22. The range of the person ability estimates ranged from -1.28 to 2.13 for Group B. The person ability estimates in Group B (Rasch person separation index = 2.27; Rasch person reliability = .84) were not as separated as the estimates in Group A (Rasch person separation index = 2.91; Rasch person

reliability = .89) but were still widely separated. The means of the person ability estimates for Group A on the pretest and posttest were -.27 and .46, respectively. That is, Group A students were able to improve their estimates by .73. The means of the person ability estimates for Group B on the pretest and posttest were .11 and .54, respectively. Thus, they were able to improve their estimates by .43. Horizontal lines in the maps show arbitrary cut-points that can be used to make pass/fail decisions set. They are set at 60% or -.08 in Figure 14 and .32 in Figure 15.

Tables 66 and 67 display the item measures for Group A who took the Form A pretest and Form B posttest and Group B who took the Form B pretest and Form A posttest. The Rasch item separation indices of 4.22 and 3.30 and the Rasch item reliability estimates of .95 and .94 indicate that the item difficulty estimates vary widely. Item difficulty estimates in bold denote that the items appear in both test forms and they were used as common items to equate the two test forms. That is why the estimates are equivalent in the pretest and posttest. For instance, the estimates for item 16 on the Form A pretest and item 11 on the Form B posttest are equivalent at .54. Based on the infit mean square values, only one item, item 41 on the Form B posttest, misfits the Rasch model. The outfit mean square values showed that six items taken by Group A and one item taken by Group B misfit the Rasch model. The outfit values are more sensitive to outliers than the infit values.

When making pass/fail decisions based on this test, having a number of items with item difficulty estimates around a cut-point would result in more dependable decisions. The item difficulty estimate for item 33 on the Form B posttest was extremely high at 2.82. It was the most difficult item and therefore did not help make dependable pass/fail decisions. Quest [version 2.1] was used to investigate how the

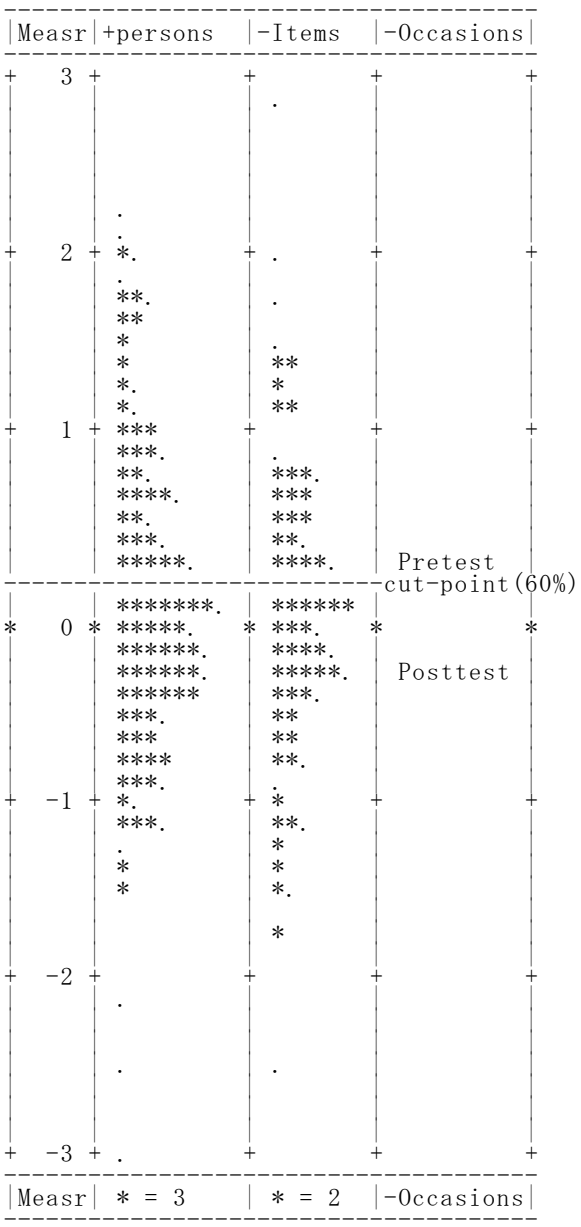

Figure 14. FACETS map for the Form A pretest and the Form B posttest taken by Group A with three facets: persons, items, and occasions.

item was functioning with the Group A students. The Rasch distractor analysis showed that the item discrimination values for choices a, b, c, and d were .24, -.23, -.07, and .04, respectively. The mean ability estimates of the persons who chose a, b, c, and d were .59, .11, .27, and .68, respectively. The correct choice was choice c.

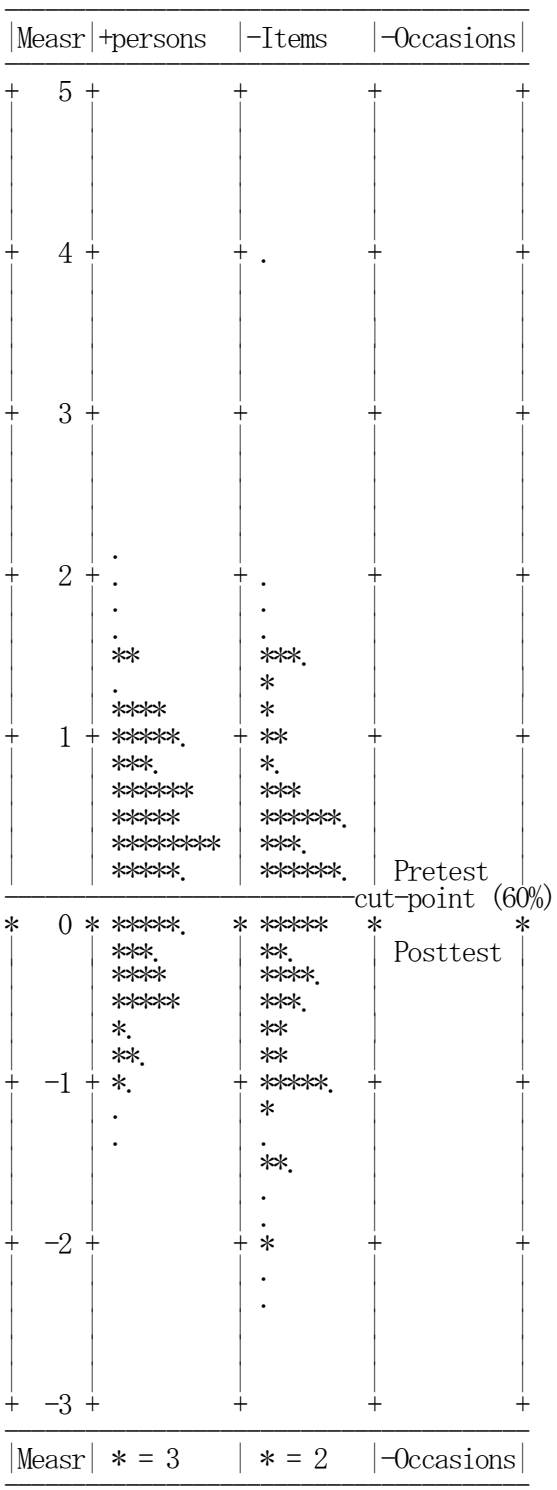

Figure 15. FACETS map for the Form B pretest and the Form A posttest taken by Group B with three facets: persons, items, and occasions.

Table 66

Rasch Item Statistics for the Form A Pretest and Form B Posttest Taken by Group A

	Form A Pretest					Form B Posttest				
Item	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ
1	.60	-.30	.18	1.10	1.10	.33	1.39	.19	1.30	1.50
2	.56	-.20	.18	1.00	1.00	.78	-.69	.21	1.00	1.10
3	.48	.16	.19	1.00	1.00	.35	1.25	.19	1.00	1.00
4	.66	-.58	.18	1.00	1.00	.88	-1.39	.26	1.00	.80
5	.58	-.20	.18	1.00	1.00	.38	1.11	.19	1.10	1.10
6	.74	-.93	.18	.90	.90	.48	.73	.18	.90	.80
7	.54	-.10	.18	1.00	1.00	.50	.59	.18	1.10	1.10
8	.88	-1.78	.21	.90	.80	.54	.46	.19	1.10	1.10
9	.48	.13	.18	1.10	1.20	.89	-1.53	.27	.90	.60
10	.36	.68	.20	1.00	1.20	.71	-.31	.20	.90	.70
11	.76	-1.07	.18	1.00	1.00	.70	**.54**	**.13**	**.80**	**.90**
12	.62	-.42	.18	1.00	1.10	.39	**.80**	**.14**	**1.10**	**1.20**
13	.52	-.01	.18	1.10	1.10	.67	**-.43**	**.13**	**1.10**	**1.20**
14	.54	-.10	.18	1.00	.90	.63	**-.11**	**.13**	**1.00**	**1.00**
15	.49	.13	.18	1.00	1.00	.59	**.01**	**.13**	**1.00**	**1.10**
16	.16	**.54**	**.13**	**.80**	**.90**	.72	-.27	.20	.90	.90
17	.39	**.80**	**.14**	**1.10**	**1.20**	.79	-.78	.22	.90	.90
18	.67	**-.43**	**.13**	**1.10**	**1.20**	.54	.49	.19	.90	.80
19	.59	**-.11**	**.13**	**1.00**	**1.00**	.71	-.27	.20	1.10	1.20
20	.56	**.01**	**.13**	**1.00**	**1.10**	.38	1.07	.19	1.20	1.30
21	.24	1.26	.24	1.10	1.20	.60	.18	.19	.90	.90
22	.23	1.31	.24	1.00	1.20	.21	1.94	.21	1.00	1.00
23	.59	-.26	.18	.90	.90	.58	.25	.19	1.00	1.00
24	.67	-.61	.18	.80	.80	.79	-.69	.21	.90	.90
25	.44	.34	.19	1.00	1.00	.67	-.11	.19	.80	.80
26	.56	-.42	.18	1.00	1.00	.33	1.54	.19	1.00	1.00
27	.68	-.97	.18	1.00	1.00	.53	.73	.18	1.00	.90
28	.58	-.45	.18	1.10	1.10	.67	.14	.19	1.00	1.00
29	.66	-.80	.18	1.10	1.10	.72	-.15	.20	1.00	1.10
30	.13	1.79	.29	1.10	1.40	.55	.63	.18	1.20	1.20
31	.61	-.64	.18	1.10	1.10	.96	-2.55	.39	.90	.70
32	.54	-.33	.18	1.00	1.10	.75	-.31	.20	1.00	1.00
33	.36	.41	.19	1.10	1.20	.13	2.82	.26	1.20	2.00
34	.42	.16	.19	1.10	1.10	.77	-.47	.21	1.00	1.00
35	.52	-.26	.18	1.10	1.10	.64	.28	.19	1.30	1.40
36	.77	**-1.29**	**.15**	**.90**	**.80**	.88	**-1.29**	**.15**	**.90**	**.80**
37	.78	**-1.14**	**.15**	**1.00**	**1.10**	.81	**-1.14**	**.15**	**1.00**	**1.10**
38	.49	**.05**	**.13**	**1.00**	**1.00**	.63	**.05**	**.13**	**1.00**	**1.00**
39	.43	**.28**	**.13**	**1.00**	**1.00**	.61	**.28**	**.13**	**1.00**	**1.00**
40	.50	**.35**	**.13**	**1.20**	**1.30**	.48	**.35**	**.13**	**1.20**	**1.30**
41	.73	-1.17	.19	1.00	1.00	.57	.59	.18	1.40	1.50
42	.28	.81	.21	1.10	1.30	.67	.11	.19	1.10	1.00
43	.77	-1.34	.19	1.00	.90	.74	-.19	.20	1.20	1.30
44	.74	-1.20	.19	1.00	1.00	.63	.28	.19	1.10	1.20
45	.54	-.39	.18	1.10	1.20	.51	.76	.18	.90	.90
46	.49	.16	.19	1.00	1.00	.85	-1.15	.24	.90	.80
47	.77	-1.10	.18	.90	.90	.69	-.23	.20	.90	.80
48	.87	-1.73	.21	.90	.90	.61	.11	.19	1.00	1.00

(Table 66 continues)

(Table 66 continued)

	Form A Pretest					Form B Posttest				
Item	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ
49	.84	-1.46	.19	.80	.80	.31	1.39	.19	1.10	1.00
50	.84	-1.53	.20	.90	.80	.59	.25	.19	1.00	1.00
52	.63	-.36	.18	1.00	1.00	.73	-.43	.20	.80	.70
53	.46	.30	.19	1.00	1.00	.58	.28	.19	.80	.80
54	.54	-.01	.18	.90	.80	.57	.32	.19	.80	.90
55	.70	-.64	.18	.90	.90	.61	.14	.19	.90	.80
56	.58	-.17	.18	.90	.80	.74	-.51	.21	1.00	1.00
57	.76	-1.03	.18	.80	.80	.66	-.08	.19	.90	.80
58	.72	-.77	.18	.90	.90	.51	.53	.18	1.00	1.00
59	.55	-.04	.18	1.00	.90	.66	-.11	.19	.90	.80
60	.56	-.07	.18	1.10	1.10	.76	-.56	.21	1.00	.90
61	.50	.13	.18	.90	.80	.47	.76	.18	.90	.90
62	.55	-.01	.18	.90	.90	.58	.25	.19	.80	.80
63	.42	.49	.20	1.00	.90	.59	.21	.19	.90	.90
64	.40	.52	.20	.90	1.00	.62	.07	.19	.90	.90
65	.28	**1.09**	**.23**	**1.00**	**1.00**	.58	**1.09**	**.23**	**1.00**	**1.00**
66	.38	**.60**	**.14**	**1.10**	**1.10**	.52	**.60**	**.14**	**1.10**	**1.10**
67	.36	**.01**	**.13**	**.90**	**.80**	.79	**.01**	**.13**	**.90**	**.80**
68	.22	**.69**	**.14**	**.90**	**.90**	.57	**.69**	**.14**	**.90**	**.90**
69	.16	**1.16**	**.15**	**.90**	**.90**	.44	**1.16**	**.15**	**.90**	**.90**
70	.36	**.84**	**.14**	**1.10**	**1.10**	.43	**.84**	**.14**	**1.10**	**1.10**

Note. Item difficulty estimates in bold are common items.

Table 67

Rasch Item Statistics for the Form B Pretest and Form A Posttest Taken by Group B

	Form B Pretest					Form A Posttest				
Item	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ
1	.54	.31	.21	1.00	1.00	.37	1.02	.22	1.30	1.30
2	.67	-.31	.21	.90	.90	.80	-.99	.27	1.00	1.00
3	.19	1.89	.29	1.00	.90	.73	-.54	.24	.90	.80
4	.55	.22	.21	1.00	1.00	.81	-1.07	.28	.90	.80
5	.43	.67	.22	1.00	1.00	.76	-.79	.26	1.00	.90
6	.26	1.44	.26	1.00	1.00	.73	-.54	.24	1.00	.90
7	.68	-.31	.21	1.00	1.00	.61	.00	.22	1.00	1.00
8	.46	.58	.22	1.00	1.10	.87	-1.51	.32	.90	.70
9	.91	-1.94	.29	.90	.80	.55	.24	.22	1.00	1.00
10	.61	.04	.21	1.00	1.00	.28	1.35	.22	1.00	1.00
11	.33	**.57**	**.15**	**.90**	**1.00**	.83	-1.15	.28	.90	.70
12	.46	**.83**	**.15**	**1.00**	**1.00**	.63	-.15	.23	1.00	1.00
13	.81	**-.56**	**.16**	**1.00**	**1.10**	.37	.97	.21	1.10	1.20
14	.62	**.23**	**.15**	**1.00**	**1.00**	.36	1.07	.22	1.20	1.30
15	.65	**.07**	**.15**	**.90**	**.90**	.72	-.48	.24	.90	.80
16	.47	.53	.22	1.10	1.10	.58	**.57**	**.15**	**.90**	**1.00**
17	.71	-.49	.21	1.10	1.10	.36	**.83**	**.15**	**1.00**	**1.00**

(Table 67 continues)

(Table 67 continued)

	Form B Pretest					Form A Posttest				
Item	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ	IF	Difficulty	*SE*	Infit MNSQ	Outfit MNSQ
18	.03	4.01	.72	1.00	1.70	.65	**-.56**	**.16**	**1.00**	**1.10**
19	.51	.35	.21	1.00	1.00	.50	**.23**	**.15**	**1.00**	**1.00**
20	.41	.72	.22	1.10	1.10	.53	**.07**	**.15**	**.90**	**.90**
21	.35	1.08	.24	1.10	1.10	.55	.24	.22	.90	.90
22	.26	1.44	.26	1.10	1.30	.26	1.50	.23	.90	.90
23	.59	.00	.21	.90	.90	.62	-.10	.23	.80	.80
24	.76	-.77	.22	.90	.90	.81	-1.07	.28	.90	.70
25	.54	.22	.21	.90	.90	.61	.00	.22	.80	.80
26	.25	1.44	.26	1.00	1.00	.68	-.66	.25	1.00	1.00
27	.45	.49	.21	1.00	1.00	.90	-2.13	.40	1.00	.80
28	.51	.31	.21	1.00	1.00	.60	-.37	.23	1.20	1.20
29	.77	-.92	.23	1.00	1.00	.85	-1.51	.32	1.00	1.00
30	.56	.09	.21	1.10	1.10	.62	-.37	.23	1.10	1.20
31	.90	-1.86	.29	1.10	1.10	.67	-.60	.25	.90	.80
32	.73	-.72	.22	1.10	1.10	.20	1.55	.23	1.20	1.40
33	.21	1.65	.27	1.10	1.30	.47	.24	.22	1.20	1.20
34	.41	.67	.22	1.10	1.10	.54	-.10	.23	1.10	1.10
35	.15	2.08	.31	1.00	1.00	.49	.14	.22	1.10	1.10
36	.79	**-1.24**	**.19**	**.90**	**.80**	.85	**-1.24**	**.19**	**.90**	**.80**
37	.86	**-.89**	**.17**	**1.10**	**1.30**	.59	**-.89**	**.17**	**1.10**	**1.30**
38	.65	**-.33**	**.16**	**.90**	**.90**	.63	**-.33**	**.16**	**.90**	**.90**
39	.42	**.43**	**.15**	**1.00**	**.90**	.47	**.43**	**.15**	**1.00**	**.90**
40	.47	**.52**	**.15**	**1.10**	**1.20**	.38	**.52**	**.15**	**1.10**	**1.20**
41	.24	1.51	.26	1.10	1.20	.82	-1.51	.32	1.00	1.00
42	.52	.26	.21	1.10	1.10	.31	.93	.21	1.20	1.20
43	.45	.49	.21	1.00	1.00	.88	-1.98	.38	1.00	.90
44	.65	-.31	.21	.90	.90	.91	-2.30	.43	.90	.70
45	.45	.49	.21	.90	.90	.45	.28	.22	1.20	1.20
46	.77	-.92	.23	.90	.80	.27	1.16	.22	1.20	1.30
47	.69	-.44	.21	.90	.80	.72	-.92	.27	.90	.80
48	.69	-.49	.21	.90	.90	.85	-1.72	.34	1.00	.90
49	.24	1.51	.26	1.10	1.30	.69	-.79	.26	.80	.70
50	.39	.82	.22	1.20	1.40	.82	-1.51	.32	1.00	.80
51	.43	.67	.22	1.00	1.10	.39	.52	.21	.90	.90
52	.58	.04	.21	.90	.90	.76	-1.07	.28	.90	.90
53	.55	.13	.21	1.00	1.00	.49	.09	.22	1.00	.90
54	.45	.53	.22	1.00	.90	.55	-.10	.23	.80	.80
55	.57	.04	.21	.90	.90	.72	-.92	.27	.90	.90
56	.75	-.72	.22	.90	.90	.57	-.26	.23	1.00	.90
57	.43	.67	.22	.90	.90	.73	-.99	.27	.90	.80
58	.48	.44	.21	1.00	1.00	.79	-1.32	.30	.90	.70
59	.55	.17	.21	1.00	1.00	.55	-.10	.23	1.00	.90
60	.79	-1.02	.23	1.00	1.00	.44	.33	.22	1.10	1.10
61	.46	.49	.21	.90	.90	.82	-1.51	.32	.90	.70
62	.66	-.31	.21	1.00	.90	.76	-1.07	.28	.90	.80
63	.60	-.04	.21	1.00	1.00	.43	.33	.22	1.20	1.20
64	.52	.22	.21	1.00	1.10	.59	-.26	.23	1.00	1.00
65	.56	.04	.21	1.00	1.00	.22	1.40	.22	1.20	1.30
66	.43	**.69**	**.15**	**1.10**	**1.20**	.35	**.69**	**.15**	**1.10**	**1.20**
67	.51	**-.02**	**.15**	**.90**	**.90**	.62	**-.02**	**.15**	**.90**	**.90**
68	.44	**.57**	**.15**	**1.00**	**1.00**	.39	**.57**	**.15**	**1.00**	**1.00**
69	.29	**.76**	**.15**	**.90**	**.90**	.42	**.76**	**.15**	**.90**	**.90**
70	.53	**.23**	**.15**	**1.10**	**1.10**	.45	**.23**	**.15**	**1.10**	**1.10**

Note. Item difficulty estimates in bold are common items.

Thus, the item did not discriminate well, and choices a and d attracted high ability examinees. This item should be revised so that the item difficulty estimate is nearer the cut-point and the correct choice attracts high ability examinees.

Tables 68 and 69 show the occasion difficulty estimates for Group A, who took the Form A pretest and the Form B posttest, and for Group B, who took the Form B pretest and the Form A posttest. The Form A pretest was more difficult for Group A than the Form B posttest by .52 logits. Likewise, the Form B pretest was more difficult for Group B than the Form A pretest by .28 logits. The occasion separation indices of 11.06 and 4.79, and the Rasch occasion reliability of .99 and .96 showed a significant difference in the occasion separations. The chi-square tests were both significant; thus, the differences in occasion separation were significant.

Table 68

Rasch Occasion Statistics for the Form A Pretest and Form B Posttest Taken by Group A

	Form A Pretest	Form B Posttest
Occasion difficulty	.19	-.19
SE	.02	.02

Table 69

Rasch Occasion Statistics for the Form B Pretest and Form A Posttest Taken by Group B

	Form B Pretest	Form A Posttest
Occasion difficulty	.24	-.24
SE	.03	.03

CHAPTER 7

DISCUSSION

In this chapter, the four research questions investigated in this study are discussed in light of the literature on criterion-referenced testing. This chapter is made up of four sections: In the first section, I discuss the soundness of the criterion-referenced items based on the criterion-referenced item statistics. In the second section, I consider the extent to which facets such as persons, classes, items, sections, and subtests contributed variability to the test scores based on the results of the G studies and FACETS analyses. In the third section, I discuss the dependability of the two CRT forms based on the results of the D studies. In the fourth section, I address the validity of the CRT forms based on pretest/posttest score gains and person ability estimates in the FACETS analyses.

Research Question 1: Item Effectiveness

Research Question 1 asked to what extent the criterion-referenced items functioned effectively for the students in this study. First, a summary of the results on norm-referenced and criterion-referenced item statistics such as item facility, item discrimination, the difference index, and the *B*-index is given, followed by the discussion.

Norm-Referenced and Criterion-Referenced Item Analyses

Item facility should be low on pretests because it indicates that students require instruction on the course objectives. Several items, such as items 8, 48, and 50 on Form A, and items 9, 13, 29, 31, 36, 37, and 60 on Form B, had high IF values on the

pretests (Tables 21 and 22). These items performed poorly for several reasons. Item 8 on Form A, which tested the receptive meaning of *apologize*, was known already by many of the students. Also, the distractors for that item, *thank*, *show feelings*, and *make a request*, did not attract many responses. Item 48 on Form A was a factual question in which the students had to find a specific fact in the passage. The correct choice was a paraphrase of what was in the passage. Factual questions tend to be easy (Alderson, 2000, p. 87). Item 50 on Form A was an inference question in which the students had to have an overall idea of the passage to answer the item correctly. Although inference questions tend to be relatively difficult, this item was easy possibly because the correct choice (c) was straightforward and the distracters (a, b, and d) were somewhat unrelated to the content of the passage.

Conversely, item facility should be high on posttests because it indicates that the students have learned the objective measured by the items. Several items, such as item 32 on Form A and item 33 on Form B, had low IF values on the posttest. When these items were examined with a distracter analysis, it was found that one distracter was attracting many students; this illustrates how poor distracters can detract from the main purpose of assessing whether the educational objective has been attained or not. Except for these twelve items, the rest of the items were working well as pretest and posttest items.

Item discrimination indicates how well items discriminate between high and low scoring students. The values for only four items, item 16 on the Form A pretest, and items 18, 22, and 50 on the Form B pretest, were negative; thus, those items did not discriminate well; however, many of the remaining items could be used for norm-referenced purposes, as they had point-biserial correlations above .20. However,

item discrimination is not an adequate statistic for criterion-referenced testing because spreading students' test scores is not a purpose of CRTs.

Inspecting item facility values can be useful when revising CRT items, objectives, lesson plans, and developing an in-house placement test. First, diagnostic test items with high item facility values should be omitted when revising the test because many students have already learned them. Also, teachers should consider omitting achievement test items with low item facility values if they believe that the objectives assessed by the items are too difficult for students to learn in the given teaching context.

Second, these diagnostic and achievement items also indicate whether the objective that the items are testing is too easy or difficult. Inappropriate objectives should be revised so that they are more suitable for students. In this study, three objectives for this class were for students to correctly answer at least 60% of the vocabulary questions, listening comprehension questions, and reading comprehension questions. The means of the item facility values for the vocabulary, listening, and reading subtests were .40, .47, and .30 on the Form A pretest, respectively. The objectives that the pretest items were testing were not fully mastered. Thus, the difficulty and content of the objectives were set properly.

Third, if teachers believe that the objectives are learnable but not acquired adequately, more effective lesson plans must be formulated so that better learning outcomes can be obtained. The means of the item facility values for the vocabulary, listening, and reading subtests were .66, .64 and .68 on the Form B posttest, respectively. Thus, Group A students did not fully mastered the objectives. In this case, the teachers might need to rethink their lesson plans so that more effective learning

occurs. Especially, the teachers need to improve the listening lessons because only a slight gain was observed in the mastery of the listening objectives.

Fourth, all of the items in this study were originally designed for criterion-referenced purposes and were revised to achieve those purposes, but these criterion-referenced items can also be used for norm-referenced purposes because of their generally high discriminatory power. It is possible to develop a placement test by assembling highly discriminating criterion-referenced items. This can be the most efficient strategy for developing curriculum-content-specific in-house placement tests because the items are referenced to specific curricular objectives and are administered to and revised for the students studying in that program.

The difference index is the best indicator of effective criterion-referenced items because it shows the degree of students' mastery of the objectives assessed by the items. Only four items, items 1 and 32 on Form A, and items 13, 33, and 37 on Form B, had negative DI values. This indicated that most of the students learned the objectives measured by other items to some extent with the degree of learning indicated by the magnitude of the value. Item 18 on Form B and item 30 on Form A with DI values of .63 and .60, respectively, had the highest values. The students did not know the meaning of the target word, *altitude*, on item 18 on the pretest, but effectively learned it by the time of the posttest. Item 30 was an inferential listening question in which the students had to infer the main point of the speech. The inference question was difficult on the pretest, but it became relatively easy on the posttest presumably due to the class instruction.

One caution concerns interpreting these difference index values. In this study I adopted a two-form counterbalanced design, so the values were calculated with two groups that might have differed in terms of their proficiency levels. Because three

teachers taught these classes, it was possible that they focused on slightly different aspects of the syllabus. These differences in the groups and teachers might have affected the values to some extent.

Table 15 displays the *B*-index values at the cut-points of 50%, 60%, and 70%; less than 15 items had negative values on the pretests, and only five items had negative values on the posttests. This result indicates that most of the items were able to differentiate pass/fail students well. Two findings can be pointed out here. One was that the average *B*-index values for the listening subtests were lower than the values on the other subtests. The values for the Form A pretest at the 70% cut-point were .36, .38, .30, and .38 for the total score, vocabulary, listening, and reading subtests, respectively. Even on the Form B posttest, the average value for the listening subtests was lower than the ones on the other subtests. The other finding was that the values on the pretests at the 70% cut-point were higher than the ones at the 50% cut-point. On the other hand, the values moved in the opposite direction on the posttests; the values at the 50% cut-point were higher than the ones at the 70% cut-point. If a cut-point is far from the mean, the *B*-index tends to increase.

Hudson and Lynch (1984) administered an achievement test to test students' mastery of course objectives. The participants consisted of an instructed group (n = 57) and uninstructed group (n = 42). The difference index was calculated by subtracting the item facility value for the instructed group from the value for the uninstructed group. They reported that the average DI values on listening, reading, and grammar subtests were .27, .34, and .38, respectively. The average DI values for vocabulary, listening, and reading subtests in this study were .23, .22, and .29, respectively. The values were slightly lower in this study mainly because the

pretest-posttest design used in this study was different from their instructed-uninstructed group design.

Brown (1989) reported the results of a 60-item reading placement test given to 61 students as a pretest and posttest. The item discrimination values were high, but the difference index values were relatively low and ranged from -.03 to .35. In this study, the difference index values ranged from -.15 to .63. They were generally higher because the items were designed to measure a specific objective and were, as a consequence, more sensitive to the instruction. Even compared to other studies that reported difference index and *B*-index values (Brown, 1993, 2001; Griffee, 1995), the criterion-referenced items in this study worked effectively.

There are at least four possible reasons why these criterion-referenced items worked well. First, the items were carefully designed as criterion-referenced items in that the items and test purposes were well matched. This stands in contrast to the use of norm-referenced items, such as TOEIC items, which are sometimes administered to estimate students' mastery of course content. Such items will not typically work effectively and this will be reflected in low DI values. The criterion-referenced items used in this study reflected the strong linkage among the contents of the objectives, items, materials, and teaching. Because what was learned and tested was similar, the students likely felt the need to study in class.

The second reason concerns the process of item development. As shown in Chapter 4, two pilot studies were conducted to fine tune the items. The items were administered to gather data on item performance, analyzed using criterion-referenced item analyses, and revised based on the results of the item analyses. Without engaging in that process, the items would not have worked as effectively. It is generally not possible to develop a dependable and valid CRT in one administration and/or without

piloting; teachers must gather data on criterion-referenced items, revise them based on the data, revise poor performing items, and recycle sound items.

The third reason the items worked well is that most of them were practiced by the students before actually taking the final examination. One of the two parallel test forms, which was slightly different in terms of content but the same in terms of item format and test procedure, was administered as a diagnostic pretest to give the students a chance to become more familiar with the form and content of the final examination. In addition, most of the multiple-choice items on the test were practiced as open-ended questions in class. The answers were also verified and corrected by the students. These practice effects helped the items work effectively. Some teachers believe that content that has been practiced in class should not be on a final examination because it is not challenging or they believe in using a norm-referenced approach to classroom assessment. However, from a criterion-referenced point of view, these teachers should teach to the test, aim to enhance practice effects, and test students' achievement using items directly related to class objectives.

The fourth reason is that teachers announced the test content one month before the final examination and informed the students that the test content would be similar to that on the practice test. The number of items and chapters tested, the item format, and recommended ways of studying for the test were explained in detail. This assured that the students knew what and how they were to be tested and they were able to study effectively for the test. All testing-related information should be provided to students because teachers are responsible for explaining their testing practices and students should be able to use any necessary information to do well.

Research Question 2: The Contribution of the Test Facets

Research Question 2 asked to what extent the facets of persons, items, sections, classes, and subtests contribute to the total score variation on the two CRT forms when given as pretests and posttests. In this section, the results of the G studies and FACETS analyses are synthesized, and the effects of each facet are discussed in the order of persons, items, sections, classes, and subtests.

Facets Variance

Tables 25, 26, 27, and 28 show the percentages of the persons variance components estimated in the 12 G studies with the *p X (i:s)* design. The percentages of the persons variance components on the pretests and on the posttests ranged from 0% to 6%, and from 2% to 4%. Tables 29, 30, 31, and 32 display the percentages of the persons variance components estimated in the 12 G studies with the *(p:c) X (i:s)* design. The percentages were about the same range as the ones with the *p X (i:s)* design, ranging from 3% to 9% on the pretests, and from 2% to 9% on the posttests. When the three subtests were treated as a facet in four G studies with the *p X (i:s)* design, the percentages of the persons variance components was 4% on the pretests, and 6% and 11% on the Form A and Form B posttests (Tables 33 and 34). The results of the multivariate G studies with the $p^{\cdot} X i^{\circ}$ design (Tables 35, 36, 37, and 38) showed that the percentages of the persons variance components ranged from 4% on the Form B listening subtest to 12% on the Form A reading subtest on the pretests, and from 5% on the Form A listening subtest to 17% on the Form B reading subtest on the posttests. Tables 39 and 40 showed that only the persons variance components on the Form B posttest taken by Group A was over 10%.

The FACETS analyses revealed similar results to the G studies. The Rasch person separation indices for the Form A pretest, Form B pretest, Form A posttest, and Form B posttest were 1.92, 1.90, 2.35, and 2.91, respectively. The Rasch person reliability coefficients for the pretests were .79 and .78 while the reliability coefficients for the posttests were .85 and .89. Thus, the posttest created the largest separation in person ability. Similarly, the persons variance components on the posttests ranged from 7% to 17% (Tables 37 and 38) and had larger percentages than the ones on the pretests, which ranged from 4% to 12% (Tables 35 and 36). This indicates that there were greater differences in the students' achievement levels on the posttests, mainly due to the different amount of effort they put into studying for the class.

Six observations can be made about the persons variance components. First, Group A had higher percentages of persons variance components because more students with different proficiency levels were included in this sample. Second, the persons variance components of the Form A reading subtest on the pretest had the largest value at 6%. This occurred because the test score was not graded and some students did not feel pushed to do well. It would be possible to include students' pretest scores and gain scores in their course grade because even pretests can be considered as part of learning in class. This approach might encourage students to work hard even on the pretest. Also, the students were not provided with sufficient time to finish some sections of the reading subtest. In situations where teachers are not given enough time to administer pretests, a short version of the tests could be administered so that students could finish all the items in the given time constraints. Third, the G study and FACETS analyses indicated that the persons variance components and the Rasch person separation indices on the posttests were greater

than the ones on the pretests. This could be due to differences in the amount of effort the students in the two groups made to perform well on the posttest. Fourth, the persons variance components on the listening subtests generally accounted for lower percentages than on the other subtests. This shows that the students' listening mastery on these subtests did not vary greatly. Fifth, the persons variance components were fairly small compared to the other effects. This finding is positive because in criterion-referenced testing, zero or small percentages of persons variance components should be observed, as this indicates that a homogeneous group of students with similar backgrounds and proficiency levels has taken the criterion-referenced items. In previous studies (Brown, 1999; Brown & Ross, 1996; Zhang, 2006), the persons variance components for the TOEFL and TOEIC were larger, ranging from 6.68% to 17.82%. Because a variety of examinees in terms of their proficiency levels take the TOEFL and TOEIC, it is reasonable that persons variance is relatively large. Sixth, when different G study designs were set, the magnitudes of the persons variance differed slightly. This occurred because the facets adopted and the data used to conduct the analyses were different. However, the percentages did not change drastically regardless of the difference in the designs.

In the 12 univariate G studies with the *p X (i:s)* design, the items variance components were over 35%, which was larger than the percentages of the persons variance components (Tables 25, 26, 27, and 28). Also, the percentages on the posttests were over 50%, even greater than those on the pretests. However, the items variance components in the 12 G studies with the unbalanced *(p:c) X (i:c)* design accounted for smaller percentages of the variance, ranging from 4% to 18% (Tables 29, 30, 31, and 32). The items variance components in the four G studies with the unbalanced *p X (i:s)* design accounted for about 10% of the variance (Tables 33 and

34). In the four multivariate G studies for the pretests and posttests with the $p^{\cdot} X i^{\circ}$ design, the percentages of the items variance components ranged from 6% to 17% on the pretests, and 5% to 14% on the posttests (Tables 35, 36, 37, and 38). The Form B reading subtest had lower percentages at 6% on the pretest and 5% on the posttest. The listening subtest on the Form B pretest had the largest percentages at 17%. The results of the two multivariate G studies shown in Tables 39 and 40 indicated that about 10% of the total variance was due to the items variance.

Two observations can be made concerning the items variance. First, the findings of the G studies suggest that the percentages of the items variance components in the unbalanced *(p:c) X (i:s)* design were reduced and the interaction effects increased compared to the results obtained from the balanced *p X (i:s)* design. When more facets are taken into account in the design, there is a tendency for the percentages of variance accounted for by the interaction effect to increase because the interaction effects are composed of a combination of the facets effects. Second, generally, the percentages of the items variance were greater than those of the persons variance. This result is similar to Zhang's (2006) study of the TOEIC in which the items variance was greater than the persons variance, and contrary to Brown and Ross' (1993) study of the TOEFL in which the persons variance was twice as much as the items variance. This finding suggests that items variance is generally greater than persons variance in criterion-referenced testing; this result is to be expected in criterion-referenced testing situations because the test takers are expected to have similar proficiency levels and backgrounds and should study a fairly wide range of materials. When developing a curriculum, the content should be both learnable and challenging for a particular group of students and should be presented in a well-sequenced manner so that second language acquisition is facilitated.

The FACETS analyses on the Form A and Form B pretests and posttests showed that the Rasch item separation indices were 3.58, 3.78, 2.35, and 2.91, respectively, and the Rasch item reliability coefficients were .93, .93, .85, and .89, respectively. These indices showed that the pretest item difficulty estimates varied more widely than the posttest estimates. This suggests that the students began the semester with different levels of knowledge concerning the test items, but this variance was reduced on the posttests, as the students who did poorly on the pretest were able to increase their scores and catch up with their more knowledgeable classmates to a degree.

In the univariate G studies for the balanced and unbalanced designs, no variance component was observed due to sections (Tables 25, 26, 27, 28, 29, 30, 31, and 32). This result is consistent with previous studies (e.g., Kumazawa, 2009). This occurred because the textbook was designed for a particular group of students, and the difficulty of the passages was targeted at an intermediate level in order to make the passages comprehensible. Thus, there was little difference in the readability of the different passages (Table 2). If a textbook is targeted appropriately at a particular group of students, then sections variance components should be small. When adopting materials, teachers should carefully examine textbooks to see if language points such as grammar and vocabulary and tasks used in the textbooks are well sequenced and learnable for their students. The only exception was the G study on the reading subtest on the Form A pretest, which showed that 9% of the total variance was due to the section. This occurred because the students in one particular class were not given enough time to finish all the items on the reading subtest and the subtest therefore appeared difficult. When conducting a study, research conditions should be carefully controlled so that unwanted effects can be minimized.

In the 12 univariate G study results shown in Tables 29, 30, 31, and 32, only 1% to 6% of the variance was due to classes on the pretest. In a proficiency-based curriculum where a placement test is used to place students into an appropriate level, having a classes facet in a G study design is potentially informative because the variance component can indicate how much the classes differ in terms of proficiency level. Large percentages of classes variance can be an argument that the placement test discriminates the students' proficiency levels reasonably well. However, on the posttest, the percentages of variance due to classes were large, especially for Group A, who took Form B. This result indicates that the students and teachers had different levels of motivation to do well. As a consequence, they produced different levels of achievement on this CRT.

Tables 33 and 34 show that a small proportion of the subtests variance components was extracted on the Form A pretest and posttest, but 0% of the variance was due to the subtests. This indicated that the difficulty of the vocabulary, listening, and reading subtests differed only slightly on Form A. Two G studies with the *p˙ X (i°: s˙)* design (Tables 39 and 40) showed the same results: 0% of the variance was due to the subtests.

In the FACETS analyses, the Rasch subtest reliability estimates for the Form A pretest and posttest were .58 and .72 (Tables 43 and 47). A reliability estimate of .00 was observed on Form B when it was given as a pretest and posttest (Tables 44 and 48), showing that the subtests were equal in the difficulty. Although the decision to make all subtests equal in difficulty depends on the testing purpose, it is generally undesirable to have one subtest unequal in difficulty because it makes comparisons among subtest scores difficult, particularly if raw scores are the basis of the comparison. However, using Rasch analysis (e.g., FACETS), comparisons can be

made through common item or person linking if the purpose of testing (e.g., diagnostic and achievement) is the same across the administrations.

An inspection of the interaction effects in the 12 univariate G studies with the *p X (i:s)* design indicated that more than 50% of the total variance on the pretests was due to the persons-by-items nested within sections interaction effect. Less than 45% of the variance on the posttest was due to the interaction effect. However, the unbalanced G study with the *p X (i:s)* design and the multivariate G studies indicated that more than 60% of the total variance was due to the interaction effect. This means that different students know different aspects of the language being tested with the CRTs (Brown, 2007). The size of this effect is consistent with that obtained in previous studies (Brown, 1999; Brown & Ross, 1993; Zhang, 2006). Regardless of whether the testing purpose is norm-referenced or criterion-referenced, interaction effects or undifferentiated errors are typically fairly large because various factors, such as teaching and individual differences in learning English, affect students' learning, and these factors are unidentifiable with a paper-and-pencil test. In other words, a number of unknown and/or unmeasured facets are inherent in foreign language testing, and it is difficult to identify them due to the complexity of language learning. Also, a single facet has a rather small effect on examinees' test performance, but when multiple facets interact with each other, they account for a greater amount of variance.

The G studies conducted in this study were useful for two reasons. First, because complex G study designs were adopted by taking account of not only items but also sections, classes, and subtests, the results provided a great deal of information on the facet effects and sources of error. These designs were able to show a more complete picture of what is tested, and this provides a better understanding of the

degree to which the test is valid. For example, simple single facet G study designs only provide the percentages of the persons, items, and interaction variance components. However, adding more facets in the designs allowed me to identify the amount of the facet effects and the interactions of the facets. It is highly recommended that complex designs be adopted so that researchers can obtain more information about sources of test variance.

Second, the univariate and multivariate G studies conducted in this study produced similar results. For example, both the univariate and multivariate G studies showed that the persons variance was smaller than the items variance. Obtaining similar results in both types of G studies allows researchers to have greater confidence in the results.

The results of the FACETS analyses conducted by taking account of person, item, and subtest facets were useful because they provided a vast amount of information on various measurement aspects, including person ability, item difficulty, and subtest difficulty estimates. FACETS analyses do not indicate the amount of the interaction effect as G theory does, but it can be checked by looking at Rasch person fit statistics and bias analyses. The Rasch separation indices and Rasch reliability coefficients were provided for all the facets, and the standard error of measurement showed, for example, which item difficulty estimates had greater or lesser error. Because all the facet estimates are probabilistic and are placed on the same interval scale, they show not only the difficulty but also the distances among the difficulty estimates and the probability of a person getting an item correct or incorrect. This kind of information is especially useful when constructing a measure and revising it to make it more suitable for testing purposes. When developing CRTs, test developers should select items with the item difficulty estimates around the cut-point(s) so that

dependable pass/fail decisions can be made. In Figure 14, a horizontal line shows the arbitrary 60% pass/fail cut-point. Items that have item difficulty estimates near the cut-point are useful because they point where the most crucial decision is made. However, Item 33 on the Form B posttest is not useful when making pass/fail decisions because an examinee with the person ability estimate of .00 has no chance of getting the item correct. In addition, when there are several subtests, composed of different numbers of items on each subtest, it is not easy to compare which subtests were more or less difficult by examining the means. However, because the Rasch model places subtest difficulty estimates on the same interval scale, it is easy to compare the relative difficulty of the subtests.

Research Question 3: Dependability

Research Question 3 asked to what extent the two CRT forms are dependable when administered as pretests and posttests. A summary of the results on the dependability indices estimated by Brown's short-cut formula, phi lamda dependability, and univariate and multivariate D studies, followed by a discussion of the results, is provided below.

Dependability of the Criterion-Referenced Test Forms

The dependability indices using Brown's short-cut formula are displayed in Tables 49, 50, 51 and 52. They showed that all the subtests given as a pretest and posttest had moderate or high dependability indices, ranging from .43 to .84. The listening subtests had lower indices mainly because fewer items were included on these subtests. In addition, small *CI* values, ranging from .1097 to .0890, were found on all the subtests. Thus, teachers can make dependable decisions about the students'

grades based on these test statistics. In Brown's (1993) study, the dependability indices estimated with his short-cut formula ranged from .50 to .71, and the indices of the 24-item listening CRT forms ranged from .41 to .71. Even compared with Brown's (1993) study, the indices in this study were the same or higher mainly because the items on the subtests were piloted and revised to make them suitable for the test purposes.

Phi lamda dependability indices showed that it was possible to make fairly dependable pass/fail decisions based on the students' test scores. For instance, in this study, it was best that the cut-point was set at 50% because of the high indices on the posttests (Tables 19 and 20). One way to use this index is that it can be an argument for the adequacy of the location of a cut-point set for a CRT. If a cut-point has to be set in order to make pass/fail decisions, teachers should examine this index to see how high or low these indices are for different possible locations and select the optimal cut-point.

When an additional facet is included in a *p X I* study design, the dependability is underestimated because the design adds another source of error in the estimate. For instance, a sections facet was added to the *p X I* design. The dependability of the vocabulary Form A pretest was .73 (Table 49), but it was .44 when the D study with the *p X (I:S)* design was conducted. As Li and Brennan (2007) noted, the *p X I* design is simplistic, does not take account of other facets in testing, and overestimates the dependability; thus, the design should be used cautiously.

When the three subtests were treated as one facet rather than as separate constructs in the *p X (I:S)* design, changes in the dependability indices were insensitive to changes in the number of items. For instance, the indices of a 20-20-20 and 25-25-25 item form would be .55 and .57, respectively (Table 55). Adding 15

items would only increase the index by .02. Thus, this design should be used cautiously because of the insensitivity of the indices to the changes in the number of items.

When compared to the results of previous studies in which designs with two or more facets were used, the dependability indices in this study were relatively low. For example, Zhang (2006), in his study of the TOEIC, indicated that the generalizability coefficients for absolute decisions were .96 and .98 when estimated with the *p X (I:Sections)* and *(p:Languages) X (I:S)* designs. Criterion-referenced testing indices should be lower than norm-referenced testing coefficients because the purpose is not to spread examinees' test scores and fewer items are typically given to fewer test takers in criterion-referenced contexts.

Univariate generalizability coefficients estimated in a multivariate D study with the $p^{\cdot} X I^{\circ}$ design should be equivalent to the dependability indices estimated using Brown's short-cut formula. For instance, the dependability for the vocabulary subtest on the Form A pretest was .73 (Table 17). Notice that the univariate generalizability coefficient for the 25-item vocabulary measure was .72 (Table 41). This shows that Brown's dependability formula and univariate D studies in a multivariate D study work in the same way when the $p^{\cdot} X I^{\circ}$ design is used.

A multivariate D study not only simultaneously conducts univariate D studies for each dimension, but also provides a multivariate generalizability coefficient for multiple subtests. The multivariate generalizability coefficients for absolute decisions ranged from .77 to .90 when there were 25-20-25 items on one form (see Tables 56, 57, 58, and 59). These values were equivalent to the indices of the total scores estimated with Brown's short-cut formula (see Tables 17, 18, 19, and 20). This design is useful in determining the number of items needed to maximize the index. For

instance, to achieve an index of .82, three possible combinations, 25-15-25, 25-20-25, and 25-25-25 item tests, can be used. Increasing or reducing the 20-item listening subtest by five items would not change the index. Although it might not be desirable to use a 15-item listening test because the univariate generalizability coefficient for absolute decisions would fall to below .50, it would be a poor use of time to develop five more items and add them to the 20-item test if the index would not change.

Tables 60 and 61 show that the multivariate D studies with the *p X (I°:S˙)* design provided composite generalizability coefficients for absolute decisions of .83 and .79 when the Form A pretest and Form B posttest were administered to Group A and when the Form B pretest and Form A posttest were administered to Group B. In this design, the vocabulary, listening, and reading subtests were treated as a single facet. Estimating a univariate generalizability coefficient for multiple subtests tends to reduce the magnitude of the coefficients because it lacks consistency. For instance, the multivariate generalizability coefficient for the Form A pretest was .82 when the form contained 25-20-25 items. The univariate generalizability coefficient for the Form A pretest with 25-20-25 items was .50. Thus, the dependability estimated when treating multiple constructs as a facet should be interpreted cautiously.

However, this *p X (I°:S˙)* design can be informative when deciding which forms should have more or fewer items to achieve a certain generalizability coefficient. Students' test scores on diagnostic tests are not usually part of the course grade, so the test does not have to be long and highly dependable. However, generalizability coefficients for posttests should be high because teachers need to make dependable decisions concerning students' grades. For example, the univariate generalizability coefficient for a 20-15-20-item pretest was .48 (Table 61). This is not highly dependable, but teachers can diagnose students' mastery level with the test.

Conversely, the univariate generalizability coefficient for a 25-20-25-item posttest was high at .86; teachers can have confidence in the test scores and make dependable decisions regarding students' grades when the coefficient is this high. The univariate generalizability coefficient for the pretest was low but the composite generalizability coefficient was still high at .83.

Research Question 4: Validity

Research Question 4 asked to what extent the two CRT forms were valid when administered as pretests and posttests. Three aspects of validity are discussed: content validity, construct validity, and the washback effect.

Validity of the Criterion-Referenced Test Forms

Content validity is the systematic investigation of the degree to which the items on a test and the resulting scores are a representative and relevant sample of whatever content or abilities the test has been designed to measure (Brown & Hudson, 2002, p. 213). For these arguments, the test specification of the CRT in Appendix I can be used as one source of the theoretical argument approach. In this approach, teachers have to carefully plan a test on the basis of theoretical, research, or practical perspectives and argue for the content validity of the test on the basis of those perspectives (Brown & Hudson, 2002, p. 213).

The CRT specification in Appendix I was originally designed by a teacher who was in charge of teaching the class. Based on the specification, two full-time Japanese teachers who taught the class worked together to develop new CRT forms in 2005. The grammar and content of the CRT forms were checked by an English teacher who is a native speaker of English and taught the class, too. Thus, the items were not only

checked to ensure that they match the objectives described in Appendix G, but also to match the content of the class materials and teaching. Most of the items on the CRT were included in the class materials and were taught in class. For instance, one of the objectives described in Appendix I is to comprehend the 400-word passages in the textbook, which have a Flesch reading ease readability index of about 70, and answer at least 50% of the multiple-choice comprehension questions correctly. The textbook passages were read by the students in class. The CRT items were developed to test how well the students comprehended the passages. These items were practiced in class, and the teachers explained the answer to some of the test questions. Those strong links among the objectives, items on the CRT, class materials, and teaching supported the high content validity of the CRT.

Construct validity involves approaches that rely on experimental studies investigating the degree to which a test is measuring the psychological construct(s) it claims to be measuring (Brown & Hudson, 2002, p. 225). The three types of experimental studies are intervention construct validity, differential-groups construct validity, and hierarchical-structure construct validity studies. In this study, an intervention construct validity study was conducted. This study is the most appropriate approach because teachers can calculate and examine pretest and posttest score means and argue for the construct validity of their CRT when score gains are observed. The major goal of any curriculum is to maximize students' learning, and this study provides evidence for teaching effects on learning in the form of gain scores. The following statistics can be used to build an argument for construct validity: the difference index, repeated-measures MANOVA, and person ability estimates in ANOVA-like FACETS analyses. In addition, divergent/convergent validity arguments are given.

Tables 21 and 22 show that only six out of 140 items had negative DI values. These results indicate that the majority of the items were plausibly measuring the constructs taught in the class. The average DI values of .25 and .23 on Form A and Form B are high when compared to previous studies (e.g., Brown, 1993, 2001). This shows that the items are strongly linked to the class materials and that the students learned them successfully.

Tables 62 and 63 show that the means of the total scores on the pretest for Group A and Group B were 29.69 and 31.08, respectively, while the means on the posttest total scores were 42.01 and 47.28, respectively. That is, Group A and Group B were able to improve their score by 12.36 and 16.20 points after taking thirteen 90-minute lessons once a week. Also, the means for the vocabulary, listening, and reading subtests on the posttests for Group A and Group B were higher than the pretest means. The magnitude of the score gains was high compared to those in previous studies (e.g., Brown, 1993, 2001). The repeated-measures MANOVA showed that those score gains were statistically significant, indicating that the teaching affected the students' learning. Notice in Tables 64 and 65 that the eta square values are high. This shows that students' learning in the 13 weeks of the lessons best explained the increases in the test scores.

Table 64 shows that there was a significant interaction effect. This effect is illustrated in Figure 12, which shows that while the three lines rise, the line for the listening subtest is not parallel to the other lines. This shows that teaching the listening materials to the Group A students did not have any noticeable effect on their learning. A possible reason is that the listening lesson plans did not effectively make the lessons sufficiently comprehensible to the students. The other possible reason is

that it was rather difficult for the students to improve their listening within one semester.

The results of the two ANOVA-like FACETS analyses (Figure 14 and 15) show that the means of the person ability estimates for Group A and Group B increased by .73 and .43 logits, respectively. The results of the occasion difficult estimates (Tables 68 and 69) show that the Form A pretest and Form B pretest were more difficult than the Form B prettest and Form A posttest. The greater the difference on the pretest and posttest, the better the instruction can be said to have worked.

Correlation-based statistics are a generally weak argument for construct validity. Because the purpose of criterion-referenced testing is not to spread the test scores, but to maximize the pretest/posttest score gain, the variance is suppressed and this suppresses correlation-based statistics (Brown, 1990b). However, attenuated correlation coefficients estimated in a multivariate G study can be informative when building criterion-referenced construct validity arguments. Theoretically, the correlation coefficient between listening and vocabulary subtests should be lower than the coefficient between vocabulary and reading subtests because the latter two skills are more interrelated. The results of the multivariate G studies shown in Tables 35, 36, 37, and 38 provide attenuated correlation coefficients among the subtests. In all the results, the correlation coefficients between the vocabulary and listening subtests are lower than those between the vocabulary and reading subtests, and between the listening and reading subtests. These results support the construct validity of the CRT forms, showing that the subtests are tapping into different constructs and plausibly measuring what they are claimed to be measuring.

A positive washback effect, which is the positive “influence of testing on teaching and learning” (Gates, 1995, p. 101), is created when content validity is

supported. A well-known example of the washback effect concerns the mismatch between the content of the Japanese entrance examinations and Japanese high school teachers' use of the communicative approach as prescribed in *the Course of Study* in 1993 (Watanabe, 1996). In the present study, the high content validity of the test forms was ensured through a strong linkage among the objectives, testing, materials, and teaching; this enhanced the positive washback effect, and created better learning and teaching contexts. This claim is supported because the students were able to learn the content of the test context as indicated by the gain scores. Therefore, high content validity created a positive washback effect, which in turn led to high construct validity.

To foster positive washback, students' test scores and feedback about their performance were returned to them so that they could use the information for their further learning. For instance, the pretest scores were returned to the students, and on their answer sheet, the number correct for each section, percentage of items correct, and total score were indicated. The teachers reported the means for each of the subtests and the total score, and told them that these means were the cut-point they had to meet on the posttest. This provided a clear goal, showed my expectations to them, and ensured that they would recognize how much they had to study in order to receive the course grade they desired.

A major negative consequence of performing poorly on the posttest was possibly receiving a failing course grade. To avoid making invalid decisions about their course grade, teachers should not make pass/fail decisions based on one invalid indicator of achievement and based on grading-on-a-curve. In this study, the teachers did not make decisions solely on the basis of a single indicator of achievement. The students' posttest score was only one indicator of achievement, as it accounted for

40% of their final course grade. The number of lessons they attended (30%), class participation, and the points they received on their portfolio (30%) were also used to determine their final course grade. When no score gains were observed by comparing the pretest and posttest means, it might not be appropriate to use posttest grades because the CRT might not be a valid indicator of the students' achievement. In this study, there were a fair number of score gains, so it was appropriate to use posttest grades. Teachers should not decide their students' grade based on their test score ranking. Regardless of their score ranking, if students surpass a certain magnitude of mastery, they should get credit for their achievement. In this study, if their total achievement score was higher than 60%, they were able to get credit for the class because the decisions were absolute, not relative.

CHAPTER 8

CONCLUSION

This chapter consists of four sections. First, the limitations of this study are pointed out. Second, implications that can be drawn from this study are discussed. Third, suggestions for future research are made. Lastly, final conclusions are provided to sum up the study.

Limitations

Threats to internal validity concern issues such as history, testing, and mortality issues. With this in mind, the first limitation is that the participants took three other classes apart from the Reading and Listening Regular class, which was the research context in this study. As a result, the gain scores observed in this study might not be solely due to the instruction provided in the Reading and Listening Regular class. They could be due in part to the instruction the students received in other classes. Thus, it is not possible to interpret the cause of the score gains with complete confidence. Second, although counterbalancing reduced the practice effect, the score gains observed in this study could be due to the participants taking the CRT twice as a pretest and posttest. However, in criterion-referenced testing, a practice effect should be encouraged because it means that the students know what and how they are tested beforehand, so they can study for the test in an appropriate way. Third, some participants only took a pretest or posttest, and their test scores were omitted in the analysis of the score gains. Omitting the students' scores in the analysis, too, might

have had an effect on the degree of score gains. Lastly, unknown factors or flaws in the research design might have affected the results.

Implications

The results provide implications for CRT development as one part of curriculum development: (a) writing, (b) piloting and editing criterion-referenced items, (c) administering CRTs, (d) performing statistical analyses of CRTs, and (e) using the CRT results for program evaluation. First, writing criterion-referenced items takes time and practice. Before actually writing the items, teachers need to conduct a needs analysis in order to obtain sufficient information about the context they are dealing with such as the students' proficiency and motivation levels. The teachers then need to develop class objectives derived from the goals of the program. Based on the existing information, teachers need to think about the best test designs and develop a test specification that can be used as a guide for the CRT development and as a basis for the validity arguments used to defend their diagnostic and achievement decisions. Finally, actual item writing takes place by following the content of the test specification. Implementing this process skillfully takes practice because teachers need an in-depth understanding of their students and the context of the curriculum. With teaching experience and practice, they can write sound criterion-referenced items that are directly connected to the needs of their students, class objectives, materials, and teaching.

Second, because writing sound criterion-referenced items is difficult, it is necessary to pilot and edit the items several times. The data obtained from piloting have to be analyzed statistically in order to know how best to revise the items. This means that criterion-referenced item analyses such as item facility, the *B*-index, the

difference index, and distracter deficiency analysis need to be taught in teacher education courses. Teachers who neglect this procedure as part of teaching practice should recognize the importance of learning from the results of item analyses. Sometimes, the actual results differ significantly from the expected results; therefore, item analyses should be considered a good teaching practice that confirms the extent to which expectations and results match. Learning which items perform well based on statistical analyses can help improve teachers' testing literacy and item writing skills. Most importantly, these analyses are necessary to fine-tune items and to recycle well-performing items for future use.

Third, three features are unique to CRT administration. NRTs are administered when making proficiency and placement decisions, but CRTs are administered when making diagnostic and achievement decisions. Thus, diagnostic pretests have to be administered as part of teaching practice before actual teaching takes place. The pretest results can be a source of valuable information, helping identify misplaced students or mismatches in the objectives. For instance, if one student scores much higher than the other students, it might indicate that the student is misplaced. If a large proportion of students score above a certain cut-point, it is possible that inappropriately easy objectives have been selected. The pretest results can also be used for pedagogical purposes such as identifying students' strengths and weaknesses. Teaching should be focused on objectives that were not mastered by students at the beginning of the course to enable them to achieve a satisfactory score on a posttest.

In order to administer diagnostic pretests, CRTs have to be developed before actual teaching occurs. This enables teachers to be aware of what will be tested and leads to the implementation of successful teaching-to-the-test instruction that is designed to maximize students' achievement. In addition, teachers can tell students

what and how they will be tested on posttests, which can motivate students to study for the final examination. This leads to positive washback effects.

Achievement posttests should not be administered just as a formality but as one of the sources of information to make achievement decisions concerning students' mastery of the objectives set for the class. CRTs might not be as high-stake as NRTs, but teachers still need to consider several indicators of their students' mastery carefully because these decisions can have consequences on their future learning. To make valid decisions, it is best that two CRT forms be administered as a pretest and posttest in a counterbalanced design so that a wide range of the materials taught in class and more items can be given, items can be revised based on the results obtained from the pretests, and score gains can be calculated by comparing pretest and posttest results. In contrast, with one test form, only a narrow range of the materials can be tested. By developing more items, teachers can keep sound items and recycle them. In addition, teachers can add good items to an item bank.

Fourth, statistics performed in this study were discussed. Teachers should use item facility and item discrimination statistics for norm-referenced purposes, and the *B*-index and difference index for criterion-referenced purposes. Teachers should know how they are calculated because they have to be able to calculate these statistics in order to revise their tests and evaluate their teaching practice. Teachers should also learn how to interpret and calculate descriptive statistics, reliability estimates, and dependability indices in order to understand overall test results.

Generalizability theory provides a great source of information on CRTs for language testers. The G study shows the degree to which each facet contributes to the total score variance, and affects students' test performance. In this study, the largest amount of variance was due to the interaction, followed by the items effects. A D

study provides generalizability coefficients for absolute decisions for a number of hypothetical test designs. This is helpful in redesigning a test. In this study, the dependability indices showed that the test design used in the main study containing 25-20-25 items was optimal to achieve a certain magnitude of dependability.

When conducting a G study, the design has to be carefully planned. In most research, Cronbach alpha coefficients or generalizability coefficients with one-facet designs are reported. However, these coefficients should be interpreted cautiously because they overestimate dependability and do not take account of other possible facets in testing. When appropriate, two or more facet designs should be adopted.

When CRTs have several subtests and separate constructs are measured, using multivariate generalizability theory is highly recommended. It is possible to apply univariate generalizability theory for the analysis by treating each subtest as a facet, but it is not recommended because it underestimates the dependability, does not provide covariance among the subtests, and does not indicate the dependability for each subtest. Thus, the advantages of using multivariate generalizability theory are that in addition to the results that can be obtained from univariate generalizability theory, it gives covariate coefficients among subtests and does not underestimate the dependability. Even in a situation where there is extremely skewed data, which is common in criterion-referenced testing, estimating attenuated correlation coefficients is possible because multivariate generalizability study does not require a normality assumption (Brenna, 2000).

When applying Rasch measurement theory in a testing situation where multiple facets in the design can affect students' test performance, the use of MFRM is recommended even for a paper-and-pencil discrete test. For instance, subtests and occasions can influence students' test performance. In most cases, the dichotomous

Rasch model is used, but MFRM should be applied in order to take into account the facets.

MFRM can be used as a surrogate for ANOVA when ANOVA-like FACETS analysis is carried out. ANOVA reports statistical significance among means in raw scores. However, an ANOVA-like FACETS analysis provides not only statistical significance for the separation of the estimates but also converts the raw score means to an interval measurement scale.

The analyses used in this study, such as generalizability studies with different designs and the FACETS analyses, are different but complementary, and provide different viewpoints. Generalizability theory provides general, group-level information that is helpful in making overall decisions about test design. FACETS analyses provide more specific information that can be used for test development and improvement (Lynch & McNamara, 1998; Sudweeks, Reeve & Bradshaw, 2004). When different analyses are conducted and they show the same tendency, researchers can be more certain about the results. Thus, it is important to conduct different analyses so that the results can be triangulated.

Fifth, the results obtained from diagnostic pretests and achievement posttests can be a useful source of information for program evaluation. In order to examine program effectiveness, students' diagnostic pretest scores should be compared to their achievement posttests scores. An observation of the score gains can serve as empirical support showing that learning has taken place. If no score gains are observed for certain objectives, teachers should reconsider their students' needs, class objectives, CRTs, materials, and teaching to enable students' effective learning. If unified CRTs are administered to the entire student body, administrators can examine the score

gains to evaluate the soundness of the goals, CRTs, materials, and teaching to best satisfy the needs of the program.

Future Suggestions

This study was limited in scope and not all the issues surrounding criterion-referenced testing could be covered. Thus, directions for future research are pointed out in the following 12 questions to help establish criterion-referenced testing in foreign language education programs.

1. Would similar results be obtained if the CRT was given to other groups of students in the same program?
2. Would similar results be obtained if other teachers taught the class or other teaching methods were used to teach the class?
3. Would similar results be obtained when the class was taught in other educational settings where the curriculum is not coherently coordinated?
4. Would similar results be obtained with the CRT forms given to higher or lower proficiency students?
5. What results would be obtained if a longer time were given for students to study for the final examination?
6. Would results be similar if even more weight were put on students' achievement scores to decide their final grade?
7. To what extent is the CRT dependable and valid if other alternative forms are administered?
8. What are the percentages of variance when complex designs are conducted by adding a number of facets?

9. Are the CRT forms dependable and valid if different test designs or validation procedures are adopted?
10. What is the washback effect of this CRT administration on the students' learning?
11. What individual difference factors best predict students' achievement on this CRT?
12. Would the CRT be reliable and valid when used for norm-referenced purposes?

Final Conclusions

The primary reason that CRTs are not used more widely is that teachers and administrators are not trained to develop them properly and, as a result, lack testing literacy. This study demonstrated how CRTs were systematically developed, analyzed, and validated within the context of a university curriculum. It is suggested that teachers and administrators develop an in-depth understanding of CRT know-how and systematically develop their curriculum for the purpose of maximizing their students' learning. It is also hoped that they will improve their testing literacy so that the misuse of tests will not occur because it is unlikely that misused tests achieve their intended testing purposes. Developing and using high quality criterion-referenced tests will greatly benefit both teachers and students and will play an important role in the creation of effective educational curricula.

REFERENCES

Adams, R. & Khoo, S. (1996). *Quest* (version 2.1) [computer software]. Camberqell, Victoria: Australian Council for Educational Research.

Andrich, D. (1978). A rating formulation for ordered response categories. *Pychometrica*, *43*, 561-574.

Alderson, C. (2000). *Assessing reading*. Cambridge: Cambridge University.

Assessment Systems Corporation. (1998). *ITEMAN* (version 3.6) [computer software]. St. Paul, MN: Assessment Systems.

Arbuckle, J. L. (2006). Amos (Version 7.0) [computer software]. Chicago: SmallWaters Corporation.

Bachman, L. F. (1989). The development and use of criterion-referenced tests of language ability in language program evaluation. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 242-258). Cambridge: Cambridge University.

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

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

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

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, *12*(2), 239-257.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*(1), 101-118.

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing, 16*(2), 131-162.

Berk, R. (Ed.). (1980). *Criterion-referenced measurement: The state of the art*. Baltimore, MD: Johns Hopkins University.

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second language research. *Language Learning, 32*, 245-258.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model*. Mahwah, NJ: Erlbaum.

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

Brennan, R. L. (2000). (Mis)Conceptions about generalizability theory. *Educational Measurement: Issues and Practice, 19*(1), 5-10.

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

Brennan, R. L. (2001b). *urGENOVA* (version 2.1) [computer software]. Iowa City, IA: The American College Testing Program.

Brennan, R. L. (2001c). *mGENOVA* (version 2.1) [computer software]. Iowa City, IA: The American College Testing Program.

Brennan, R. L., Gao, X., & Colton, D. (1995). Generalizability analyses 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 at Los Angeles.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23*(1), 65-83.

Brown, J. D. (1990a). Where do tests fit into language program? *JALT Journal, 12*(1), 121-140.

Brown, J. D. (1990b). Short-cut estimators of criterion-referenced test consistency. *Language Testing, 7*(1), 77-97.

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

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development*. Boston, MA: Heinle & Heinle.

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

Brown, J. D. (1999). Relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16*(2), 216-237.

Brown, J. D. (2001). Developing and revising criterion-referenced achievement tests for a textbook series. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development* (Technical Report #21, pp. 205-228). Honolulu: University of Hawai‘i, Second Language Teaching and Curriculum Center.

Brown, J. D. (2003). Criterion-referenced item analysis: The difference index and the *B*-index. *Shiken: JALT Testing and Evaluation SIG Newsletter, 7*(3), 13-17.

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

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*, 21-42

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

Brown, J. D., & Pennington, M. C. (1991). Developing effective evaluation system for language programs. In M. C. Pennington (Ed.), *Building better English language programs* (pp. 57-74). Washington, DC: NAFSA.

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

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

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

Choi, I. C., & Bachman, L. F. (1992). An investigation into the adequacy of three IRT models for data from two EFL reading tests. *Language Testing, 9*(1), 51-78.

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

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

Cronbach, L. J., Rajaratnam, N., & Gleser, G. (1963). Theory of generalizability: A liberalization of reliability theory. *The British Journal of Statistical Psychology, 16*(2), 137-163.

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

Cziko, G. (1982). Improving the psychometric, criterion-referenced, and practice qualities of integrative language tests. *TESOL Quarterly, 16*(3), 367-378.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. London: Yale University.

Embretson, S. E., & Hershberger, S. (Eds.). (1999). *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Erlbaum.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

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

Gates, S. (1995). Exploiting washback from standardized tests. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 101-106). Tokyo: Japan Association for Language Teaching.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519-521.

Genesee, F., & Upshur, J. (1996). *Classroom-based evaluation in second language education*. Cambridge: Cambridge University.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Part, CA: Sage.

Henning, G. (1984). Advantages of latent trait measurement in language testing. *Language Testing, 1*(2), 123-133.

Henning, G. (1987). *A guide to language testing: Development, evaluation, and research.* Boston, MA: Heinle & Heinle.

Henning, G. (1988). The influence of test and sample dimensionality on latent trait person ability and item difficulty calibrations. *Language Testing, 5*(1), 83-99.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing, 9*(1), 1-11.

Hudson, T. (1991). Relationships among IRT item discrimination and item fit indices in criterion-referenced language testing. *Language Testing, 8*(2), 160-181.

Hudson, T. (1993). Surrogate indices for item information functions in criterion-referenced language testing. *Language Testing, 10*(2), 171-191.

Hudson, T., & Lynch, B. (1984). A criterion-referenced measurement approach to ESL achievement testing. *Language Testing, 1*(1), 170-201.

Kenyon, D. (1998). An investigation of the validity of task demands on performance-based tests of oral proficiency. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 19-40). Mahwah, NJ: Erlbaum.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing, 21*(1), 1-27.

Kumazawa, T. (2007a). bunsan bunseki to taso moderu o mochiita eigo toutatsudo tesuto no bunseki [Analyses of an English achievement test using ANOVA and many-faceted model]. *Kanto Gakuin University Bulletin, 18*, 17-27.

Kumazawa, T. (2007b). Criterion-referenced administration design and analyses. In T. Newfields, I. Gledall, P. Wanner & M. Kawate-Mierzejewska (Eds.), *Language acquisition-theory and pedagogy: Proceedings of the 6th Annual JALT Pan-SIG Conference* (pp. 65-74). Tokyo: JALT.

Kumazawa, T. (2009). Revision of a criterion-referenced vocabulary test using generalizability theory. *JALT Journal, 31*(1), 81-100.

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.

Lee, Y. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*(2), 131-166.

Lee, Y., & Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for new ESL writing test through G-theory. *International Journal of Testing, 7*(4), 353-385.

Li, D., & Brennan, R. (2007). *A multi-group generalizability analysis of a large-scale reading comprehension test* (Report No. 25). Iowa: Center for Advanced Studies in Measurement and Assessment.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA.

Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice, Volume 3* (pp. 85-98). Norwood, NJ: Ablex.

Linacre, J. M. (2002). *Facets* (Version 3.41) [computer software]. Chicago: MESA.

Lord, F. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing, 22*(4), 415-437.

Lynch, B. (2003). *Language assessment and programme evaluation*. Edinburgh: Edinburgh University.

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

Lynch, B., & Davidson, F. (1997). Criterion-referenced testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education, Volume 7: Language testing and assessment* (pp. 263-273). Dordrecht: Kluwer.

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

Matsuzawa, S. (2002). *Eigo kyoshi no tame no atarashii hyoukahou [A new assessment approach for English language teachers]*. Tokyo: Taishukan.

McNamara, T. (1990). Item Response Theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52-76.

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

Monsaas, J. A., & Engelhard, G. (1996). Examining changes in the home environment with the Rasch measurement model. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice, Volume 3* (pp. 127-140). Norwood, NJ: Ablex.

Negishi, M. (2005). Dou natta? kantenbetsu zettai hyouka [What happened? criterion-referenced testing]. *Eigo Kyoiku, 54*(4), 10-13.

Norris, J., Brown, J. D., Hudson, T., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*(4), 395-418.

Okubo, N. (2006). Shido to hyouka no ittaika wo mezashita shinraisei no takaiei sakubun hyouka kijyun hyou no sakusei: Tahenryo ippanka kanousei riron wo mochiite [Developing a reliable rubric for English composition to match assessment and teaching: Applying multivariate generalizability theory]. *STEP Bulletin, 18,* 14-29.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192.

Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*(1), 1-9.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago.

Sawaki, Y. (2005). The generalizability of summarization and free recall ratings in L2 reading assessment. *JLTA Journal, 7*, 21-44.

Sasaki, M. (2008). The 150-year history of English language assessment in Japanese education. *Language Testing, 25*(1), 63-83.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*(3), 355-390.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*(1), 1-30.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*(6), 922-932.

Stansfield, C. W., & Kenyon, D. M. (1992). Research of the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20*, 347-364.

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

Sudweeks, R., Reeve, S., & Bradshaw, W. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239-261.

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

Thompson, B. (Ed.). (2003). *Score reliability*. Thousand Oaks, CA: Sage.

Tyndall, B., & Kenyon, D. (1996). Validation of a new holistic rating scale using Rasch multi-faceted analysis. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 39-57). Clevedon: Multilingual Matters.

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16*(1), 82-111.

Wada, M. (2005). Mokuhyo kijyun jyunkyo tesuto no riron to jissen [Theory and practice of criterion-referenced testing]. *STEP Eigo Jyoho, 7-8*, 56-59.

Wakabayashi, S., & Negishi, M. (1993). *Musekinin na tesuto ga ochikobore o tsukuru* [Irresponsible tests make dropouts]. Tokyo: Taishukan.

Watanabe. Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13*(3) 318-333.

Webb, N. W., Shavelson, R. J., & Maddahian, E. (1983). Multivariate generalizability theory. In L. J. Fyans (Ed.), *Generalizability theory: Inferences and practical applications* (pp. 67-82). San Francisco: Jossey-Bass.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wolfe, E. (2000). Equating and item banking with the Rasch model. *Journal of Applied Measurement, 1*(4), 409-434.

Wright, B., & Masters, G. (1982). *Rating scale analysis*. Chicago: MESA.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing*, *24*(2), 251-286.

Yamamori, K. (2003). Chugakkou eigoka no kantenbetsu gakushu jyoukyou no hyouka ni okeru kanshin kanshin iyoku taido no hyouka no kentou [Investigating evaluation of students' interests, motivation, attitude in junior high school English course]. *Kyouikushinrigakukenkyu, 51,* 195-204.

Yamanishi, H. (2005). Using generalizability theory in the evaluation of L2 writing. *JALT Journal, 27*(2), 169-185.

Zhang, S. (2006). Investigating the relative effects of persons, items, sections, and languages on TOEIC score dependability. *Language Testing, 23*(3), 351-369.

APPENDIX A
READING AND LISTENING CRITERION-REFERENCED TEST (FORM A)

NAME: _____

Section 1：ボキャブラリー

Part 1:下線部の語句に最も近い意味を一つ選びなさい

1. Other stereotypes paint a picture of casual American students calling their professors by their first names.
    a. simple ideas
    b. musical concepts
    c. overall ideas
    d. strange images

2. Schools keep track of students' GPAs.
    a. evaluate
    b. control
    c. record
    d. delete

3. Also, a diploma from an American university doesn't guarantee graduates a job.
    a. experience
    b. recommendation letter
    c. graduate certificate
    d. special skill

4. Also, a diploma from an American university doesn't guarantee graduates a job.
    a. mean
    b. promise
    c. help
    d. get

5. Each system is a reflection of a different society, so it is hard to say which is better or worse.
    a. deep culture
    b. structure
    c. long history
    d. expression

6. Complaining loudly to a boy or girlfriend who arrives late for a date is a direct conflict resolution strategy.
    a. answer-solving
    b. expressive
    c. behavioral
    d. conceptual

7. Indirect communicators have the sense that by being silent, their partners will understand their feelings and adjust their behavior accordingly.
    a. decide
    b. change
    c. keep
    d. make

8. He has not been given time to find some way to apologize effectively.
    a. thank
    b. show feelings
    c. say sorry
    d. make a request

9. If this is the case, he may retreat, and become more silent.
    a. leave
    b. treat nicely
    c. repeat
    d. react

10. In English, the expression “say what you mean” reflects the assumption that words are necessary to solve conflict.
    a. understanding
    b. knowledge
    c. belief
    d. feeling

11. It weakens our body’s natural defenses against disease, making us more likely to develop colds and infections.
    a. illnesses caused by virus
    b. heat
    c. sadness
    d. bleeding

12. Because stress is physical, it’s not enough to simply distract ourselves with TV or comics.
    a. think about it
    b. make us think about it
    c. take our attention away
    d. keep our attention

13. We need an activity which engages our body and mind fully.
    a. endangers
    b. exercises
    c. encourages
    d. enables

14. Our struggle has become a fight for psychological survival as individuals.
    a. energy
    b. effort
    c. study
    d. stress

15. We are animals, it is true, but we are creative animals who can learn and adapt.
    a. change
    b. search
    c. understand
    d. create

16. It is the fourth most common ailment in the world.
    a. aim
    b. illness
    c. problem
    d. achievement

17. Everyone has emotional ups and downs or feels saddened or overwhelmed by the crises that occur in their lives.
    a. overacted by
    b. made strongly emotional by
    c. felt very angry by
    d. felt very happy by

18. Symptoms depend on the person, but depression is generally ongoing, without clearly marked causes.
    a. Unhealthy physical conditions
    b. Symbols
    c. Necessary cures
    d. Medicines

19. Symptoms depend on the person, but depression is generally ongoing, without clearly marked causes.
    a. labeled
    b. reasonable
    c. noticeable
    d. graded

20. Some people have reported remarkable cures.
    a. quick
    b. reactive
    c. amazing
    d. easy

21. Here are some tips.
    a. advice
    b. money
    c. chip
    d. note

22. The food is safe and delicious, even in stalls and in the marketplace.
    a. supermarkets
    b. stands
    c. discount stores
    d. flea markets

23. You may come across ordinary Balinese playing traditional music, or creating beautiful paintings or sculpture.
    a. statue
    b. design
    c. dance
    d. doll

24. Art is a big part of Balinese culture, and wandering on your own will give you the chance to see how it fits into the everyday life of Bali.
    a. wondering
    b. walking around
    c. shopping
    d. traveling around

25. Travel in the spirit of openness and you will find something to keep long after your tan has faded.
    a. darkened
    b. lost its color
    c. been hurting
    d. been taken care of

Section 2：リスニング

Part 2: Getting Personal Chapter 1 のビデオをみて次の問題に答えなさい

26. How was the speaker's experience in high school?
    a. He had to take many tests.
    b. He had to go to prison.
    c. He had few friends.
    d. He was tired of high school.

27. Why did he take a special test?
    a. He failed a class.
    b. He wanted to get good grades in a course.
    c. He wanted to leave high school.
    d. He wanted to get into a good university.

28. How was the speaker's experience in college?
    a. He was treated as equal as his teachers.
    b. He was treated equally with the rest of the students.
    c. He did not go to classes.
    d. He had many boring teachers.

29. What made him study hard in college?
    a. There were fewer rules.
    b. He was treated as an adult.
    c. He had a lot of free time.
    d. He wanted to be a teacher.

30. What is the main purpose of the passage?
    a. To tell how he learned to grow up in high school
    b. To tell how he learned to grow up in college
    c. To tell why his high school life was better than his college life
    d. To tell how his college life differed from his high school life

Part 3: Getting Personal Chapter 3 のビデオをみて次の問題に答えなさい

31. What is NOT important in friendship for the speaker?
    a. To be trustworthy
    b. To live far apart
    c. To feel conformable with each other
    d. To be silly together

32. Why did she mention Stonehenge?
    a. To tell her trip to Stonehenge was the best part
    b. To tell how her friend and she became the best friends
    c. To tell about one of the most famous places in England
    d. To tell why her friend and she met in London

33. Why did she buy two bottles of French brandy?
    a. To get drunk with her friend
    b. To make her friend fall asleep
    c. To see the light reflection of the bottles
    d. To keep the bottles for decoration

34. How did she and her friend spend a night in a hotel?
    a. She took care of her sick friend.
    b. She went to a pub to drink brandy.
    c. They continuously laughed hard.
    d. They continuously drank together.

35. What is the main message of the passage?
    a. To tell her about a trip in southern England
    b. To tell how to make friends
    c. To tell how she and her friend spent a night on a trip
    d. To tell her opinion about what friendship is

Part 4: Getting Personal Chapter 6 のビデオをみて次の問題に答えなさい

36. Why did Leandra become a counseling psychologist?
    a. She was good at communicating with people.
    b. She was interested in the study.
    c. She just wanted to do something for people.
    d. She wanted to be a detective.

37. Why do people need counseling?
    a. To talk about a normal problem
    b. To talk about their desire to kill someone with a gun
    c. To talk about a variety of problems
    d. To talk about their desire to kill themselves

38. Why is it that counseling is like detective work?
    a. Counselors have to be in dangerous situations sometimes.
    b. Counselors have to get a lot of experiences to deal with a serious problem.
    c. Counselors have to work with policemen.
    d. Counselors have to ask many questions to find people's actual problems.

39. What was the real problem that the family had?
    a. The child had no friend.
    b. The parents had a lot of complaints about their son's school.
    c. The child did not want to go to school.
    d. The parents did not get along well.

40. What is the main message of the passage?
    a. Life is full of problems.
    b. Life is depressing.
    c. Life is fun.
    d. Life is full of unique experiences.

Part 5: Getting Personal Chapter 7 のビデオをみて次の問題に答えなさい

41. What made Scott's life change?
    a. A phone call
    b. A letter
    c. A chance meeting with somebody
    d. A small advertisement in a newspaper

42. Why did Scott go to Korea for the first time?
    a. To go sightseeing
    b. To go to school
    c. To work
    d. To see his friend

43. Why did Scott go to Korea for the second time?
    a. To go sightseeing
    b. To go to school
    c. To work
    d. To see his friend

44. What does he like about Seoul?
    a. People are nice.
    b. Foods are delicious and cheap.
    c. American culture is naturally built into the culture.
    d. Traditional and modern cultures exist together.

45. What is the purpose of his talk?
    a. To tell what Seoul is like.
    b. To tell how nice Seoul is to live in.
    c. To tell how his life has changed.
    d. To tell how to shop in Seoul.

Section 3: リーディング

Part 6: Chapter 1 の内容を読み以下の問題に答えなさい

University life is an unavoidable step into adulthood. In “Passages: Predictable Crises of Adult Life,” Gail Sheehy describes this stage as "pulling up roots," the first of five "passages" we face throughout adult life. For 18 to 20-year-olds, it's a time of breaking away from family and departing to be on their own, to test their competency as independent people.

All five of the passages mark important changes in the challenges we face as adults. After “Pulling up Roots! come "The Turning Twenties” and “Catch Thirty.” Middle-age is “The Deadline Decade,” fol1owed by the “Comeback Decade,” and each of these is an important part of growth that we experience in adult life.

While all of these stages are important, the first step, moving beyond childhood and becoming an adult, is perhaps the stage most crucial. An important mark of this transition is a shift in how personal worth is measured. In childhood, growth tends to be measured as an external product that can be easily measured as test scores or acceptance into a good university. The results are shown clearly for everyone to see.

Although this continues in the form of finding a good job or getting a promotion, increasingly adulthood focuses on personal growth as an internal measure that is more difficult to determine. It is a journey and a process, and is reflected not in the obvious signposts along the road, but the passage itself. Our personal growth is measured in the quality of our travels through life: the attitudes and feelings we have, our courage, our openness to new experiences, our ability to love others.

University life offers students the opportunity to shift focus from the external growth of childhood to the internal growth of adulthood. It offers young people the opportunity to learn about the world and to accept new challenges. It also offers them the opportunity to learn about themselves, to think about the fundamental question—"Who am I?" It is possible to learn this through academic studies, the study of foreign languages, as well as talking with professors and friends—meeting new people and new challenges.

In order to grow, to travel new roads, people need to have a willingness to take risks and to accept the possibility that they may fail at first. For those that learn to be independent and confront the unknown, the other passages of adult life are smoother. Each step takes them farther along in their personal journey.

46. What do you have to do to grow?
    a. Travel around the world
    b. Meet many people
    c. Take chances
    d. Make many mistakes

47. How is personal growth measured in childhood?
    a. The number of challenges one takes
    b. The number of friends one has
    c. The university one goes to
    d. The job one has

48. How is personal growth measured in adulthood?
    a. The quality of experience
    b. Test scores
    c. The university one goes to
    d. The job one has

49. What does university life offer?
    a. Opportunities to get a good job
    b. Opportunities to study foreign culture and languages
    c. Opportunities to learn about the world and oneself
    d. Opportunities to get good grades and take on new challenges

50. What is the best title for this passage?
    a. Study Hard in Your University Life
    b. Enjoy Your University Life
    c. Go to Many Foreign Countries in Your University Life
    d. Grow Up Internally in Your University Life

Part 7: Chapter 3 の内容を読み以下の問題に答えなさい

Friends are special. Although we meet and talk with many people every day, we usually consider only a few of them true friends. Whether we realize it or not, making friends is a complicated process.

There are four basic qualities that influence who we choose to be our friends. The first quality is proximity—the geographic closeness of the person. If a person lives far away and we can't meet them regularly, we are less likely to form or maintain a relationship with that person. Interestingly, if we are around a person often, we are more likely to overlook things that may cause conflict. The opposite is also true; the farther a person is from us, the less forgiving of differences we tend to be.

Similarity is also important for friendship. Although we often hear the phrase "opposites attract," this doesn't seem to be the case. People usually look for a person with similar interests. Whether people actually have the same interests is not as important as the perceived similarity. This is called the balance theory. People want their beliefs, values, and feelings to fit with others around them. When the balance does not exist, they try to create it.

Attractiveness also counts both physically and socially. People with high status are often considered attractive. Though we might not like to admit it, humans are social animals who care about our position within our social groups. This often influences how we make friends.

The fourth quality is that of complementary needs. We like friends with similar interests, but not too similar. Instead we tend to look for people who complement us. For example, a person who is quiet at parties can be a perfect companion to someone who is outgoing and social.

Most people will tell you they choose friends simply based on personality. There are, however, many unconscious factors which cause us to find some people more or less attractive. As we understand these factors more, we not only can find better friends, but can have more to offer to the people we are close to.

51. If we are close to a person, what are we more likely to do?
    a. We don’t make a strong relationship.
    b. We don’t like to make a difference.
    c. We don’t try to see our differences.
    d. We try to make a difference.

52. Which is NOT a basic quality that influences who we choose to be friends with?
    a. Proximity
    b. Oppositeness
    c. Attractiveness
    d. Complementary needs

53. Which theory best explains the following case?
    “I have a good friend. He has something that I don’t have.”
    a. Balance theory
    b. Proximity theory
    c. Complementary needs theory
    d. Similarity theory

54. Why is making friends a complicated process?
    a. Because all people are special
    b. Because we can only make a few true friends
    c. Because we have very different kinds of personality
    d. Because there are many factors that are involved in making friends

55. What is the main purpose of the passage?
    a. To explain factors in making good friends
    b. To explain how special friends are
    c. To explain how to get along with friends
    d. To explain what makes people attractive

Part 8: Chapter 5 の内容を読み以下の問題に答えなさい

You have deadlines. Sometimes you argue with your friends. Maybe you're frustrated with the packed train you ride every morning. Perhaps you've fallen in love, and then had your heart broken. At times like this you feel tension in your shoulders, tightness in your stomach. You can't concentrate. You are under stress.

In the information age, we tend to forget that humans are animals suited to surviving in nature, hunting, or defending our home from outside threats. Physically, though, we haven't changed. We still have the same stress reaction as humans did in the past. Our bodies prepare to fight or run away. These days, though, we force ourselves to control our feelings and actions, and stress builds up.

Stress is not simply in our minds. It damages us physically by weakening our natural defenses against disease, making us more likely to develop colds and infections. It also causes heart disease, and it can even cause us to gain weight. Under ongoing stress, the body releases fats and sugar for the muscles to use for energy. If they don't get used up, fat gets stored near our liver, giving us a larger waist. If we are not careful, stress can shorten our life.

Awareness is the first step in fighting stress. In particular, young people may not recognize the stress caused by new types of relationships, new responsibilities and the challenges of moving into adult society. Paying attention to one's body helps us notice stress. Then we can take steps to release it by exercising or using relaxation techniques. Because stress is physical, it's often not enough to simply distract ourselves with TV or comics. We need an activity which engages our body and mind fully.

There is, however, an important difference between the challenges and stress we face today, compared to those in our more animal past. The battles we face in our everyday life are not truly threatening. Food, shelter, and community—our basic needs for survival are assured. Our struggle has become instead a fight for psychological survival as individuals. Fortunately, this takes place inside of us, giving us the chance to take control of our own stresses.

We are animals, it's true, but we are creative animals that can learn and adapt As a species, humans have become successful using our ability to make decisions and find new solutions to old problems. As individuals, our task is to learn to use this talent to deal with the stresses we face everyday, on the train, with our friends―and in our own minds.

56. What is one thing we haven't changed as compared with humans in the past?
    a. Our mind
    b. We forget things.
    c. We survive in nature.
    d. Our stress reaction

57. Which is NOT a true statement?
    a. Stress causes heart problems.
    b. Stress causes us to be fat.
    c. Stress causes us to gain muscles.
    d. Stress can shorten our life.

58. What is necessary in fighting stress?
    a. We have to take responsibilities.
    b. We have to make good relationships.
    c. We have to pay attention to TV.
    d. We have to exercise and relax.

59. What is the biggest difference between stress now and in the past?
    a. We have more physical stress.
    b. We have more stress to keep us alive.
    c. We have more stress in our mind.
    d. We experience more stress to be part of a community.

60. What is the main message of this passage?
    a. We have to have a better understanding of stress.
    b. We have to learn to use our ability to deal with stress.
    c. We have to use our mind to deal with stress.
    d. We have to understand that we are an animal.

Part 9: Chapter 7 の内容を読み以下の問題に答えなさい

"If you have a horse, send it to Cheju Island; if you have a son, send him to Seoul." So an old Korean saying advises. Send a horse to Cheju Island where the grass is green and lush. Send a son to Seoul to go to school: the city has thousands of students attending its 18 universities and 15 colleges. Send him to Seoul to get a job: all the head offices of any enterprise commercial, financial or educational are in Seoul. Send him to Seoul for a chance at the best opportunities.

That's what Seoul is: the center of the nation, the heart of the country with the best of everything. Seoul is like a giant magnet attracting business, education, culture, government-- all the occupations of any nation. Many people both sons and daughters---have responded to Seoul's pull, and now the city is home to one quarter of the country's population. Its residents now number over 10 million.

If you are downtown in Seoul at 6 or 7 p.m., you will feel yourself caught up in this rush-hour city. Seoul is the central city and has been the capital for 5,000 years (the word Seoul means capital). And if Seoul is the center of the nation, then where is its center-the soul of Seoul?

In terms of altitude, the center of Seoul is Nam-san (South Mountain), from which you can look across and see Pugak-san (North Peak Mountain). Between these two mountains is the old walled city. There used to be 9 gates which led into the ancient capital. Of those, 5 still remain. The Namdae-mun (Great South Gate) and Tongdae-mun (Great East Gate) are elegant reminders of the past, sitting in the middle of a busy city.

Some think of the center of the city as City Hall Plaza, the fountain square near the City Hall and the Plaza Hotel. From this intersection traffic runs into and out of the city. Running under the plaza are two subway lines as well.

There are other central landmarks as well, each one important for understanding Seoul. Still, no matter where we say the center is, no single place can express the total feeling of this city. You need to explore, look down from the mountains and out from the plazas. Seoul will reward you with all the charms that bring everyone here.

61. Why do people go to Seoul?
    a. To see horse racing
    b. To graduate from a good university
    c. To earn money
    d. To get the best chances

62. Where is the geographic center of Seoul?
    a. City Hall Plaza
    b. Pugan-san
    c. Nam-san
    d. The city hall

63. Where is the most crowded place in Seoul?
    a. The fountain squire near the Plaza Hotel
    b. Near Namdae-mun
    c. Near Tongdae-mun
    d. The intersection near the city

64. What do you have to do to fully understand Seoul?
    a. To understand the soul of Seoul
    b. To live there for a long time
    c. To look up at the mountains from the plazas
    d. To travel around the city and visit many places

65. Why did the author talk about the center of Seoul?
    a. To express the heart of Koreans
    b. To show what attracts people there
    c. To show there is no center
    d. To show many places are the center

Part 10: Chapter 8 の内容を読み以下の問題に答えなさい

Bali is like a gentle drop of clear water---a jewel on a magnificent necklace called Indonesia. You've probably seen pictures of Bali in the travel pamphlets that people read on crowded trains in Tokyo.

The packaged vacations offered in these pamphlets are carefully planned; the hotel is modern, you are led on tours, and you are even told where to sit for the scheduled cultural performances. Even your souvenir shopping is pre-planned. If you choose this kind of trip, you will get a tan and a few memories, but both will quickly fade. For something more, try going to Bali as an independent traveler. Here are some tips.

First, base yourself in the town of Ubud, rather than the tourist capital of Kuta, and travel to other places from there. The Ubud area is the cultural center of Bali. It has tourist conveniences, but is also a place where Balinese live ordinary lives.

Also, don't worry about getting sick, having your money stolen, communication problems, getting lost, etc. Bali is an easy-going, safe place. The food is safe and delicious, even in stalls, and in the marketplace. Your money will not be stolen, and a little English will take you everywhere.

Avoid the tourist spots and tour groups because they attract Vendors and taxi drivers. Instead, walk five minutes in the opposite direction. Get lost and open yourself to the chance for adventure and new discoveries. Stop at a fruit stand and eat something you've never tried before. If you're not sure how, the vendor will show you.

Notice the little things. For example, Besakih is the most important temple on Bali, and is a beautiful popular place. However, to enter the inner part, foreigners must be accompanied by a guide. Many

guides compete for business, and the, crowds can make Besakih tiring and annoying. Instead, walk through the outer grounds where anyone is allowed. Pathways lead through the nearby neighborhood. In the courtyards you may notice a group of women preparing food and offerings. This is the true spirit of Besakih.

Rent a bicycle. It's cheap and gives you the chance to meet people in smaller villages. You may come across ordinary Balinese playing1 traditional music, or creating beautiful paintings or sculpture.

Art is a big part of Balinese culture, and wandering on your own will give you the chance to see how it fits into the everyday life of Bali. Finally, keep a journal of your impressions or bring a sketch book and stop to draw what you see. You will remember the moment because you stopped and noticed. Travel in the spirit of openness and you will find something to keep long after your tan has faded.

66. Why does the author recommend not to go to Bali on a packaged tour?
    a. Because the tours are too well-planned
    b. Because the tourists visit too many places
    c. Because the tourists cannot fully experience the culture there
    d. Because the tourists can easily lose their tan

67. What is the first tip to travel in Bali?
    a. To stay in the tourist capital
    b. To stay where tourists stay and Balinese live
    c. To stay in Kuta and travel to other places
    d. To stay in Ubud and experience Balinese ordinary lives

68. What is the third tip to travel in Bali?
    a. Not to worry much
    b. To get lost and be adventurous
    c. To go to tourist spots
    d. To take taxis to go in the opposite direction

69. What is the true spirit of Besakih?
    a. What is in the temple
    b. Crowds going in to see the beauty of the temple
    c. What is outside of the temple
    d. Balinese religious practices such as offerings

70. How can we discover the attraction of Bali?
    a. Finding information in the pamphlets
    b. Going to many places
    c. Having an open mind to be with the culture
    d. Writing a journal to better remember the places

APPENDIX B
READING AND LISTENING CRITERION-REFERENCED TEST (FORM B)

NAME: _____

Section 1:ボキャブラリー

Part 1:下線部の語句に最も近い意味を選びなさい

1. In "Passages: Predictable Crises of Adult Life," Gail Sheehy describes this stage as "pulling up roots," the first of five "passages" we face throughout adult life.
    a. cries
    b. turning points
    c. pleasure
    d. steps

2. An important transition is a shift in how personal worth is measured.
    a. change
    b. chance
    c. challenge
    d. charm

3. The first step is perhaps the stage most crucial.
    a. hard
    b. basic
    c. central
    d. challenging

4. It is a time of breaking away from family and departing to be on their own to test their competency as an independent people.
    a. courage
    b. ability
    c. communication skill
    d. personality

5. Middle age is "The Deadline Decade," followed by the "Comeback Decade," and each of these is an important part of growth that we experience in adult life.
    a. age you die
    b. busy
    c. line between being alive and dead
    d. due date

6. The first quality is proximity.
    a. short distance
    b. estimate
    c. kindness
    d. close friendship

7. If we are around a person often, we are more likely to overlook things that may cause conflict.
    a. overestimate
    b. examine
    c. ignore
    d. keep

8. The fourth quality is that of complementary needs.
    a. different
    b. supplementary
    c. similar
    d. thankful

9. People with high status are often considered attractive.
    a. social position
    b. statue
    c. salary
    d. appearance

10. A person who is quiet at parties can be a perfect companion to someone who is outgoing and social.
    a. entertainer
    b. server
    c. partner
    d. enemy

11. It is the fourth most common ailment in the world.
    a. aim
    b. illness
    c. condition
    d. achievement

12. Everyone has emotional ups and downs or feels saddened or overwhelmed by the crises that occur in their lives.
    a. overacted by
    b. get strongly emotional by
    c. felt very angry by
    d. felt very happy by

13. Symptoms depend on the person, but depression is generally ongoing, without clearly marked causes.
    a. unhealthy physical conditions
    b. symbols
    c. necessary cures
    d. medicines

14. Symptoms depend on the person, but depression is generally ongoing, without clearly marked causes.
    a. labeled
    b. reasonable
    c. noticeable
    d. graded

15. Some people have reported remarkable cures.
    a. quick
    b. reactive
    c. amazing
    d. easy

16. Send a horse to Cheju Island where the grass is green and lush.
    a. poor
    b. rich
    c. cleanly shaped
    d. conditioned

17. Send him to Seoul to get a job: all the head offices of any enterprise.
    a. business
    b. states
    c. professions
    d. entries

18. In terms of altitude, the center of Seoul is Nam-san.
    a. beauty
    b. history
    c. height
    d. weight

19. The Namdae-num and Tondae-mun are elegant reminders of the past.
    a. remembrance
    b. reformed buildings
    c. reconsiderations
    d. remakes

20. There are other central landmarks as well, each one important for understanding Seoul.
    a. towers
    b. monuments
    c. paces
    d. markets

21. Here are some tips.
    a. advice
    b. money
    c. chip
    d. note

22. The food is safe and delicious, even in stalls and in the marketplace.
    a. supermarkets
    b. stands
    c. discount stores
    d. flea markets

23. You may come across ordinary Balinese playing traditional music, or creating beautiful paintings or sculpture.
    a. statues
    b. designs
    c. dance
    d. dolls

24. Art is a big part of Balinese culture, and wandering on your own will give you the chance to see how it fits into the everyday life of Bali.
    a. wondering
    b. walking around
    c. shopping
    d. traveling around

25. Travel in the spirit of openness and you will find something to keep long after your tan has faded.
    a. darkened
    b. lost its color
    c. been hurting
    d. been taken care of

Section 2：リスニング

Part 2: Getting Personal Chapter 2 のビデオをみて次の問題に答えなさい

26. How was Adrian's experience last year in high school?
    a. He was physically tired of studying.
    b. He was sick of studying.
    c. He studied hard to go to university.
    d. He studied many kinds of subjects.

27. Why did he take a year off to go to university?
    a. He wanted to earn money.
    b. He wanted to have fun.
    c. He wanted to expand his horizon.
    d. He wanted to study hard.

28. Why did he work in a bank?
    a. He wanted to be independent.
    b. He wanted to learn to get along with people.
    c. He eventually wanted to be a banker.
    d. He wanted to save money.

29. What was the biggest achievement on his trip to India?
    a. He learned to cook various kinds of curry.
    b. He met various kinds of people.
    c. He climbed high mountains.
    d. He traveled various parts of the country.

30. What is the main message of the passage?
    a. Be adventurous.
    b. Be creative.
    c. Be studious.
    d. Be enthusiastic.

Part 3: Getting Personal Chapter 4 のビデオをみて次の問題に答えなさい

31. Why does Ian look a little worried?
    a. Scott wants to borrow his money.
    b. His friend hasn't paid back his money.
    c. His friend wouldn't lend him money.
    d. Scott refuses to pay back his money.

32. How did Ian give some hints to his friends?
    a. He brought up a key word in a conversation.
    b. He used some gestures.
    c. He gave some examples.
    d. He dropped some money on a table.

33. Why doesn't Ian tell his friend directly?
    a. He doesn't want to make his friend feel bad.
    b. He doesn't want to lose a friend.
    c. He doesn't want to tell his friend what to do.
    d. He doesn't want to be rude.

34. What happened after Ian told his friend directly?
    a. He gave Ian a donut.
    b. He gave Ian some advice.
    c. He gave Ian the money.
    d. He gave Ian a reminder.

35. What is this conversation about?
    a. Friendship
    b. Trouble solving
    c. Personality
    d. Communication style

Part 4: Getting Personal Chapter 6 のビデオをみて次の問題に答えなさい

36. Why did Leandra become a counseling psychologist?
    a. She was good at communicating with people.
    b. She was interested in the study.
    c. She just wanted to do something for people.
    d. She wanted to be a detective.

37. Why do people need counseling?
    a. To talk about a normal problem
    b. To talk about their desire to kill someone with a gun
    c. To talk about a variety of problems
    d. To talk about their desire to kill themselves

38. Why is it that counseling is like detective work?
    a. Counselors have to be in dangerous situations sometimes.
    b. Counselors have to get a lot of experiences to deal with a serious problem.
    c. Counselors have to work with policemen.
    d. Counselors have to ask many questions to find people's actual problems.

39. What was the real problem that the family had?
    a. The child had no friend.
    b. The parents had a lot of complaints about their son's school.
    c. The child did not want to go to school.
    d. The parents did not get along well.

40. What is the main message of the passage?
    a. Life is full of problems.
    b. Life is depressing.
    c. Life is fun.
    d. Life is full of unique experiences.

Part 5: Getting Personal Chapter 8 のビデオをみて次の問題に答えなさい

41. What is Jegog?
    a. Balinese traditional dance
    b. Balinese traditional music
    c. Balinese traditional dance and music
    d. Balinese traditional instrument

42. Why did the speaker go to a village?
    a. To see the Jegog instruments
    b. To see the Jegog musicians practice
    c. To see the traditional dance
    d. To see the Jegog concert

43. How did she get the musicians to play?
    a. Paying money to all the musicians
    b. Asking the musicians nicely to practice
    c. Paying money to the coordinator of the musicians
    d. Shaking hands together happily

44. How was the concert?
    a. Like a classical music concert
    b. Like a village summer festival
    c. Like a music performance for the foreign travelers
    d. Like a music practice for the musicians

45. Why did Mono Arta teach the instrument to the speaker?
    a. To make the villagers laugh
    b. To surprise the villagers
    c. To entertain the speaker
    d. To celebrate the day

Section 3: リーディング

Part 6: Chapter 2 の内容を読み以下の問題に答えなさい

The university entrance exam system in Japan is famous as a "test hell." This can create the impression in Japan that American universities are easy to get into. Other stereotypes paint a picture of casual American students calling their professors by their first name. These professors are thought to teach by debating with their students in a lively give and take.

In fact, things are not so simple. It's true that in the United States, if you don't care what school you go to, and if you have enough money, any high school graduate can be admitted to a college. But this is true in Japan as well. However, as in Japan, American students and their parents are interested in getting into the top schools. High-level American universities have extremely competitive entry requirements.

High school grades are important. Schools keep track of students' GPA -grade point average. Students receive four points for an "A," three for a "B," etc. Their average overall grade is calculated and used as a way for universities to evaluate applicants. This means that students who didn't have good grades at the beginning of high school-in 9th or 10th grade- sometimes have trouble entering a good university.

The image of casual American students is also misleading. Students are casual in dress and speech, but still work hard. Also, a diploma (a degree) from an American university doesn't guarantee graduates a job. There's a lot of pressure for students to do well at school, work outside, and also develop skills to sell to future employers. There's pressure and opportunity to succeed, but also the real possibility for failure.

The image of teachers involved in a give and take with their students may be true in graduate school, but not generally for undergraduates. Freshman and sophomore classes are often very large-sometimes 200-300 students. This makes a one-to-one relationship with teachers nearly impossible. Many teachers do call their students by first names, but usually not the other way around. In large universities especially, professors are busy, and often have little time for students.

The reality of university life in Japan is more complicated than the image of "entrance exam hell." In the same way, the reality of university life in the United States is not as simple as the stereotypes that many people have. Each system is a reflection of a different society, so it's hard to say which is better or worse. This doesn't mean, of course, that there isn't room for improvement in both.

46. What do American students have to do to get into a good university?
    a. Enter a high-level high school.
    b. Get a high score on the entrance exam.
    c. Get good grades in all the classes in high school.
    d. Debate with teachers actively in class.

47. Why is the casual image of American students wrong?
    a. Most of them are under constant pressure to do well.
    b. Most of them don’t dress and talk casually.
    c. Most of them enjoy their lives.
    d. Most of them don’t have to experience failure.

48. Why is the image of American teachers wrong?
    a. Most teachers like to discuss ideas with their students.
    b. Most teachers don’t have time to spend with their students.
    c. Most teachers are called by their first names.
    d. Most teachers are busy with their students.

49. Choose a statement that is true about the passage.
    a. American university students can easily graduate from university.
    b. American high school students hardly ever study.
    c. Almost all high school students can get into college in Japan.
    d. American university students hardly ever study.

50. What is the main purpose of the passage?
    a. Arguing which educational system is better.
    b. Arguing which university is more difficult to enter, Japanese or American.
    c. Arguing that stereotypes about American education can be wrong.
    d. Arguing that American and Japanese educational systems need to make progress.

Part 7: Chapter 3 の内容を読み以下の問題に答えなさい

Friends are special. Although we meet and talk with many people every day, we usually consider only a few of them true friends. Whether we realize it or not, making friends is a complicated process.

There are four basic qualities that influence who we choose to be our friends. The first quality is proximity—the geographic closeness of the person. If a person lives far away and we can't meet them regularly, we are less likely to form or maintain a relationship with that person. Interestingly, if we are around a person often, we are more likely to overlook things that may cause conflict. The opposite is also true; the farther a person is from us, the less forgiving of differences we tend to be.

Similarity is also important for friendship. Although we often hear the phrase "opposites attract," this doesn't seem to be the case. People usually look for a person with similar interests. Whether people actually have the same interests is not as important as the perceived similarity. This is called the balance theory. People want their beliefs, values, and feelings to fit with others around them. When the balance does not exist, they try to create it.

Attractiveness also counts both physically and socially. People with high status are often considered attractive. Though we might not like to admit it, humans are social animals who care about our position within our social groups. This often influences how we make friends.

The fourth quality is that of complementary needs. We like friends with similar interests, but not too similar. Instead we tend to look for people who complement us. For example, a person who is quiet at parties can be a perfect companion to someone who is outgoing and social.

Most people will tell you they choose friends simply based on personality. There are, however, many unconscious factors which cause us to find some people more or less attractive. As we understand these factors more, we not only can find better friends, but can have more to offer to the people we are close to.

51. If we are close to a person, what are we more likely to do?
    a. We don’t make a strong relationship.
    b. We don’t like to make a difference.
    c. We don’t try to see our differences.
    d. We try to make a difference.

52. Which is NOT a basic quality that influences who we choose to be friends with?
    a. Proximity
    b. Oppositeness
    c. Attractiveness
    d. Complementary needs

53. Which theory best explains the following case?
    “I have a good friend. He has something that I don’t have.”
    a. Balance theory
    b. Proximity theory
    c. Complementary needs theory
    d. Similarity theory

54. Why is making friends a complicated process?
    a. Because all people are special
    b. Because we can only make a few true friends
    c. Because we have very different kinds of personality
    d. Because there are many factors that are involved in making friends

55. What is the main purpose of the passage?
    a. To explain factors in making good friends
    b. To explain how special friends are
    c. To explain how to get along with friends
    d. To explain what makes people attractive

Part 8: Chapter 4 の内容を読み以下の問題に答えなさい

Falling in love is easy. The real test of a romance, though, is not when things are going well, but when we argue and fight. If we can work out our differences, the relationship deepens, but if not, we run the risk of building up frustrations until the whole relationship is in danger.

Different people, and different cultures, have different strategies for dealing with conflict. One key is the degree of directness during communication. Complaining loudly to a boy or girlfriend who arrives late for a date is a direct conflict resolution strategy. These direct communicators feel that if they can't express their frustration openly, putting their ideas into words, then the other person will not understand them and the problem will only get worse.

Being quiet when one is upset is also a way to communicate feelings. Indirect communicators have the sense that by being silent, their partner will understand their feelings and adjust their behavior accordingly. Too much talk is thought to get in the way. When two indirect communicators have conflict they give each other space to figure out the best way to change the situation over time.

Problems start when partners' communication styles don't match. If the indirect-style boyfriend is late, the direct-style girlfriend may complain openly. He then may feel his space is being invaded-that he's not been given time to find some way to apologize effectively. If this is the case, he may retreat, and become more silent. Unfortunately, she may then feel her feelings are being ignored, and complain even more. This vicious cycle can continue, with both people feeling that the other person is being unreasonable.

These differences are not just a reflection of our personalities. We learn communication strategies from our family and friends as we grow up - from our culture. Japanese is said to emphasize indirect communication strategies. Expressions like "ichi wo kiite, juu wo shiru" (hear one, understand ten) reflect this. In English, the expression "say what you mean " reflects the assumption that words are necessary to solve conflict.

Relationships don't end simply because we disagree or argue. Conflict is a natural part of human interaction, and when it is managed well we learn to appreciate others. No one is always right, and only conflict can teach us our limitations. Conflict can make us grow. Understanding our own communication style is a first step in this process. It may not be as much fun as falling in love, but it is a skill which will last us a lifetime.

56. How can we deepen our love with a partner?
    a. By falling in love
    b. By understanding our differences
    c. By testing our relationship
    d. By building up frustrations

57. What do direct communicators do to deal with conflict?
    a. Being quiet
    b. Understand others' feelings
    c. Complain
    d. Ignore others

58. What is NOT the cause of the difference in our communication style?
    a. Our personalities
    b. Our culture
    c. Conflicts
    d. People surrounding us

59. What is conflict according to the passage?
    a. It leads to our frustration.
    b. It is a learning opportunity.
    c. It can lead to the end of relationship.
    d. It teaches us which person is right or wrong.

60. What is a good title for this passage?
    a. Avoid Conflict
    b. Understand Communication Styles
    c. Directness in Communication Styles
    d. Falling in Love

Part 9: Chapter 6 の内容を読み以下の問題に答えなさい

Depression is more common than most people realize. It's the fourth most common ailment in the world, and number two in Western countries, behind heart disease. In spite of this, scientists still don't fully understand it, and talking about serious depression remains taboo for many people.

Some things are clear. As a medical condition, depression is different than simply feeling "down" or "blue." Everyone has emotional ups and downs, or feels saddened or overwhelmed by the crises that occur in our lives. Depression is different. Symptoms depend on the person, but depression is generally ongoing, without clearly marked causes. It involves an inability to feel a usual range of emotions, both positive and negative. Someone who is depressed loses interest in the world, and may shut themselves away from others.

Until recently, serious depression was thought to be caused entirely by an imbalance of chemicals in the brain called "neurotransmitters." Researchers now understand brain chemistry better, and have been able to develop many new drugs to treat depression. Some people have reported remarkable cures. Despite these successes, researchers have started to realize that there is more to the problem than simply mixed-up chemicals.

Scientists are finding differences in how the brains of depressed people function. For example, the area of the brain just behind the forehead, the pre-frontal cortex (PFC), is critical in determining our emotional lives. The left side of the PFC is associated with positive feelings, while the right side is associated with negative ones. It appears that the left side of the PFC in the brains of depressed people is less active than normal.

It also appears that trauma and stress can affect this part of the brain. These brain functions are much more stable in adults than in children. One key to understanding depression appears to lie in childhood. Some research suggests that if children experience emotional trauma and stress before they reach their teenage years, their brains change in a way which could make depression more likely in later life.

This insight of course, is not far from the common idea in psychology that childhood experiences shape our mental health. Scientists are starting to find some physical support for this idea. Hopefully, this increased understanding will also help us develop better treatments.

There's another possible benefit to all of this research. Science is taking some of the mystery out of a problem that many people are uncomfortable talking about We need to be open to this new understanding, to better control our mental health, and make sure anyone suffering from depression gets whatever help they need.

61. What is NOT depression?
    a. It is a taboo that people don't talk about.
    b. It is an ailment that scientists don't know much about.
    c. It is feeling "down".
    d. It is one of the most common ailments.

62. What did researchers think of the cause of depression in the past?
    a. Feeling sad
    b. Taking many drugs
    c. Losing interest in the world
    d. Losing chemical balance in the brain

63. What did researchers find out about depressed people?
    a. They have many crises in their lives.
    b. They are very emotional.
    c. Some part of their brain is not active.
    d. They are very critical about human rights.

64. Choose a person who likely gets depressed in later life.
    a. A stressful student
    b. A person with brain damage
    c. A person who experienced child abuse
    d. A physically unhealthy child

65. What is the purpose of the passage?
    a. To make us understand what depression is
    b. To explain that children who experience a trauma are likely to get depression
    c. To tell research findings on depression
    d. To show how to control our mental health

Part 10: Chapter 8 の内容を読み以下の問題に答えなさい

Bali is like a gentle drop of clear water---a jewel on a magnificent necklace called Indonesia. You've probably seen pictures of Bali in the travel pamphlets that people read on crowded trains in Tokyo.

The packaged vacations offered in these pamphlets are carefully planned; the hotel is modern, you are led on tours, and you are even told where to sit for the scheduled cultural performances. Even your souvenir shopping is pre-planned. If you choose this kind of trip, you will get a tan and a few 10 memories, but both will quickly fade. For something more, try going to Bali as an independent traveler. Here are some tips.

First, base yourself in the town of Ubud, rather than the tourist capital of Kuta, and travel to other places from there. The Ubud area is the cultural center of Bali. It has tourist conveniences, but is also a place where Balinese live ordinary lives.

Also, don't worry about getting sick, having your money stolen, communication problems, getting lost, etc. Bali is an easy-going, safe place. The food is safe and delicious, even in stalls, and in the 20 marketplace. Your money will not be stolen, and a little English will take you everywhere.

Avoid the tourist spots and tour groups because they attract Vendors and taxi drivers. Instead, walk five minutes in the opposite direction. Get lost and open yourself to the chance for adventure and new discoveries. Stop at a fruit stand and eat something you've never tried before. If you're not sure how, the vendor will show you.

Notice the little things. For example, Besakih is the most important temple on Bali, and is a beautiful popular place. However, to enter the inner part, foreigners must be accompanied by a guide. Many guides compete for business, and the, crowds can make Besakih tiring and annoying. Instead, walk through the outer grounds where anyone is allowed. Pathways lead through the nearby neighborhood. In the courtyards you may notice a group of women preparing food and offerings. This is the true spirit of Besakih.

Rent a bicycle. It's cheap and gives you the chance to meet people in smaller villages. You may come across ordinary Balinese playing1 traditional music, or creating beautiful paintings or sculpture.

Art is a big part of Balinese culture, and wandering on your own will give you the chance to see how it fits into the everyday life of Bali. Finally, keep a journal of your impressions or bring a sketch book and stop to draw what you see. You will remember the moment because you stopped and noticed. Travel in the spirit of openness and you will find something to keep long after your tan has faded.

66. Why does the author recommend not to go to Bali on a packaged tour?
    a. Because the tours are too well-planned
    b. Because the tourists can visit too many places
    c. Because the tourists cannot fully experience the culture there
    d. Because the tourists can easily lose their tan

67. What is the first tip to travel in Bali?
    a. To stay in the tourist capital
    b. To stay where tourists stay and Balinese live
    c. To stay in Ubud and travel to other places
    d. To stay in Ubud and experience Balinese ordinary lives

68. What is the third tip to travel Bali?
    a. Not to worry much
    b. To get lost and be adventurous
    c. To go to tourist spots
    d. To take taxis to go in the opposite direction

69. What is the true spirit of Besakih?
    a. What is in the temple
    b. Crowds going in to see the beauty of the temple
    c. What is outside of the temple
    d. Balinese religious practices such as offerings

70. How can we discover the attraction of Bali?
    a. Finding information in the pamphlets
    b. Going to many places
    c. Having an open mind to be with the culture
    d. Writing a journal to better remember the places

APPENDIX C
PROCTOR MANUAL (JAPANESE VERSION)

今回 R & L レギュラー用テストの実施にご協力いただきどうも有り難うございます。
簡潔にテストの実施過程に関して説明させていただきます。

1. 第一回目の授業で以下の点を学生に伝達してください。
   - 第二回目の授業 90 分を使って練習テストを行う。
   - テスト結果は成績には関係ない。
   - 実施目的としては
     1. 自分のレベルを把握する。
     2. 期末テストではどのような内容が出題されるのかを知る。しかし、期末テストでは違う問題が出題される。
     3. 練習テスト結果は研究目的で用いられるが、個人情報は厳守される。
2. 一週間前には第二回目の授業で実施していただきますテストのコピーと解答用紙を人数分、メールボックスに入れておきますので事前にご確認ください。
3. 第二回目の授業で練習テストとして実施してください。
   - 事前にご用意していただくものは以下の通りです。
     1. リスニング・セクションで使います Information, Please の DVD
     2. メールボックスに入れますテストと解答用紙などテスト一式
   - 学生を着席させ、携帯のスイッチの確認、筆記用具の準備など練習試験を受験する準備させてください。
   - 練習テストの主旨と手順を説明してください。
     1. 実施目的としては
        - 自分のレベルを把握する。
        - 期末テストではどのような内容が出題されるのかを知る。しかし、期末テストでは違う問題が出題されます。
     2. 試験時間は 90 分とする。
     3. テスト結果は成績には関係ない。
     4. 試験開始 15 分後リスニングを開始する。
   - 解答用紙を配布し、次に問題用紙を配布してください。
   - はじめの合図からおよそ 15 分経過したら、セクション 2 のリスニングを開始してください。まず、Part 2 にある指定の Chapter の内容を DVD で流してください。その Chapter が終わりましたら 2 分間問題を解く時間を与えてください。そして、2 分経過したら次の Part 3 にある指定の Chapter の内容を DVD で流してください。
   - チャイムが鳴ったらテスト問題を全部回収してください。そして解答用紙を回収してください。＊問題が漏れないためテスト問題は回収してください。
4. テスト実施後、テスト問題は破棄してください。解答用紙は用意いたします封筒に入れ、お手数ですが、12 号館講師控え室内私のメールボックスへ入れてください。
5. 私のほうでテストを採点し、結果を集計後、フロッピー・ディスクにてテスト結果をお渡しいたします。2 週間ほどで集計ができると思います。
6. 期末試験 2 週間前には以下の点を学生に伝達してください。
   - 期末テストでは第 2 回目に実施した練習テストと同じような問題を出題する。
   - テスト結果は全体の成績の 40％となる。
   - テスト範囲は Chapter 8 までとする。
   - 習熟度テストなので、授業内容を学習すればよい点をとれるであろう。
7. 期末テスト実施日 1 週間前にはテスト問題と解答用紙を先生方のメールボックスへ入れ

ておきますので事前にご確認ください。

8. 最終授業で授業内にて期末テストを実施してください。
   - 事前にご用意していただくものは以下の通りです。
     1. リスニング・セクションで使います Information, Please の DVD
     2. メールボックスに入れますテストと解答用紙などテスト一式
   - 学生を着席させ、携帯のスイッチの確認、筆記用具の準備など練習試験を受験する準備させてください。
   - 期末テストの主旨と手順を説明してください。
     1. 実施目的としては
        - 学生の習熟度を測定する。
        - 成績の 40%の判断材料として用いる。
        - 練習テスト結果は研究目的で用いられるが、個人情報は厳守される。
     2. 試験時間は 90 分とする。
     3. 試験開始 15 分後リスニングを開始する。
   - 解答用紙を配布し、次に問題用紙を配布してください。
   - はじめの合図からおよそ 15 分経過したら、セクション 2 のリスニングを開始してください。まず、Part 1 にある指定の Chapter の内容を DVD で流してください。その Chapter が終わりましたら 2 分間問題を解く時間を与えてください。そして、2 分経過したら次の Part 2 にある指定の Chapter の内容を DVD で流してください。
   - チャイムが鳴ったらテスト問題を全部回収してください。そして解答用紙を回収してください。＊問題が漏れないためテスト問題は回収してください。
9. テスト実施後、テスト問題は破棄してください。解答用紙は用意いたします封筒に入れ、お手数ですが、12 号館講師控え室内私のメールボックスへ入れていただければと思います。
10. 私のほうでテストを採点し、結果を集計後、フロッピー・ディスクでテスト結果をお渡しいたします。2 週間ほどで集計ができると思います。

ご質問やお気づきの点がありましたら何なりとメールや電話をいただければと思います。
何かとお手数をお掛けいたします。何卒、よろしくお願い致します。

APPENDIX D
PROCTOR MANUAL (ENGLISH VERSION)

Thank you very much for your help in administering the Reading and Listening regular test. I would like to briefly explain the procedure for the test administration.

1. Please announce the followings in the first lesson.
    - A 90-minute practice test will be given in the second lesson.
    - Students' test scores will not be graded.
    - The purpose of this test administration is:
        1. to diagnose students' mastery and non-mastery of the materials.
        2. to let students know what will be tested on the final examination, but the exact same test will not be given.
        3. to collect data for research but information such as the students' number and scores will be kept confidential.
2. One week before the administration day, I will put the question sheets and answer sheets in your mail box. Please check to see if there are enough sheets for all of your students.
3. In the second lesson, please administer the test following the procedure.
    - Beforehand, you need to take:
        1. a DVD for the assigned textbook, Information Please. That will be used in the listening section of the test.
        2. the question and answer sheets that I have prepared and put in your mail box.
    - Have students be seated. Have them turn off their cell phones. Have them take out pencils and erasers.
    - Explain the purpose of the test and procedure.
        1. The purposes are:
            - to diagnose students' mastery and non-mastery of the materials.
            - to let students know what will be tested on the final examination, but the exact same test will not be given.
        2. Students can take 90 minutes to complete the test.
        3. Students' test scores will not be graded.
        4. The listening section will start 15 minutes after the test has started.
    - Distribute the answer sheets, and then question sheets.
    - Fifteen minutes after the test has started, start the listening section. First, play the DVD to show the assigned chapter in part 2. After the chapter is

finished, give them two minutes to answer the corresponding questions. Then, go on to the next part.

- Ninety minutes after the test has started, collect the test sheets and answer sheets. Be sure to count the number of the test sheets and answers sheets to see if you collected them from all the students.

4. After the test administration, throw away the question sheets. Put the answer sheets in an envelope that I have put in your mail box. Put the envelope in my mail box.
5. I will score your students' test and return their scores on EXCEL saved on a floppy disk two weeks after I get your students' answer sheets
6. Two weeks before the final examination, please announce the following to your students:
    - The content of the final examination will be almost the same as the one given on the second week.
    - Your test score on the final examination will account 40% of your final grade.
    - The test content covers unit 1 to unit 8.
    - This is an achievement test designed to test your mastery of the materials. Study the materials.
7. I will put the question and answer sheets in your mail box a week before the day of the final examination. Please be sure to check if there are enough sheets for all your students.
8. In the last lesson, please administer the test following the procedure.
    - Beforehand, you need to take:
        1. a DVD for the assigned textbook, Information Please. That will be used in the listening section of the test.
        2. the question and answer sheets that I have prepared and put in your mail box.
    - Have students be seated. Have them turn off their cell phones. Have them take out pencils and erasers.
    - Explain the purpose of the test and procedure.
        1. The purposes are:
            - to estimate students' mastery and non-mastery of the materials.
            - to use students' test score to decide 40% of their final grade.
            - to collect students' test score as data for research, but students' private information will be securely kept.

2. Students can take 90 minutes to complete the test.
3. The listening section will start 15 minutes after the test has started.

- Distribute the answer sheets, and then the question sheets.
- Fifteen minutes after the test has started, start the listening section. First, play the DVD to show the assigned chapter in part 2. After the chapter is finished, give them two minutes to answer the corresponding questions. Then, go on to the next part.
- Ninety minutes after the test has started, collect the question sheets and answer sheets. Be sure to count the number of the question sheets and answers sheets to see if you collected them from all the students.

9. After the test administration, throw away the question sheets. Put the answer sheets in an envelope that I have put in your mailbox. Put the envelope in my mailbox.
10. I will score your students' test and return you their scores on EXEL spreadsheets saved on a floppy disk in two weeks after I get your students' answer sheets.

If you have further questions, please do call or email me. Thank you very much for your cooperation in advance.

APPENDIX E
READING AND LISTENING CLASS SYLLABUS

Meeting time: Friday 2nd period
Instructor: Takaaki Kumazawa 熊沢　孝昭
E-mail: takaakikumazawa@hotmail.com
Office hours: Tuesday & Friday lunch periods

Course description:
This textbook-based course is designed to provide ample opportunities for students to improve listening and reading skills. Vocabulary development is also an important part of this course. The other aspect of this course is that in the reading session you will practice strategies that help you understand the passages.

Specific goals:
1. You will be able to improve your reading skills.
2. You will be able to improve your listening skills.
3. You will be able to understand my instruction in English.
4. You will be able to learn strategies that help you understand the passages such as guessing, phrase reading, focusing on stressed words, inference, skimming, scanning, identifying topic sentences, and paragraph reading.
5. You will be able to increase the number of vocabulary items.

Course requirements:
1. 80% attendance
2. In-class exams at the beginning and the end of this course
3. Turn in activity sheets

Grading:
30% attendance
* Considered late after I finish taking the attendance.
* Three late comings are considered to be one absence.

40% Final Examination
* All the questions on the test are based on the content of the textbook. Study the textbook!!!
* Two tests are given at the beginning (pre-test) and end (final exam) of the class. The first exam is NOT graded but your test score is part of your participation grade.

30% class participation

* Turn in all the activity sheets (Do not lose them)

* If you get a high score on the tests given at the beginning and end, you will get extra participation grade.

Textbook:

Information, Please. Shohakusha. ¥2952

Tentative schedule:

1. 4/13 Syllabus
2. 4/20 Pretest
3. 4/27 Chapter 1
4. 5/11 Chapter 2
5. 5/18 Chapter 3
6. 5/25 Chapter 4
7. 6/1 Chapter 5
8. 6/8 Chapter 7
9. 6/15 Chapter 8
10. 6/22 Chapter 9
11. 6/29 Chapter 10
12. 7/6 Final Exam

APPENDIX F
SAMPLE ACTIVITY SHEET

R & L Regular Lesson Plan
NAME:_____

1. Attendance
2. Warm up
   (a) Form a group of 4 and work together to find out the meanings. Write the meanings in the parenthesis.
   (b) The following words and phrases are used to describe either high school life or university life. Guess what belongs to which category. Write H or U next to each word or phrase.
   _____ prison ( )
   _____ want to study harder ( )
   _____ treated as equals ( )
   _____ no attendance ( )
   _____ bored ( )
   _____ felt trapped ( )

3. 1st Listening
   Watch the video, and answer the questions that the teacher reads, either A or B.
   1._____ 2._____ 3._____

4. 2nd Listening
   Watch the video again, and answer the questions that the teacher reads.
   1.

   2.

   3.

5. 3rd Listening: Fill in the blanks in the transcript.

APPENDIX G
HOLISTIC SCALE FOR PORTOFOLIO ASSESSMENT
(JAPANESE VERSION)

得点	内容
30	全授業回数分のプリントを提出しており、指定の記入個所を丁寧にかつ詳細に全部記入しており、授業に熱心に参加したことがうかがえる。
～	
26	全授業回数分のプリントを提出しており、指定の記入個所をほぼ全部記入しており、授業に参加したことがうかがえる。
25	数枚プリントが提出されていが、指定の記入個所を全部記入しており、よく授業に参加したことがうかがえる。
～	
21	数枚プリントを提出していなく、指定の記入個所を記入していない場合があるが、まあ授業に参加したことがうかがえる。
20	半分程度しかプリントを提出していなく、指定の記入個所も記入していない場合があるが、まあ授業に参加したことがうかがえる。
～	
16	半分程度しかプリントを提出していなく、指定の記入個所も記入していない場合があり、授業にあまり参加していないことがうかがえる。
15	多数プリントを提出していなく、指定の記入個所も記入していない場合があり、授業に参加していないことがうかがえる。
～	
11	多数プリントを提出していなく、指定の記入箇所も記入していない場合があり、授業にまったく参加していないことがうかがえる。
10	数枚しかプリントを提出していなく、指定の記入個所もあまり記入していなく、授業にまったく参加していないことがうかがえる。
～	
6	数枚しかプリントを提出していなく､指定の記入もまったく記入していなく、授業にまったく参加していないことがうかがえる。
5	プリントを提出していなく、授業も参加していなかった。
～	
0	プリントを提出していなく、授業態度も悪かった。

APPENDIX H
HOLISTIC SCALE FOR PORTOFOLIO ASSESSMENT (ENGLISH VERSION)

Score	Description
30 ~ 26	All the handouts are turned in and all the places where the students are supposed to fill in are written in detail, so the teacher can tell that he or she participated in lessons actively. All the handouts are turned in and almost all the places where the students are supposed to fill in are written, so the teacher can tell that he or she participated in lessons.
25 ~ 21	Some of the handouts are not turned in, but all the places where the students are supposed to fill in are written, so the teacher can tell that he or she participated in lessons. Some of the handouts are not turned in and some of the places where the students are supposed to fill in are not written, but the teacher can tell that he or she fairly participated in lessons.
20 ~ 16	About half of the handouts are not turned in and some of the places where the students are supposed to fill in are not written, but the teacher can tell that he or she fairly participated in lessons. About half of the handouts are not turned in and some of the places where the students are supposed to fill in are not written, so the teacher can tell that he or she did not actively participate in lessons.
15 ~ 11	Many of the handouts are not turned in and not many places where the students are supposed to fill in are written, so the teacher can tell that he or she did not actively participate in lessons. Many of the handouts are not turned in and not many places where the students are supposed to fill in are written, so the teacher can tell that he or she did not participate in lessons at all.
10 ~ 6	Only few handouts are turned in and not many places where the students are supposed to fill in are written, so the teacher can tell that he or she did not participate in the lessons at all. Only few handouts are turned in and places where the students are supposed to fill in are not written, so the teacher can tell that he or she did not participate in the lessons at all.
5 ~ 0	No handouts are turned in and the student did not participate in the lessons at all. No handouts are turned in and the student had a bad attitude toward learning.

APPENDIX I
TEST SPECIFICATION FOR READING AND LISTENING CRITERION-REFERENCED TEST

General Description

The R & L CRT is a criterion-referenced test that is designed to measure students' mastery of the class content of the Reading and Listening Regular and the cut-points are set at sectional means in the pretest. Two forms of the CRT are developed so that the tests can be given as a pretest and posttest. It is a 90 minute in-class examination given on the second and last days of the class.

The class is basically textbook-based in which students are given ample opportunities to improve listening and reading skills. Vocabulary development is also an important part of the class. The other aspect of this class is that in the reading session students practice strategies that help them understand the passages. Students' level is about intermediate.

The goals of the class are: (a) to improve students' reading comprehension, (b) to improve listening comprehension, (c) to understand my instruction in English, (d) to learn strategies that help them understand the passages such as guessing, phrase reading, focusing on stressed words, inference, skimming, scanning, identifying topic sentences, and paragraph reading, and (d) to increase the number of vocabulary items.

Except goals (c) and (d), the objectives that are to be tested are developed. The objectives are as follows: Students taking this R & L regular class will be able to (a) comprehend the 400-word passages in the textbook with the readability of about 70 in FRE and answer multiple-choice questions 60% correctly, (b) comprehend the 350 word dialogs spoken by native speakers of English with the readability of about 80 in FRE and answer multiple-choice questions 60% correctly, (c) comprehend the receptive meanings of the vocabulary words in the textbook that are categorized as frequent and academic word lists and answer multiple-choice items 60% correctly.

Test Design

Within each test form, there is a total of 70 items, composed of three sections including vocabulary, listening, and reading. The multiple-choice receptive item type is chosen because listening and reading are considered to be receptive skills. In the vocabulary section, students read a sentence with a target word underlined and then choose the word that is closest in meaning from four choices. There are 25 items. The listening section comprises four parts. Within each part, students watch a video and answer five questions based on the passage. In total, there are 20 questions. The

reading section has five parts. Within each part, students are required to read a passage that is taken exactly from the text; they are made to answer five questions that are supposed to measure their comprehension on the passages in the textbook. The vocabulary section contains five linking items. One part of the listening section appears in both the test forms, and the five items can be used as a linking item. Two parts of the reading section are inserted for linking the two test forms.

Sample Items

(a) Listening comprehension of the passages in the textbook
Choose the best answer for the questions
Why does Joseph like university life better than high school life?

(a) busier schedule
(b) more freedom
(c) more boring teachers
(d) less studying

(b) Receptive context-dependent vocabulary knowledge
Reading: choose the closest meaning of the underlined word
University life is an <u>unavoidable</u> step into adulthood.

(a) inescapable
(b) dangerous
(c) incapable
(d) fast

(C) Reading comprehension questions
What is the main message of the passage?

(a) life
(b) university life
(c) adulthood
(d) growing up

Prompt Attributes

1. All the dialogs, vocabulary, and reading passages are adopted from the assigned textbook.

2. Listening questions involve the following:
   - Understand specific information
   - Understand main ideas for each paragraph
   - Understand the overall idea of the dialogs

3. Vocabulary questions involve the following:
   - Understand the meaning of a sentence
   - Understand the receptive meaning of the target vocabulary words

4. Reading questions involve the following:
   - Understand specific information
   - Understand main ideas for each paragraph
   - Understand the overall idea of the passage

5. Each listening dialog is about 300-600 words in length. Native speakers of English talk rather slowly on the topics such as college life, friends, and conflict.

6. The vocabulary words tested on the CRTs are selected from frequent and academic word lists.

7. The reading passages are about 300-500 words in length. The readability ranges from 60-80. The topics include college life, friends, and conflict.

Response Attributes

1. The student listens to the dialogs and answers the items. The correct answer is: for details and for main ideas.
2. The student reads the sentences and selects the best choice that is close to the target word.
3. The student reads the passages and answers the items. The correct answer is: for details and for main ideas.

Test Analysis

The following analyses are utilized: (a) descriptive statistics, (b) item facility, (c) the *B*-index, and (d) the difference index.

APPENDIX J
A FACETS CONTROL FILE

```
title = Equivalence of Test forms
output = Equivalence of forms.txt
Facets = 4
gstat = yes
write = no
arrange = N
vertical = 1*,2*,3A

Models = ?,?,?,?,D

Positive = 1

LABELS=
1, persons
1-139 = Pre-test,,1
140-278 = Post-test,,2
*
2,Items
1-125
*
3,Occasions
1=Pretest
2=Posttest
*
4,subtests,A
1-6,,0
1=Vocabulary(pre)
2=Listening(pre)
3=Reading(pre)
4=Vocabulary(post)
5=Listening(post)
6=Reading(post)
*
Data=
1,1-45,1,1,0,0,1,1,1,1,1,1,0,1,1,1,0,1,1,0,1,1,1,0,0,1,1,1,0,,,,,,,,,,,,,,,,,,,
|
139,1-45,1,1,0,0,1,1,0,0,0,0,1,0,0,1,0,1,1,0,0,1,1,0,0,0,0,0,1,,,,,,,,,,,,,,,,,,,
140,1-45,2,4,,,,,,,,,,,,,,,,,,,,0,0,1,0,0,0,1,1,1,0,0,0,1,1,1,0,1,0,0,0,0,0,1,1,1
|
278,1-45,2,4,,,,,,,,,,,,,,,,,,,,1,0,1,1,1,0,1,0,1,0,1,1,1,1,1,1,0,1,0,1,0,0,0,0,1,1
1,46-80,1,2,0,0,0,0,0,0,0,1,0,1,1,1,1,0,1,1,1,1,1,1,,,,,,,,,,,,,
|
139,46-80,1,2,0,1,1,0,0,1,1,0,1,0,0,0,1,1,1,0,1,1,0,0,,,,,,,,,,,,,
140,46-80,2,5,,,,,,,,,,,,,,,,1,1,1,0,0,0,0,0,1,0,0,0,1,1,0,0,0,1,1,0
|
278,46-80,2,5,,,,,,,,,,,,,,,,1,0,1,1,0,1,1,0,0,0,1,1,0,0,1,1,1,1,1,1
1,81-125,1,3,0,1,1,1,1,1,1,1,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,,,,,,,,,,,,,,,,,,
|
139,81-125,1,3,1,1,1,1,1,0,1,1,1,1,0,1,1,0,1,1,1,0,0,0,1,1,0,0,1,,,,,,,,,,,,,,,,,,
|
278,81-125,2,6,,,,,,,,,,,,,,,,,,,,0,1,0,0,0,1,1,1,0,1,0,1,1,1,1,1,0,0,1,0,1,0,1,1,0
```