THE JAPANESE PLACEMENT TESTS

AT THE UNIVERSITY OF HAWAI'I:

APPLYING ITEM RESPONSE THEORY[1]

KIMI KONDO-BROWN

JAMES DEAN BROWN

University of Hawai'i at Manoa

ABSTRACT

First-rate placement procedures are important for effective teaching and learning in any language program because they help create classes that are relatively homogeneous in terms of the language proficiency of the students. The main purpose of this study was to investigate how effectively and efficiently the current norm-referenced Japanese Placement Test (JPT) battery for the Japanese language program (three multiple-choice tests and essay test) at UHM separates the incoming students of Japanese into different course levels. The *XCalibre*^{tm} computer software program (Assessment Systems, 1997) was used to estimate the discrimination, difficulty, and guessing parameters for each item on each multiple-choice test. Based on these IRT analyses, we were able to suggest that all three multiple-choice tests be reduced in length while maintaining the same, if not better, level of reliability. Additional analyses of the interrater reliability of the essay tests using the Spearman-Brown prophecy formula led us to suggest that the essay test might be made more efficient by using two raters instead of three. The pattern of correlation coefficients between the tests indicates a certain degree of convergent validity for all the tests in this study, especially the subscales within the essay test. At the same time, when factor analysis was applied, support for divergent validity was also found based on language skills and testing method.

INTRODUCTION

This study provides analyses of the Japanese Placement Tests for purposes of suggesting revisions in the tests themselves and the placement procedures associated with them. Naturally, this project does not exist in a vacuum; indeed, it was conducted within a framework of other previous research including work on: (a) language testing in general, (b) Japanese language testing in particular, and (c) item response theory. Thus the literature on those three topics will be discussed at the outset in order to properly situate the study.

The Literature on Language Testing in General

A great deal has been written about language testing over the last fifty years. Most notably a number of books have been published on the topic. These books have ranged from practical guides to help teachers write good classroom language tests (e.g., Harris, 1969; Heaton, 1977; Valette, 1977; Madsen, 1983; Underhill, 1987; and Carroll & Hall, 1985) to books aimed at helping language professionals develop tests for decision-making in language programs (e.g., Brown, 1996; and Alderson, Clapham, & Wall, 1995), to more theoretical tomes (like Lado, 1961; Henning, 1987; Bachman, 1990; and Bachman & Palmer, 1996). Indeed, a journal, the *Language Testing* journal has been dedicated to publishing research on the testing of languages for over 16 years and more recently the International Language Testing Association was founded with the purpose of promoting sound research and practice in language testing. Hence, we can say with some confidence that language testing is a well-established field.

The Literature on Japanese Language Testing

Specifically within the area of Japanese language testing, we note that a few books have appeared in the last decade about testing students of Japanese as a second or foreign language (e.g., Ishida, 1992; and Society for Teaching Japanese as a Foreign Language, 1991). However, we feel that, in these books, the development of Japanese language placement tests has not been adequately discussed. In fact, to our knowledge, the development of Japanese language placement tests for university students of Japanese as a second language (JSL) or Japanese as a foreign language (JFL) is an area that has received little attention in the literature. Notable exceptions are (a) the recent research efforts that investigated the potential for using the SPOT (Simple Performance Oriented Test) as a placement test for university students in Japan and the United States and (b) the development of a computer-adaptive placement test for university students of Japanese in Australia. These studies will be reviewed next.

The SPOT was developed by a group of researchers at the Tsukuba University. The SPOT is an indirect, integrative test designed to measure the overall Japanese language proficiency level of the candidates (Kobayashi, Ford, & Yamamoto, 1996). The SPOT consists of 60-65 contextually unrelated sentences; each sentence (written in Japanese scripts) has one blank, which the candidates are required to fill in with one *hiragana* letter. In each case, the *hiragana* letter is related to a particular grammatical item. The candidates fill in the blanks while listening to a taped reading of the sentences. Each sentence is read only once. The developers of the SPOT reported that correlation coefficients between SPOT and various other kinds of placement tests at the Tsukuba University ranged from moderate (.75 with the listening test, .69 with the

reading test, and .61 with the *kanji* test) to high (.81 with the grammar test) (Kobayashi & Ford, 1992, cited in Kobayashi, et al., 1996, p. 202).

Hatasa and Tohsaku (1997) administered the SPOT to students, who were enrolled in undergraduate Japanese language programs at the University of Iowa and the University of California at San Diego. In order to explore the potential for using the SPOT for placement decisions with university students of JFL, they examined (a) the reliability and discrimination indexes of items on the SPOT based on classical test theory, (b) correlation coefficients between the SPOT scores and the scores on the other placement tests used at Tsukuba University, (c) the rank-order correlation between the SPOT scores and instructors' evaluations of speaking abilities, and (d) correlation coefficients between the SPOT scores and students' achievement levels measured by their final course grades and final examination scores. Based on the results of these studies, they concluded that the "SPOT is highly effective as a placement test of Japanese in the US" (Hatasa & Tohsaku, 1997, p. 94). However, at the same time, they questioned the construct validity of the SPOT: they pointed out that it is not clear yet what exactly the SPOT measures.

Brown and Iwashita (1996) investigated the influence of learners' first languages (Chinese versus English) on their performances on a computer-adaptive grammar placement test of Japanese. They used a placement test developed by experienced teachers of Japanese at the University of Melbourne. With a large pool of participants in Australia and China (644 native speakers of English and 456 native speakers of Chinese), they analyzed how the candidates' first languages influenced their performances on the placement test. This issue of differential item functioning (see Sasaki, 1991) is particularly important in the development of computer-adaptive tests, because the candidates' abilities are calculated on the basis of fixed measures of item

difficulty. Brown and Iwashita found that the item difficulties were very different for candidates with English and Chinese language backgrounds. They suggest that in developing computer-adaptive placement tests, the test developers need to carefully consider the candidates' language background(s).

Another relevant study involved using placement test data for curriculum development at the university level. Tanaka (1995) at International Christian University conducted a correlational study examining the scores on one of the university's placement tests (a reading comprehension and grammar test) and the scores on a production test in order to restructure the curriculum for their summer courses in Japanese. Based on the results of her study, Tanaka made recommendations for how best to restructure their Japanese language summer program.

Item Response Theory

Classical test theory does not make any assumptions about how students' levels of ability affect the way they perform on a test: the only information for predicting an individual's performance on a given item is the index of difficulty, which is simply the percentage of the individuals in a group who responded correctly to a particular item. In other words, in classical test theory, the difficulty level of an item is assumed to be the same for every individual in a group regardless of the individual's level of ability. Also in classical test theory, the difficulty of each item changes if the average performance of the group of examinees changes.

In contrast, item response theory (IRT) models consider each individual's expected performance on a particular test item based on information about both the level of difficulty of

the item and the individual's level of ability (Hambleton, Swaminathan, & Rogers, 1991; Ohtomo, 1996; McNamara, 1996). Other advantages of the IRT approach are that:

1. The logit scale used to represent the student ability levels in an IRT model is "a true interval scale" (Henning, 1987, p. 129), unlike raw test scores in which the distances between intervals may not be equal (Brown, 1996, p. 97).
2. The fit validity statistic in IRT analyses indicates the total fit for a given item, which helps to detect unexpected answers. The higher the fit statistic, the less likely the item is to conform to the expected patterns of the model.
3. The deletion of misfitting items based on an IRT model results in better overall reliability estimates than does item deletion based on item statistics in classical test theory.
4. The standard error of measurement is estimated for every item, and this is more accurate than a single overall estimate of standard error of measurement given in classical test theory (Henning, 1987, pp. 129-130).
5. Once items are calibrated using an IRT model, fewer items are required to estimate student ability and student abilities can be determined without using all items. These characteristics of IRT item calibration make it ideal for item banking or test tailoring for computer adapted tests (Henning, 1987, p. 131; Brown, 1992, 1997).
6. IRT models have been found to be useful for both NRTs and CRTs (Rentz & Rentz, 1978, p. 5-7).

The one-, two-, and three-parameter models are the most frequently used IRT models for dichotomously scored test data. While IRT has been used in analyzing the items of university

language placement tests (e.g., Lozier & Chalhoub-Deville, 1995), to our knowledge, IRT has not been applied to the item analysis of a Japanese language placement test in a JFL context.

This study will use a three-parameter model, which typically requires a large sample size of at least 1,000 for the results to be accepted with confidence (Henning, 1987). Three-parameter models allow researchers to estimate difficulty, discrimination, and guessing parameters for each item.

The *difficulty parameter* of an item in a one- or two-parameter model is the point on the ability scale where there is a .50 probability of a correct response. In three-parameter models, instead of the difficulty parameter being that point on the ability scale where there is a .50 probability of a correct answer, it is that point on the ability scale where there is a (1+guessing parameter)/2 probability of a correct response (Ohtomo, 1991, p. 84; Harris, 1989, p. 35). For instance, for a four-option multiple-choice item with a guessing parameter of .22, the difficulty parameter would be that point on the ability scale where there is a .61 [(1 + .22)/2 = .61] probability of a correct response.

Positive difficulty logits indicate items of above average difficulty (the higher the value, the more difficult the item is), and negative logits indicate items of below average difficulty (the lower the value is, the easier the item is). Typically, the difficulty parameters range between –2.00 and +2.00, but some items may be considerably outside that range.

The *discrimination parameter* of an item is expressed in units that typically range from 0 to 2. When the discrimination parameter turns out to be a negative value, it is typically flagged by the IRT computer program and dropped as an item that does not fit the model. The higher the value of the discrimination parameter for a particular item, the better the item discriminates. For

example, an item that has a discrimination parameter of 1.96 discriminates much better than one with an item parameter of .16.

The *guessing parameter* (also sometimes called the *pseudochance* parameter) is expressed in probabilities. The guessing parameter accounts for the probability that students, even those of low ability, may correctly answer moderate or even high difficulty items. The lower the probability is, the less likely it is that students are correctly answering an item above their ability level. For instance, an item with a guessing parameter of .11 has a lower probability of lower ability students (i.e., of ability levels lower than the difficulty of the item) correctly answering it than an item with a guessing parameter of .24.

Purpose

First-rate placement procedures are extremely important for effective teaching and successful learning because they can help create classes that are relatively homogeneous in terms of the language proficiency levels of the students. The main purpose of the present study is to investigate how effectively and efficiently the current norm-referenced JPT battery for the Japanese language program at the University of Hawai‘i at Manoa (UHM) separates the incoming students of Japanese from diverse language abilities into different course levels. Another purpose of this study is to explore the potential of item response theory for analyzing the JPT battery for university students of Japanese. Based on three-parameter model item response theory analyses, recommendations will be made about ways the current JPT battery might be improved. To those ends, three central research questions were posed:

1. Based on the three-parameter item response theory model, which (and how many) multiple-choice items on the listening, grammar, and recognition tests are at the appropriate levels of discrimination, difficulty, and guessing?
2. For the analytic writing scale, how many raters are most efficient in terms of the trade off between number of raters and test reliability? In other words, how can the essay test also be made more efficient?
3. To what degree are the four tests valid in terms of their relationships with each other?

METHOD

Participants

The *multiple-choice data* for this study were gathered from all incoming students of Japanese at UHM in 1998 and 1999. During that time, 939 students took the listening section, 1294 took the grammar section, and 1124 took the recognition section.

Essay data were also collected from the essays that incoming students of Japanese wrote at the time they took the multiple-choice placement test. There were asked to describe: (a) how they like to spend their vacations, (b) themselves, their background, and their family, or (c) Hawaii to a friend in Japan. In 1998 and 1999, a total of 428 students wrote an essay: 78 chose prompt (a), 253 prompt (b), and 83 prompt (c). Among them, all essays written by heritage language students (i.e., students with at least one native Japanese speaking parent) were rated because the number of these students was much smaller than that for non-heritage language students and because a study of these students was planned for a later date. In order to make the sample sizes for each prompt equal, 78

(the number for the smallest prompt) were randomly selected from the compositions written on each of the other two prompts.

Table 1 shows the distribution in percentage terms of gender for each of the four tests. For instance, 58.7% of the students taking the listening test were female, while 35.7% were male, and 5.6% are missing (in the sense that they gave no information about gender) for a total of 100%.

[Insert Table 1 About Here]

Similarly, Table 2 shows the distribution in percentage terms of major for each of the four tests. Table 3 shows the percentages for academic status (or year in school) for the four tests. Table 4 gives the percentages for self-identified native language.

[Insert Tables 2 to 4 About Here]

Table 5 presents the percentages for the Japanese language background of the students' families. For instance, for 83.4% of the students on the listening test, no parent or grandparent was a native speaker of Japanese; for 4.4%, both parents were native speakers of Japanese; for 5.4%, only the mother was a native speaker of Japanese; for 1.9%, only the father was a native speaker of Japanese; and for 4.8%, at least one grandparent was a native speaker of Japanese. Only .1% were missing data on this variable.

[Insert Table 5 About Here]

Table 6 shows the distribution in percentage terms for number of years (rounded to the nearest integer) of high school Japanese taken by the students who took each of the four tests. The means (and standard deviations) were 3.12 (1.15), 2.80 (1.35), 2.95 (1.32), and 3.29 (1.27) for the listening, grammar, recognition, and essay tests, respectively.

[Insert Table 6 About Here]

Table 7 shows the percentages for number of years (rounded to the nearest integer) of other Japanese language programs (for instance, programs at Japanese language schools, community colleges, etc.) taken by the students who took each of the four tests. The means (and standard deviations) were 1.14 (2.23), 1.23 (2.28), 1.27 (2.32), and 1.47 (2.57) for the listening, grammar, recognition, and essay tests, respectively. These distributions appear to be positively skewed probably because two-thirds of the students attended no other Japanese language programs.

[Insert Table 7 About Here]

Table 8 shows the distribution in percentage terms for the number of years (rounded to the nearest integer) students had lived in Japan for each of the four tests. The means (and standard deviations) were .20 (1.30), .37 (1.85), .40 (1.96), and .37 (1.79) for the listening, grammar, recognition, and essay tests, respectively. Again, the distributions appear to be skewed because the vast majority of students had never lived in Japan.

[Insert Table 8 About Here]

Table 9 shows the ultimate percentages of placements into each of the Japanese courses for the four tests. Notice that about half to two-thirds of the students were placed into 100 or 100/102.

[Insert Table 9 About Here]

Materials

Incoming students of Japanese at UHM are required to take the multiple-choice parts of the JPT battery if they wish to be placed in a class other than the first-semester Japanese (JPN 101). The current JPT battery was developed and revised by faculty in the Japanese language program of the Department of East-Asian Languages at UHM. The JPT battery consists of three

multiple-choice tests (listening, grammar, and recognition) and an essay. The listening comprehension test has 14 items, the grammar test has 70, and the *kana* (a mixture of the Japanese *hiragana* syllable symbols and *katakana* syllable symbols used to spell out foreign words in Japanese) and *kanji* (Chinese characters) recognition test has 50.[2] In all three multiple-choice tests, the items all have four options (one correct answer and three distractors). We will now describe each of the three multiple-choice tests and the essay in more detail, as well as how all the tests were scored.

Listening test. The directions for the listening test are in English and the test is played from an audiotape. During the test, the students listen to five prompts in Japanese: four conversations (between a man and a woman) and one narrative (read by a woman). The prompts are recorded at a natural rate of speech, and the enunciation is very clear. Each prompt is read three times, and the students answer questions, which are written in English. The entire test takes approximately 10 minutes. An ample 45 seconds was allowed for each question, so this test is considered a power test rather than a speeded one.

Grammar test. The grammar test questions are written in both *kana* and *roomaji* (with which Japanese scripts are spelled out in Roman letters). Since students are taught in various Japanese language programs to use one or the other (or both) of these systems, the grammar test allows them to read and answer items in whichever way is most comfortable for them. The grammar items are designed to test the syntactic and morphological rules of the Japanese language (one

exception is a question that tests a lexical rule of *Keigo* [honorific Japanese]). No time limit was imposed. Hence, this test is considered a power test rather than a speeded one.

Recognition test. In the recognition test, ten questions are about *kana*-word spelling, and 40 questions ask students to identify the meanings of words written in *kanji*. As in the case of the grammar test, no time limit is imposed. Hence, this test is viewed as a power test rather than a speeded one.

Essay test. As mentioned above, in the *essay test*, students were asked to describe: (a) how you like to spend your vacation; (b) yourself, your background, and your family; or (c) Hawaii to a friend in Japan. The students wrote these essays at the same time they took the multiple-choice placement test. No time limit was imposed. Hence, this test is considered a power test rather than a speeded one.

Scoring. The multiple-choice tests were all scored by hand at the time of the actual placement decisions. For the purposes of the present project, the answers were entered into an *Excel*^{tm} spreadsheet program (Microsoft, 1999) and then scored within that program.

In addition, three raters scored each of the 234 essays in this study. All three raters were experienced Japanese language instructors at UHM. Using a modified version of the Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981) essay scoring grid, students' essays were scored in five categories: content, organization, vocabulary, language use, and mechanics (see Appendix A, or Kondo-Brown, K., 2000, p. 33). Each subscale ranged from 7 to 20 possible

points. The raters received training in the scoring procedures on the first day of scoring. Each rater scored the entire set of 234 essays under the same conditions over two days working together in a quiet academic setting. Each started with one prompt, then moved to the second prompt, and finished with the third.

Decision making based on the JPT battery. Students who choose to take the JPT are required to answer items in the grammar test, but those students who claim to have no recognition ability in *kana* and *kanji* are not required to take the recognition test. The listening test is given only to those students who take the JPT in a large group. The essay test is optional for students who wish to provide additional evidence of their proficiency. At present, placement decisions are usually based primarily on the students' performances on the grammar and recognition tests with the scores on the other tests being used for back-up information.

Analyses

The first step in analyzing the data for this project was to enter the data into an *Excel*^{tm} spreadsheet program (Microsoft, 1999). Each correct answer was coded as *1*, and each incorrect answer as *0*. We decided to interpret unanswered items as a failure to answer the item successfully, and therefore missing answers were coded as *0*. As mentioned earlier, since these were power tests (i.e., students were given as much as necessary to complete them), we felt it was unlikely that unanswered items were a result of time constraints.

The next step was to screen the data. There were no test *items* that everyone answered correctly, nor any that everyone answered incorrectly. However, in the listening section, there were 50

students who answered all questions correctly and one student who did not have any correct answers. In the grammar section, six students answered all questions correctly and four students had no correct answers. In the recognition section, four students answered all questions correctly. These students who either had perfect scores or zero scores were eliminated from the data pool because an IRT model cannot estimate such students' abilities. Put simply, for students who had perfect scores, we know their abilities are high, but we do not know how high; for students who had zero scores, we know their abilities are considerably below guessing and that does not fit the model.

The resulting numbers of participants were 888 for the listening test, 1284 for the grammar test, and 1120 for the recognition test. Thus, while the grammar and recognition tests had sufficiently large samples for the three-parameter model, the listening test had a sample that was somewhat smaller with 888 students than the recommended sample of at least 1,000. This means that the results of the listening test should be interpreted cautiously especially give the small number of items.

We used the *XCalibre*^{tm} computer program (Assessment Systems, 1997) to estimate the discrimination (usually symbolized by *a*), difficulty (symbolized by *b*), and guessing (symbolized by *c*) parameters for each item on each test. The *XCalibre*^{tm} program automatically flags problem items. Items are flagged if the item has: an *a* parameter that is too low (less than .30), a *b* parameter that is too low (less than –2.95) or too high (more than 2.95), a *c* parameter that is too high (more than .40), a possible keying error, or a standardized residual statistic is more than 2.0.

RESULTS

As is often the case in such studies, the analyses in this study produced an overabundance of results. As a consequence, we have had to be both organized and selective in our approach to them. In order to make that organization clear to readers, we will present our results in five sections, four focused on the results for the listening, grammar, recognition, and essay tests, and one comparing all four tests.

Results for the Listening Test

We began the IRT analyses by using the *XCalibre*^{tm} program to analyze the original 14 items on the listening test. The results are shown in Table 10. From the left, Table 10 shows one column each for item number, flagging, difficulty parameter, discrimination parameter, guessing parameter, and residual. The output indicated only one flagged item (item 2). In this case, the flagging indicated that the standardized residual statistic exceeded a value of 2.0. In other words, this item did not fit the model, and therefore, we eliminated it. The reliability of the original version of the listening test (k = 14) was .797, so we conclude that the listening test was 79.7% reliable (and by extension, 20.3% unreliable). The reliability of the revised version of the test (k = 13) was .801, which means that the revised version of the test is slightly shorter and more reliable (and therefore more efficient).

[Insert Table 10 About Here]

We ran the *XCalibre*^{tm} program again without item 2. Table 11 shows the results of the three-parameter IRT analysis of the remaining 13 items. From the left, it shows the original item number, difficulty parameter, discrimination parameter, guessing parameter, and residual. The difficulty parameters ranged from –1.95 to 2.51. The items are sorted from the item that had the lowest difficulty parameter to the item that had the highest difficulty parameter. In the original version of the test, test items were numbered from the easiest item to the most difficult item based on the test

developers' judgments. The results here indicate that the test developers' original difficulty estimates for certain items did not necessarily match the difficulty parameters we found for these items.

[Insert Table 11 About Here]

Figure 1 is a visual representation of the difficulty parameters for the 13 items on the revised version of the listening test. From the left, the items are arranged from the easiest items to the most difficult items. Figure 1 indicates that the difficulty parameters for about two-thirds of the items were above the average logit of zero. Figure 1 also indicates that about half of the items were within the difficulty range between .00 and 1.00, which suggests that additional items with a difficulty parameter of below .00 and above 1.00 need to be developed. The discrimination parameters ranged from .52 (item 7) to 1.60 (item 12). Ten items or most of the items had an *a* (discrimination) parameter of below 1.00, and only two items had an *a* parameter of above 1.00. These results suggest that more items with an *a* parameter of 1. 0 or above should be added. Guessing parameters seemed to be low for this four-option multiple-choice test ranging from .10 to .15 (even though, overall, the guessing parameters for the listening test were higher than those for the other two multiple-choice tests).

[Insert Figure 1 About Here]

Results for the Grammar Test

When we ran the *XCalibre*^{tm} program on the original grammar test (see Table 12), six items out of 70 were flagged, indicating that the standardized residual statistic exceeded a value of 2.00. These items were eliminated first. A second preliminary run of *XCalibre*^{tm} indicated that two more items

were flagged for the same reason, so we eliminated these two items as well. Then, in order to reduce the number of items to the 50 most efficient items, we removed an additional 12 items. The items we removed in this case had both discrimination parameters of 1.00 or less and difficulty parameters of 1.00 or more. We decided to keep items with *b* parameters of less than 1.00 because items in this difficulty range were small in number. The reliability of the original version of the grammar test (k = 70) was .955, and the reliability of the revised shorter version of the test (k = 50) was .948. Thus, although the length of the test was considerably shortened, the reliability remained approximately the same, which means that the revised version of the test would probably be shorter and more efficient.

[Insert Table 12 About Here]

After we removed 20 items, we ran the *XCalibre*^{tm} program again on the remaining grammar items. Table 13 shows the results of the three-parameter analysis of the 50 items in the revised version of the grammar test. The difficulty parameters ranged from -.90 to 2.22. The items are sorted in Table 13 from the item that had the lowest difficulty parameter to the item that had the highest difficulty parameter. In the original version of the test, test items were numbered from the easiest item to the most difficult item based on the test developers' judgments. However, as in the listening test, the test developers' perceptions of the relative difficulties for certain grammar items did not necessarily match the relative values of the difficulty parameters for these items.

[Insert Table 13 About Here]

Figure 2 is a visual representation of the difficulty parameters for the 50 items in the revised version of the grammar test. From the left, the items are arranged from the easiest item to the most difficult. Figure 2 shows that the difficulty parameters for about four-fifths of the items were above the average logit of zero. This result indicates that the overall difficulty of the grammar test was

high, and that the test did not have a sufficient number of easy items to effectively discriminate among low ability students. The discrimination parameters ranged from .63 (item 31) to 2.17 (item 75). A total of 37 items had an *a* parameter above 1.00, and three items had an *a* parameter above 2.00. Items with the highest discrimination parameters tended to be of above average difficulty. The remaining items had *a* parameters below 1.00. The guessing parameters were low for the revised grammar test, ranging from .06 to .14.

[Insert Figure 2 About Here]

Results for the Recognition Test

Next we ran the *XCalibre*^{tm} program on the original recognition data (see Table 14). A total of 25 items or half of 50 items were flagged indicating that the standardized residual statistic exceeded a value of 2.0. Since we wanted to keep at least 30 items in the revised version, we decided to eliminate only 20 items out of the 25 flagged items. The items that we decided to keep (items 9, 28, 30, 31, & 41) had higher discrimination parameters than other flagged items. The reliability of the original version of the test (k = 50) was .951, and the reliability of the revised version of the test (k = 30) was .922. Thus, although the length of the test was considerably shortened, the reliability remained about the same, which means that this revised version of the test would probably be shorter and more efficient.

[Insert Table 14 About Here]

After 20 items were eliminated, we run the *XCalibre*^{tm} program on the recognition again. Interestingly, this time, the program flagged no items—not even items 9, 28, 30, 31, and 41, which had been flagged in the analysis of the original 50-item version. Table 15 shows the results of the

three-parameter analysis of the 30 items in the revised version of the recognition test. The difficulty parameters ranged from –2.83 to 1.76. The items are sorted from the item that had the lowest difficulty parameter to the item that had the highest difficulty parameter. Eight of the easiest items (items 86, 89, 92, 87, 94, 93, 91, 96) were *kana* spelling questions. The rest of the items were *kanji* recognition. Again, the items in the original version of the test had been numbered from the easiest item to the most difficult based on the test developers' judgments. In the analysis, the test developers' intended difficulties for certain items did not necessarily match the difficulty parameters we found for these items.

[Insert Table 15 About Here]

Figure 3 is a visual representation of the difficulty parameters for the 30 items in the revised recognition test. From the left, the items are arranged from the easiest item to the most difficult. Figure 3 indicates that, compared to the listening and grammar tests, the difficulty parameters were more balanced, that is, the numbers of items above and below the average item difficulty logit of .00 were approximately the same. Figure 3 also indicates that, unlike the listening and grammar tests, the recognition test had very easy items with *b* parameters of -2.00 or less (these were *all kana* spelling questions). The discrimination parameters ranged from .54 (item 86) to 1.83 (item 119). In total, 24 items had *a* parameters above 1.00, and only five items (they were all *kana* spelling items) had *a* parameters below 1.00. The guessing parameters seemed to be consistently very low ranging from .05 to .07.

[Insert Figure 3 About Here]

Results for the Essay Test

Four issues arose in analyzing the results of the essay test: the reliability of the subscales used to rate the essays, interrater reliability, differences in rater severity, and differences in topics. Naturally, we will explore each of these issues in turn.

Subscale reliability. In this case, Cronbach alpha was used to estimate the reliability, or consistency, within each rater's judgments across the five subscales (content, organization, vocabulary, language use, and mechanics). The Cronbach alpha reliability estimates found here for each of the three raters (across the five subscales as items) were .964, .977, and .961, respectively. These three estimates are all very high indicating that the raters were very consistent in the scores they assigned across subscales: students who scored high on one subscale tended to score high across all subscales, and students who scored low on one subscale tended to score low on all subscales, and so forth.

Interrater reliability. Interrater reliability is used to estimate the consistency of ratings between raters. Table 16 presents the interrater reliabilities of the essay subscales and total scores in a variety of ways. The first three rows show the interrater correlation coefficients for each pair of raters with the first five columns showing the interrater correlation coefficients for the five subscales (which ranged from of .770 to .886) and the last column showing the coefficients for the total scores (which ranged from .881 to .904). The fourth row shows an average (averaged across the three raters using the Fisher *z* transformation) of .810 to .860 for the five subscales and .895 for the total scores. The fifth and sixth columns show the reliabilities for subscales and total scores (estimated using the

Spearman-Brown prophecy formula) for cases where the scores of two raters or three raters are combined, respectively. For two raters, the reliability ranged from .898 to .925 for the five subscales and was .945 for the total scores. For three raters, the reliability ranged from .927 to .949 for the five subscales and was .962 for the total scores. Summarizing these results for total score reliabilities (which would be the basis for any decision making), the one-rater reliability found for this analytic writing scale was .895, two-rater reliability was .945, and three-rater reliability was .962.

[Insert Table 16 About Here]

Differences in rater severity. Given the generally high interrater reliabilities reported above, some readers might be tempted to infer that one rater would be sufficient for applying this analytic writing scale in all situations. Such an inference would be incorrect. First, the interrater reliability information given above is only applicable to situations where the students are as widely dispersed in terms of ability as they were in this placement-testing situation. Second, raters can produce highly correlated scores that are at different levels on the scale. In cases where only one rater is used, unless exactly the same rater is used for all students, the scores that students get could depend to a large extent on which rater happened to rate their essays. Naturally, such a situation would not be fair to all students.

[Insert Tables 17 & 18 About Here]

Table 17 shows the means and standard deviations for each of the three raters. Notice that the three means are not exactly the same. Table 18 shows the means and standard deviations for each of the five subscales on the analytical essay-rating grid. Notice again that the five means are not exactly the same. Table 19 shows the means and standard deviations for each of the five subscales as they were applied by each of the three raters. And, again, quite naturally, the fifteen means are not all

same. One large question begs to be addressed with regard to the descriptive statistics shown in Tables 17, 18, and 19: are these observed differences among the means for raters, subscales, and their interaction chance fluctuations, or are they something potentially more meaningful?

[Insert Table 19 About Here]

Two-way repeated-measures analysis of variance (ANOVA) procedures (with raters and subscales as the repeated-measures independent variables and scores as the dependent variable) were used to examine the differences among means for raters, subscales, and their interaction to determine whether they were chance fluctuations, or rather were probably outside of the distribution of chance fluctuations at the preset probability level of less than one percent, or _ < .01. Table 20 provides the results of this ANOVA analysis.

[Insert Table 20 About Here]

Notice in Table 20 that the main effects for raters, subscales, and their interaction are all significant at $p < .005$ as indicated by the asterisks next to the *F* ratios and the footnote below the table.[3] The fact that the main effects for raters and subscales are significant means that there is less than a one percent probability that the observed fluctuations in means for raters and subscales occurred by chance alone. The significant interaction effect means that the observed differences between raters and subscales are not parallel or systematic. Further investigation of that interaction effect should help clarify the relationships among the means for raters and scales.

Figure 4 graphically shows the interaction effect. Notice how each line represents one rater across the five subscales. The fact that the lines cross for raters one and three as well as for raters two and three illustrates how raters and categories are interacting. If the lines were parallel (that is, systematic), the interaction effect would not have been significant, but since they are not parallel and

the interaction effect is significant, we must carefully think about what such an interaction effect indicates.

Notice first of all that, on average, the raters disagree the most on the first subscale (content) and disagree the least on the third subscale (vocabulary). The lines cross because the three raters are more or less severe in their judgments on the five scales, and because they differ in this severity. The effect of such individual rater differences tends to be lessened when two or three raters' judgments are pooled by averaging them. The average of any two raters, would clearly lessen the differences between raters through the averaging process. The effects of such differences are moderated even more if three raters are used, and so forth. Thus, even though the inter-rater reliability was high for one rater, in most situations, it would probably be irresponsible to use only one rater.

[Insert Figure 4 About Here]

We should also explain the eta^2 values shown in the column furthest to the right in Table 20. Eta^2 helps put into perspective the relative importance of factors, interactions, and error in such an ANOVA. Here eta^2 indicates that a total of 4.97% of the variance in the scores of students in this design was accounted for by differences in raters, while 1.47% was accounted for by differences in subscales, and 2.36% was accounted for by their interaction, while the remaining 91.20% (32.57+ 34.04 + 24.59 = 91.20) is not accounted for in this design and may be due to differences among students, or chance, or other factors not considered here. In any case, raters accounted for nearly 5% of the variance in scores, and raters, subscales, and their interaction together accounted for a total of 8.80% of the variance. These are hardly trivial amounts of variance that can easily be ignored, especially since averaging across at least two raters and across subscales can help moderate these differences.

Differences in topics. Another issue that arose while we were doing this research was the degree to which there might be differences in the performances of students due to differences in the topics and the fact that students were allowed to select their own topic. Table 21 shows the descriptive statistics for the three topics. Recall that those three topics involved describing (a) how you like to spend your vacation; (b) yourself, your background, and your family; or (c) Hawaii to a friend in Japan. Notice in Table 21 that topic (b) has the highest mean (M = 73.41) and topic (c) the lowest (M = 67.15) which is more than a six point difference in means based on topic alone. Notice also that topic (a) has the greatest dispersion of scores as indicated by the standard deviation (SD = 12.20) and that the dispersion is approximately equal for topics (b) and (c) as indicated by their standard deviations of 10.21 and 9.86, respectively.

[Insert Table 21 About Here]

The question that remains is whether the observed differences in means shown in Table 21 occurred by chance alone or are significantly different, and if the later, which pairs are significantly different? Table 22 shows the results of a one-way ANOVA procedure, which indicate a p value for differences between topics of less than .005.[4] This means that a significant difference exists between at least one pair of the three means. Scheffé post hoc analyses indicated a significant difference (at below .005) between topics (b) and (c), but no significant differences for either of the two other possible pairings of means. Note also that the column furthest to the right in Table 22 shows an eta^2 value between topics of .0544, which indicates that 5.44 percent of the variance in writing scores is accounted for by differences in topics.

However, we have to interpret the results very cautiously. Recall that among 428 students who wrote an essay, 78 chose prompt (a), 253 prompt (b), and 83 prompt (c) and that all essays

written by heritage language students (i.e., students with at least one native Japanese speaking parent) were rated because such students were much smaller in number and also because an analysis focused on these students was planned for future research. Also recall that in order to make the sample size for each prompt equal to the number for the smallest group (*N*=78), the rest of the students (all non-heritage students) were randomly selected for each of the other two prompts. As the result of this sample selection procedure, each group of 78 had a different number of heritage students: 10 heritage students for topic (a), 31 for topic (b), and 12 for topic (c). It is possible that topic (b) had the highest mean and the greatest dispersion because the number of heritage students who chose that topic was about three times as large as the number who chose the other two topics. Because of this limitation, we cannot say that the observed significant difference in means between topics (b) and (c) was due strictly to a difference in topic. Nonetheless, for whatever reason this significant difference occurred, it signals the importance of focusing future research on whether topic differences can be a significant source of error variance and need to be controlled (for an example of such results in English L1 composition, see Brown, Hilgers, & Marsella, 1991).

[Insert Table 22 About Here]

Comparing All Four JPT

Descriptive statistics and reliability for the original tests and subscales. Table 23a shows the descriptive statistics and reliability estimates for the three multiple-choice tests, as well as the five subscales and total scores for the essay test in their *original* versions. Notice that the first column (*N*) of numbers gives the differing sizes of the groups of students taking each test. The successive columns provide the total possible points on each test, the means, the standard deviations, the

minimum scores attained, and the maximum scores attained for the various tests and subscales. Note that the skew statistic did not exceed 1.00 in a positive or negative direction for any test or subscale, which means that none of the distributions reported in Table 23a were markedly skewed. Notice also that the listening test had a mean of 7.77 out of 14 for an average score equivalent to 55.50%; the grammar test had a mean of 26.62 out of 70 for an average score equivalent to 38.03%; and the recognition test had a mean of 22.63 out of 50 for an average score equivalent to 45.26%. Thus the listening test was reasonably well centered, but the grammar and recognition tests appeared to be very difficult for the students as a group and were not as well centered. Such difficult tests might prove to be very discouraging experiences for the average student and are certainly not contributing to the norm-referenced placement decision as well as might be expected. The average score on the writing test was 70.05, which reflects a difficulty level in line with typical uses of such writing scales.

Table 23a also shows that the original version of the listening test was moderately consistent with a reliability estimate of .797, meaning that the test was 79.7% reliable when administered under these conditions to these students. The results for the original versions of the grammar, recognition, and essay tests would be considered *very* reliable with estimates of .955, .951, and .962, respectively.

The standard error of measurement (*SEM*) is also a reliability statistic. However, instead of estimating the proportion of consistent variance accounted for by a test as reliability estimates do, the *SEM* estimates the distribution of unreliable variance in score points. Thus the *SEM* is an estimate of how much we can expect students' scores to vary by chance alone, given what we know about the descriptive and reliability characteristics of a give test. The *SEM* can be especially valuable in making decisions at cut-points. Since we know that 68% of the students will vary within one

SEM of the cut-point (on average) by chance alone, it is often good policy to make decisions about those students within one *SEM* of a cut-point with special care. For example, let's say that a cut-point of 35 points is the dividing line between two courses in a placement decision based on the original grammar test shown in Table 23a and that the *SEM* is therefore 3.36 points. Based on this knowledge, we should look especially carefully at students who score between 31.64 and 38.36 points because they are very close to the cut-point (in terms of random score fluctuations) and indeed, could be on the other side of the cut-point by chance alone if they were to take the test again. The effects of such random fluctuations can be reduced by getting additional information (in the form of additional test scores, written essays, interviews, etc.) about the students within this band of potential random fluctuations, and basing the placement decisions on all of the available information. Thus a low *SEM* is a good *SEM* because low means that the fluctuations will be narrow around our cut-points.

With the exception of the listening test which had an *SEM* of 1.49, which is large relative to the magnitude of the mean (7.77) and standard deviation (3.31), the *SEM*s for the grammar, recognition, and essay tests (at 3.36, 2.39, and 2.16, respectively) were all reasonably low in light of the associated means and standard deviations.

[Insert Table 23a About Here]

Descriptive statistics and reliability for the revised tests and subscales. Table 23b shows the descriptive statistics for the three multiple-choice tests, as well as the five subscales and total scores for the essay test in the *revised* versions. Notice that Table 23b is laid out the same as Table 23a for easy comparison. Note also that the listening test has a mean of 6.79 out of 13 for an average score equivalent to 52.23%; the grammar test had a mean of 18.43 out of 50 for an average score equivalent

to 36.86%; and the recognition test had a mean of 15.23 out of 30 for an average score equivalent to 50.77%. Thus, the revised versions of the listening and recognition tests are better centered than the original versions. However, the grammar test is still very difficult for the student and probably needs more work (particularly the addition of easier items that discriminate well at the lowest levels of proficiency. As shown in Table 23b, the reliability estimates and *SEM*s for the revised versions of the tests were all very similar for the original and revised versions of the tests (naturally, they are exactly the same for the original and revised versions of the essay subscales and total test scores because no revisions took place on this test). Thus, even though the revised versions of the listening, grammar, and recognition tests are shorter than the original versions, the reliability of the shorter versions is about the same, meaning that overall the revised versions of the tests are as reliable as the original versions, but are considerably more efficient.

[Insert Table 23b About Here]

Correlational analysis for original tests and subscales. Table 24a shows the correlation coefficients *above the diagonal* (the 1.000s running from the top-left corner to bottom-right corner) for each possible pairing of tests and subscales. Note that all these correlation coefficients are statistically significant ($p < .01$, one-tailed). Thus the probability of their having occurred by chance alone is less than one percent. Notice that the listening, grammar, and recognition tests correlate moderately with each other and with the essay subscales and total in the range between .577 and .754. Note also that, with the exception of mechanics, the essay subscales are highly intercorrelated in the range between .945 and .977. The coefficients for mechanics with the other four subscales are somewhat lower but still moderately high ranging from .801 to .831. The essay subscales correlated

with the total essay scores (of which each was a part) at .882 to .987 with mechanics once again being somewhat lower than the other subscales.

[Insert Table 24a About Here]

The coefficients given *below the diagonal* are adjusted for attenuation in both measures, that is, they are estimates of what the correlation coefficient would be if both measures involved in each were perfectly reliable. Naturally, all the coefficients are somewhat higher than the unadjusted coefficients given above the diagonal. The adjusted coefficients show exactly the same pattern of correlations as the unadjusted coefficients, with (a) moderate correlations for listening, grammar, and recognition with each other and with the essay subscales and total, (b) high intercorrelations among the essay subscales and total, and (c) mechanics being somewhat lower than the other essay subscales. Since the pattern is similar for the adjusted and unadjusted coefficients, it is unlikely that the observed differences in correlation among tests and subscales are due to differences in the reliability of the measures. The overall pattern of correlations tends to support the convergent validity of the various tests and subtests involved. In other words, to some degree the tests are all measuring related constructs.

Correlational analysis for revised tests and subscales. Table 24b again shows the correlation coefficients *above the diagonal* for each possible pairing of the *revised* tests and subscales. Notice that, once again, the listening, grammar, and recognition tests correlate moderately with each other and with the essay subscales and total in the range between .598 and .755. Note also that, with the exception of mechanics, the essay subscales are highly intercorrelated in the range between .945 and .973. The coefficients for mechanics with the other four subscales are somewhat lower but still moderately high ranging from .807 to .831. The essay subscales correlated with the total essay

scores (of which each was a part) at .882 to .987 with mechanics once again being somewhat lower than the other subscales. In addition, the pattern of coefficients adjusted for attenuation in both measures in Table 24b (below the diagonal) is exactly the same as the related pattern of unadjusted coefficients, so it is likely that any observed differences in correlation among tests and subscales are due to factors other than reliability of the measures. In short, the very similar patterns for the correlations among tests and subscales found in Tables 24a for the original versions and Table 24b for the revised versions indicate that the convergent validity of the revised versions of the tests and subscales is very similar to that of the original versions.

[Insert Table 24b About Here]

Factor analysis for original tests and subscales. Another way to investigate the underlying patterns in a set of correlation coefficients like those shown above the diagonal in Table 24a is to use factor analysis procedures. We began these analyses of the *original versions* of the tests and subtests by running a principal components analysis. In our first analysis, the decision about the number of factors to extract was left up to the SPSS computer program, which selects the number of factors with Eigen values of 1.00 or higher. The result was the single factor solution shown in Table 25a. In this case, the Eigen value for this single component was 6.33 (while the next three factors had Eigen values of .721, .471, and .234), and the proportion of variance accounted for by the single component analysis was .792, or 79.2%. Notice that the loadings on the single component range from a moderate .795 to a high of .964. The communalities, which in this case are just the squared loadings, range from .599 to .925 indicating that this single component accounts for 59.9% to 92.5% of the variance among the tests and subscales.

[Insert Table 25a About Here]

Another approach for deciding the number of factors to extract in this sort of analysis is to examine a scree plot of Eigen values plotted against the factor numbers. The scree plot for these data (shown in Figure 5) supports the notion that one factor is appropriate for these data.

[Insert Figure 5 About Here]

Still other approaches for deciding on the number of factors are often used. In this study, we chose to make use of a combination of two other approaches: (a) investigating varying numbers of factors and (b) stopping when all nontrivial amounts of variance are all accounted for (see Gorsuch, 1983, pp. 164-171). The result was a series of three factor analyses, our second, third, and forth, each of which is based on the same PCA and each of which is illuminating in its own right.

The second analysis was the two-factor solution with Varimax rotation shown in Table 25b. We selected Varimax as our rotation procedure because we wanted to investigate the degree to which orthogonal factors might underlie the correlations shown above the diagonal in Table 24a, while at the same time simplifying factors "by maximizing the variance of the loadings within factors, across variables [tests and subscales in this case]" (Tabachnick & Fidell, 1996, p. 666).

Notice in Table 25b that the proportion of variance accounted for by the first factor is .523 and by the second factor is .359 for a total of .882. We decided this solution added a non-trivial amount of variance because: (a) the 88.2% accounted for in this two factor solution is 9.0% more than the 79.2% accounted for in the one component solution, (b) the second factor accounted for 35.9% of the variance, and (c) all of tests or subscales loaded above .30 on each factor (after Gorsuch, 1983, pp. 164-171). Notice that the communalities in Table 25b range from a moderate .764 to a very high .976 and that they are generally higher and less varied than the squared loadings in the single component solution. The loadings in bold-faced italics in this and the subsequent two tables are those over .60.

The writing subscales all load heavily on the first factor, while the listening, grammar, and recognition tests all load heavily on the second factor. Note also however that the listening, grammar, and recognition tests all load above the traditional cut point of .30 on factor 1 and the five essay subscales load even more heavily on factor 2.

[Insert Table 25b About Here]

The third analysis was the three-factor solution with Varimax rotation shown in Table 25c. Notice that the proportion of variance accounted for by the first factor is .487, by the second factor is .255, and by the third factor is .198 for a total of .941. We decided that this solution added a non-trivial amount of variance because: (a) the 94.1% accounted for in this three-factor solution is 5.9% more than the 88.2% accounted for in the two-factor solution, (b) the third factor accounted for 19.8% of the variance, and (c) at least two tests or subscales loaded above .30 on each factor. Notice that the communalities range from a fairly high .885 to a very high .979 and that these values are both higher and less varied than the ones in the single component and two factor solutions. Again, the loadings in bold-faced italics are those over .60. Note also that the writing subscales all load heavily on the first factor, and the listening and grammar tests load heavily on the second factor, while the recognition test and mechanics subscale load heavily on the third factor. Note also however, that, in addition to the patterns in the previous sentence, at least one of the other tests or subscales loads above the traditional cut point of .30 on each of the factors.

[Insert Table 25c About Here]

The fourth analysis was the four-factor solution with Varimax rotation shown in Table 25d. Notice that the proportion of variance accounted for by the first factor is .491, by the second factor is .201, by the third factor is .156, and by the fourth factor is .121 for a total of .969. We decided

cautiously (see warning in the next paragraph) that this solution added a non-trivial amount of variance because: (a) the 96.9% accounted for in this four-factor solution is 2.8% more than the 94.1% accounted for in the three-factor solution, (b) the fourth factor accounted for 12.1% of the variance, and (c) at least two tests or subscales loaded above .30 on each factor. Notice that the communalities range from a very high .931 to an even higher .997 and that these values are all high and certainly less varied than the ones in the other three solutions. With communalities of 93.1% to 99.7%, virtually all the variance in each test or subscale is accounted for. The writing subscales all load heavily on the first factor, while recognition and mechanics load heavily on the second factor, listening on the third factor, and grammar on the fourth factor. Note also however, that, in addition to the patterns in the previous sentence, at least one of the other tests or subscales loads above the traditional cut point of .30 on each of the factors.

[Insert Table 25d About Here]

We must be very careful in interpreting the four-factor solution since the results of this solution may largely be a result of the analyses done. What we mean is that the essay scales may be so highly intercorrelated that they are working largely as one variable. If that is the case, then the four variables (listening, grammar, recognition, and essay) would each naturally load very highly on one factor of their own because of the form of analysis alone.

Nonetheless, the patterns of factor loadings in Tables 25b, 25c, and 25d show at least some degree of divergent construct validity into the four separate skills beyond what would be expected due to simple test method effects alone. The fact that mechanics loads moderately with both the essay subscales and with the recognition test in two of the analyses makes perfect sense given the emphasis in both on the ability to read or write *kana* and *kanji*.

Factor analysis for revised tests and subscales. Similar factor analyses were conducted for the *revised versions* of the tests and subtests based on the correlation coefficients shown above the diagonal in Table 24b. We began by running a principal components for the single factor that showed an Eigen value of 1.00 or higher. The result was the single factor solution shown in Table 26a. In this case, the Eigen value for this single component was 6.33 (while the next three factors had Eigen values of .681, .481, and .252), and the proportion of variance accounted for by the single component analysis was .789, or 78.9%. Notice that the loadings on the single component range from a moderate .777 to a high of .964. The communalities, which in this case are just the squared loadings, range from .601 to .929 indicating that this single component accounts for 60.1% to 92.9% of the variance among the tests and subscales.

[Insert Table 26a About Here]

Recall that another approach for deciding the number of factors to extract in this sort of analysis was to examine a scree plot of Eigen values plotted against the factor numbers. The scree plot was almost exactly the same for these data as that shown in Figure 5, which means to things: (a) there is no need to present it here and (b) this scree plot also supported the notion that a one factor solution is appropriate for these data.

The second analysis of the revised versions was the two-factor solution with Varimax rotation shown in Table 26b. Notice in Table 26b that the proportion of variance accounted for by the first factor is .518 and by the second factor is .358 for a total of .877. We decided this solution added a non-trivial amount of variance because: (a) the 87.7% accounted for in this two factor solution is 8.8% more than the 78.9% accounted for in the one component solution, (b) the second factor accounted for 35.8% of the variance, and (c) all of tests or subscales loaded above .30 on each factor

(after Gorsuch, 1983, pp. 164-171). Notice that the communalities in Table 26b range from a moderate .743 to a very high .976 and that they are generally higher and less varied than the squared loadings in the single component solution. Note also that the writing subscales all load heavily on the first factor, while the listening, grammar, and recognition tests all load heavily on the second factor. Notice also, however, that the listening, grammar, and recognition tests all load above the traditional .30 cut point on factor 1 and the five essay subscales load even more heavily on factor 2.

[Insert Table 26b About Here]

The third analysis was the three-factor solution with Varimax rotation shown in Table 26c. Notice that the proportion of variance accounted for by the first factor is .482, by the second factor is .252, and by the third factor is .203 for a total of .936. We decided that this solution added a non-trivial amount of variance because: (a) the 93.6% accounted for in this three-factor solution is 5.9% more than the 87.7% accounted for in the two-factor solution, (b) the third factor accounted for 20.3% of the variance, and (c) at least two tests or subscales loaded above .30 on each factor. Notice that the communalities range from a fairly high .869 to a very high .979 and that these values are both higher and less varied than the ones in the single component and two factor solutions. Note also that the writing subscales all load heavily on the first factor, and the listening and grammar tests load heavily on the second factor, while the recognition test and mechanics subscale load heavily on the third factor. Note also however, that, in addition to the patterns in the previous sentence, at least one of the other tests or subscales loads above the traditional cut point of .30 on each of the factors.

[Insert Table 26c About Here]

The fourth analysis was the four-factor solution with Varimax rotation shown in Table 26d. Notice that the proportion of variance accounted for by the first factor is .485, by the second factor

is .202, by the third factor is .153, and by the fourth factor is .128 for a total of .968. We decided cautiously (see warning in the next paragraph) that this solution added a non-trivial amount of variance because: (a) the 96.8% accounted for in this four-factor solution is 3.2% more than the 93.6% accounted for in the three-factor solution, (b) the fourth factor accounted for 12.8% of the variance, and (c) at least two tests or subscales loaded above .30 on each factor. Notice that the communalities range from .916 to .998 and that these values are all high and certainly less varied than the ones in the other three solutions. With communalities of 91.6% to 99.8%, virtually all the variance in each test or subscale is accounted for. The writing subscales all load heavily on the first factor, while recognition and mechanics load heavily on the second factor, listening on the third factor, and grammar on the fourth factor. Note also however, that, in addition to the patterns in the previous sentence, at least one of the other tests or subscales loads above the traditional cut point of .30 on each of the factors.

[Insert Table 26d About Here]

Once again, we would like to remind readers that we must be very careful in interpreting the four-factor solution since the results of this solution may largely be a result of the analyses done. What we mean is that the essay scales may be so highly intercorrelated that they are working largely as one variable. If that is the case, then the four variables (listening, grammar, recognition, and essay) would each naturally load highly on one factor each simply because of the form of analysis.

Nonetheless as for the original versions of the tests, the patterns of factor loadings for the revised versions (shown in Tables 26b, 26c, and 26d) show at least some degree of divergent construct validity into the four separate skills beyond what would be expected due to simple test method effects alone. And again, the fact that mechanics loads moderately with both the essay subscales and

with the recognition test in two of the analyses makes perfect sense given the emphasis in both on the ability to read or write *kana* and *kanji*.

DISCUSSION

In this section, we will begin by directly addressing the three research questions posed at the beginning of this study. The research questions will serve as headings in order to keep the discussion clearly organized. Our focus in this discussion will be on the practical implications of our results for test development.

Based on the Three-Parameter Item Response Theory Model, Which (and How Many) Multiple-Choice Items on the Listening, Grammar, and Recognition Tests Are at the Appropriate Levels of Discrimination, Difficulty, and Guessing?

The results of the first IRT analysis of item discrimination, difficulty, and guessing indicate that the listening test could be made more efficient by shortening it from 14 items to 13 without sacrificing reliability. However, the relatively low reliability estimates for either of these lengths indicates that items should be added (especially relatively easy items that discriminate at the lowest levels of ability) to make a new revised version of the test of perhaps 30-40 items. The test would then have to be reanalyzed to once again make it shorter and more efficient and then validate it.

The results of the second IRT analysis of item discrimination, difficulty, and guessing indicate that the grammar test could be made more efficient by shortening it from 70 items to 50 without sacrificing reliability. At the same time, relatively easy items that discriminate at the lowest

levels of ability might profitably be added early in the test. The test would then necessarily be reanalyzed to once again make it shorter and more efficient.

The results of the third IRT analysis of item discrimination, difficulty, and guessing indicate that the recognition test could be made more efficient by shortening it from 50 items to 30 without sacrificing reliability.

In short, in terms of practical test development implications, the results of this study suggest that the three multiple-choice tests could be made much more efficient without sacrificing reliability to any great extent.

For the Analytic Writing Scale, How Many Raters Are Most Efficient in Terms of the Trade off Between Number of Raters and Test Reliability?

In terms of practical test development implications, the results of the interrater reliability and Spearman-Brown formula analyses of the essay ratings indicate that the analytic writing scale used here can be applied with fair reliability by one rater (.895), with a great deal of reliability by two raters (.945), and even more reliability by three raters (.962). However, in terms of the trade-off between efficiency and reliability, using three raters appears to be a waste of resources, at least for norm-referenced purposes with a sample of students across the entire range of abilities tested in this study. In addition, the five percent gain in reliability afforded by using two raters would seem to be worthwhile, especially in view of the mean differences observed among raters that would be moderated by averaging at least two raters.

However, even with high interrater reliability, a significant mean difference was also observed between topics (b) and (c) in this study. At the moment, students choose which topic they will

write on, so it is impossible to say whether the observed differences are due to differences in the difficulty of the tasks or to differences in the types of students who choose each of the three tasks. However, if future research can pin down the source of those differences, they might have important policy implications. Alternatively, this issue could be avoided completely in any one of several ways:

1. All students could be assigned to one topic (which might have security implications)
2. The students could be randomly assigned to the three topics and scores could be equated using standardized scores and/or regression analysis.
3. The students could be randomly assigned to the three topics and scores could be equated using item response theory

To What Degree Are the Four Tests Valid in Terms of Their Correlations with Each Other?

In terms of practical test development implications, the pattern of correlation coefficients shown in Tables 24a and 24b indicate a certain degree of convergent validity for all the tests in this study, especially the subscales within the essay test. In addition, when factor analyses was applied, support was found for the divergent validity of the tests and subscales based on the language skills of listening, grammar, recognition, and essay writing. Further research is always warranted in support of the validity of such tests (see ***Suggestions for Future Research***).

CONCLUSIONS

In conclusion, we would like to (a) summarize the practical test development implications of our results, (b) explore some of the decision-making and policy implications of this study, and (c) make some suggestions for future research into the effectiveness and efficiency of the listening, grammar, recognition, and essay tests.

Practical Test Development Implications

In terms of practical test development, the results of this study suggest that all four tests could be made more efficient without sacrificing reliability to any appreciable degree. To maximize the effectiveness of these tests, we would suggest the following:

1. The grammar test would be more efficient if shortened from 70 items to 50; at the same time, relatively easy items that discriminate at the lowest levels of ability might profitably be added early in the test.
2. The listening test would be more efficient if shortened from 14 items to 13; the relatively low reliability estimates for both of these lengths indicates that items should be added, especially relatively easy items that discriminate at the lowest levels of ability to make a new revised version of the test of perhaps 40 items.
3. The recognition test would be more efficient if shortened from 50 items to 30.
4. The items in the listening, grammar, and recognition tests should all be arranged from easiest to most difficult on the basis of actual student performances rather than on the basis of test developer intuitions.

5. The results of the interrater reliability and Spearman-Brown formula analyses indicate that using two raters (instead of one rater, or three raters) would probably be most efficient and yet would still be reliable.

Decision-Making and Policy Implications

The results of this study have raised three separate sets of policy issues related to: using the *SEM*, decision reliability, and test validity. We will end this section by addressing the following question: did we achieve our purpose?

Using the SEM. The standard error of measurement for whatever tests are used, should be incorporated as part of the decision making process. Recall, as we explained earlier in this document, that the SEM provides an estimate of how many test score points of fluctuation can be expected by chance alone. Typically, if a decision-making policy is designed to protect the students from false negative decisions, students who score within one SEM below a particular cut-point would be treated differently from those who score even lower. Further information could be gathered about these students in the form of additional test scores, grades in previous Japanese language studies, interview scores or impressions, previous teachers' comments, etc. to help in deciding if the student probably does belong below the cut-point or simply arrived there by chance alone. Indeed, if this policy issue is considered important, item response theory could be used to provide a much more accurate standard error for each student.

Decision reliability. In a similar vein, decision making, whatever it's purpose, will typically be more reliable if multiple sources of information are used (see Brown & Hudson, 1998). This would be an argument for actually using the scores on two or more of the four tests in making

decisions about students' placement, instead of basing the decision on the grammar test and using the other tests as "back-up information."

Given that new Japanese courses with more oral/aural content are being developed at this time, it may become essential to use at least the listening, grammar, and recognition tests in concert when making placement decisions for these new courses. Indeed, it might even become desirable to develop some form of task-based oral examination, or interview procedure, if oral language is indeed to be a focus of these courses. As a matter of policy, it could also become useful and even necessary to actually use essay test scores for placement into high level courses or for placement into courses focused specifically on writing. Naturally, many other options in language testing could also be considered, such as cloze testing, dictation, performance tests, etc. (for more ideas, see Brown & Hudson, 1998).

Validity. The last issue we need to discuss here is that of validity. The pattern of correlation coefficients shown in Table 24 for the revised tests indicates a certain degree of convergent validity for all the tests in this study, especially the subscales within the essay test. When factor analysis was applied, support for divergent validity was also found based on language skills and testing method. However, as a matter of policy, it would probably be a good idea to encourage further research on the item characteristics, reliability, and validity of the four tests of the JPT battery (including content validity and other construct validity strategies, as well as the social consequences and value implications of score interpretation and use).

Did we achieve our purpose? As we argued at the beginning of this report, first-rate placement procedures are extremely important for effective teaching and successful learning because they can help create classes that are relatively homogeneous in terms of the language

proficiency levels of the students. The main purpose of the present study was to investigate how effectively and efficiently the current norm-referenced JPT battery for the Japanese language program at the University of Hawai‘i at Manoa (UHM) separates the incoming students of Japanese from diverse language abilities into different course levels. We found that basically the tests were reasonably reliable and valid for measures that had never previously been analyzed statistically. We also found that some minor revisions and changes in administration and scoring procedures as well as testing policies could make the JPT battery even more efficient, fairer, and more professional in terms of making placement decisions about the students in the Japanese language courses.

Another purpose of this study was to explore the potential of item response theory for analyzing the JPT battery for university students of Japanese. Clearly, item response theory helped inform the analysis reported here. Many of the revision decisions we made would not have been possible without IRT. Indeed, from classical theory alone, we would have had no idea that two of the tests (listening and grammar) did not have enough lower-level items that discriminate effectively among low proficiency students. We would also never have realized that guessing is *not* a major factor on any of the multiple-choice tests studied here.

Suggestions for Future Research

1. To what degree would analyses of actual administrations of the revised versions of the listening, grammar, and recognition tests produce similar results to those estimated here (without re-administering the revised tests) in terms of item discrimination, difficulty, and guessing?

2. To what degree are the scores on the four tests in this study content, criterion-related, and construct valid?
3. To what degree are the decisions made with the four tests in this study valid? And, what are the values implications and social consequences of these decisions (see Messick, 1988)?
4. What are the relative contributions to the reliability of the essay test of raters and rating categories as determined by generalizability theory (see Brown, 1999 for more on this type of analysis)?
5. What are the relative differences in severity for the essay test of raters, rating categories, and topics as determined by Facets theory (see Linacre, 1994 for more on this topic)?
6. What differences are observed between the performances of the heritage and non-heritage students on the four placement tests at UHM?
7. What are the contributions to error variance of topics and how can that error variance be eliminated or controlled?

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Assessment Systems. (1997). *XCalibre* (version 1.10e). St. Paul, MN: Assessment Systems.

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

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

Brown, A., & Iwashita, N. (1996). Language background and item difficulty: The development of a computer-adaptive test of Japanese. *System, 24*, 199-206.

Brown, J. D. (1992). Using computers in language testing. *Cross Currents, 19*(1), 92-99.

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

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning and Technology, 1*(1), 44-59.

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

Brown, J. D., Hilgers, T., & Marsella, J. (1991). Essay prompts and topics: Minimizing the effect of differences. *Written Communication, 8*(4), 532-555.

Brown, J. D., & Hudson, T. (1998). Alternatives in language assessment. *TESOL Quarterly, 32*, 653-675.

Canadian Association for Japanese Language Education. (2000). *Kodomono kaiwaryoku no mikata to hyooka [Oral proficiency assessment for bilingual children]*. Ontario, Canada: Soleil.

Carroll, B. J., & Hall, P. J. (1985). *Make your own language tests: A practical guide to writing language performance tests*. Oxford: Pergamon.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and practice, 8*(1), 35-40.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill.

Hatasa, Y. A., & Tohsaku, Y. (1997). SPOT as a placement test. In H. M. Cook, K. Hijirida, & M. Tahara (Eds.), *New trends & issues in teaching Japanese language and culture*—Technical report #15 (pp. 77-98). Honolulu: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

Heaton, J. (1977). *Writing English language tests*. London: Longman.

Henning, G. (1987). *A guide to language testing: development, evaluation, research*. Cambridge, MA: Newbury House.

Ishida, T. (1992). *Nyuumon Nihongo tesutohoo [Introductory Japanese language testing method]*. Tokyo: Taishuukan.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach.* Rowley, MA: Newbury House.

Kobayashi, N., & Ford, J. (1992). Bunpoo koomoku no onsee chooshuu ni kansuru jisshooteki kenkyuu [An empirical study of aural recognition of grammatical items]. *Nihongo Kyooiku [Journal of Japanese Language Teaching], 78*, 167-177.

Kobayashi, N., Ford, J., & Yamamoto, H. (1996). Nihongo nooryoku no atarashii sokuteihoo [SOPT][SPOT: A new method of testing Japanese language proficiency. *Sekaino Nihongo Kyooiku [Japanese-Language Education Around the Globe], 6*, 201-218.

Kondo-Brown, K. (2000) How is high school foreign language study related to the proficiency of heritage language students of Japanese? Unpublished manuscript. Honolulu, HI: University of Hawaii at Manoa.

Kondo-Brown, K., & Brown, J. D. (forthcoming). Investigating the Japanese language placement test. In *Report on the Educational Improvement Fund 1999/2000.* Honolulu: University of Hawai‘i at Manoa.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests.* London: Longman.

Linacre, J. M. (1994). *Many-facet Rasch measurement.* Chicago: Mesa.

Madsen, H. (1983). *Techniques in testing*. Oxford: Oxford University Press.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning of consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum.

Microsoft. (1999). *Excel*^{tm}. Redman, WA: Microsoft.

Rentz, R. R., & Rentz, C. C. (1978). *Does the Rasch model really work? A discussion for practitioners. Technical Memorandum (No. 67)*. Princeton, NJ: ERIC Clearinghouse on Tests and Measurement and Evaluation, Educational Testing Service.

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing*, *8*(2), 95-111.

Society for Teaching Japanese as a Foreign Language (Ed.). (1991). *Nihongo tesuto handobukku [Handbook for Japanese language test]*. Tokyo: Taishuukan.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

Tanaka, M. (1994). Kaki nihongo kyooiku koose saikoo [Toward a reconsideration of the SCJ curriculum]. *The Research Center for Japanese Language Education Annual Bulletin* (International Christian University, Tokyo), *4*, 63-86.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge: Cambridge University Press.

Valette, R. M. (1977). *Modern language testing* (2nd ed.). New York: Harcourt Brace Jovanovich.

APPENDIX A

JAPANESE COMPOSITION SCORING SHEET
(from Kondo-Brown, K., 2000, p. 33, adapted from Jacobs Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981)

SCORE	LEVEL	CRITERIA
CONTENT	20-18 excellent to very good	knowledgeable; substantive; thorough development of thesis; relevant to assigned topic
	17-14 good to average	some knowledge of subject; adequate range; limited development of thesis; mostly relevant to topic, but lacks detail
	13-10 fair to poor	limited knowledge of subject; little substance; inadequate development of topic
	9-7 very poor	does not show knowledge of subject; non-substantive; not pertinent ; OR not enough to evaluate
ORGANIZATION	20-18 excellent to very good	fluent expression; ideas clearly stated/supported; succinct; well-organized; logical sequencing; cohesive; consistent style
	17-14 good to average	somewhat choppy; loosely organized but main ideas stand out; limited support; logical but incomplete sequencing; inconsistent style
	13-10 fair to poor	non-fluent; ideas confused or disconnected; lacks logical sequencing and development
	9-7 very poor	does not communicate; no organization; OR not enough to evaluate
VOCABULARY	20-18 excellent to very good	sophisticated range; effective word/idiom choice and usage; word form mastery; appropriate register
	17-14 good to average	adequate range; occasional errors of word/idiom form, choice, usage but meaning not obscured
	13-10 fair to poor	limited range; frequent errors of word/idiom form, choice, usage; meaning confused or obscured
	9-7 very poor	essentially translation; little knowledge of Japanese vocabulary, idioms, word form; OR not enough to evaluate
LANGUAGE USE	20-18 excellent to very good	effective complex constructions; few errors of agreement, tense, number, word order/function, pronouns, inflections, particles.
	17-14 good to average	effective but simple constructions; minor problems in complex constructions; several errors of agreement, tense, number, word order/function, pronouns, inflections, particles
	13-10 fair to poor	major problems in simple/complex constructions; frequent errors of negation, agreement, tense, number, word order/function, pronouns, inflections, particles; run-ons, deletions; meaning confused or obscured
	9-7 very poor	virtually no mastery of sentence construction rules; dominated by errors; does not communicate; OR not enough to evaluate
MECHANICS	20-18 excellent to very good	Kana and Kanji are well-formed and used appropriately; few errors of spelling, punctuation, paragraphing
	17-14 good to average	occasional errors in the use of Kana and Kanji; occasional errors of spelling, punctuation, paragraphing but meaning not obscured; occasional use of English
	13-10 fair to poor	infrequent or no use of Kanji; frequent errors of spelling, punctuation, paragraphing; poor handwriting; meaning confused or obscured; frequent use of English
	9-7 very poor	no mastery of Kana; dominated by errors of spelling, punctuation, paragraphing; handwriting illegible; Or not enough to evaluate

Table 1

Percents for Gender on the Four Tests

Gender	LIST *N*=939	GRAM *N*=1294	RECOG *N*=1124	ESSAY *N*=234
Female	58.7	54.5	56.6	60.7
Male	35.7	38.8	37.4	36.3
Missing	5.6	6.7	6.0	3.0
Total	100.0	100.0	100.0	100.0

Table 2

Percents for Major on the Four Tests

Major	LIST *N*=939	GRAM *N*=1294	RECOG *N*=1124	ESSAY *N*=234
Undecided	11.8	12.8	12.2	9.4
Architecture	1.0	1.0	1.0	.9
Arts & Sciences (excluding Japanese)	31.8	33.6	32.5	25.2
Business	8.8	8.9	8.8	12.4
Education	4.2	3.9	3.7	3.4
Engineering	5.4	5.2	5.5	4.3
Hawaiian, Asian, Pacific Studies	.3	.8	.6	.0
Health Science &Social welfare	.1	.2	.2	.0
Japanese	2.0	2.9	2.7	3.4
Law	.3	.2	.3	.0
Medicine	3.8	3.1	3.5	4.3
Nursing	1.7	1.6	1.7	1.3
Social Work	.0	.1	.1	.0
TIM	1.6	2.2	2.0	2.1
Tropical Agriculture & Human Resources	.0	.2	.1	.0
Subtotal	72.9	76.5	74.7	66.7
Missing	27.1	23.5	25.3	33.3
Total	100.0	100.0	100.0	100.0

Table 3

Percents for Academic Status on the Four Tests

Academic Status	LIST *N*=939	GRAM *N*=1294	RECOG *N*=1124	ESSAY *N*=234
Freshman	81.5	68.8	73.5	83.3
Sophomore	4.0	9.8	7.0	2.6
Junior	1.8	5.6	3.8	2.6
Senior	1.5	2.6	2.4	1.7
Graduate	.4	2.7	2.5	.0
HS Senior	5.1	3.8	4.4	5.6
Subtotal	94.4	93.4	93.6	95.7
Missing	5.6	6.6	6.4	4.3
Total	100.0	100.0	100.0	100.0

Table 4

Percents for Native Language on the Four Tests

Native Language	LIST *N*=939	GRAM *N*=1294	RECOG *N*=1124	ESSAY *N*=234
English	91.5	88.3	89.9	93.6
Chinese	3.5	4.8	4.0	.9
Korean	2.1	2.6	2.2	2.6
Japanese	.7	.9	1.1	2.6
Japanese & English	.1	.3	.4	.0
Philipino	.4	.5	.4	.4
Vietnamese	1.0	.9	.7	.0
English & Hawaiian	.0	.1	.0	.0
Vietnamese & English	.2	.2	.2	.0
French	.0	.1	.1	.0
Farsi	.1	.1	.1	.0
Tahitian	.0	.1	.1	.0
Subtotal	99.7	98.8	99.1	100.0
Missing	.3	1.2	.9	.0
Total	100.0	100.0	100.0	100.0

Table 5

Percents for Family Japanese Language Background on the Four Tests

Family Language Background	LIST *N*=939	GRAM *N*=1294	RECOG *N*=1124	ESSAY *N*=234
No JPN (Grand)Parent	83.4	81.9	81.5	70.1
Both Parents JPN	4.4	4.6	5.1	9.4
Mother JPN	5.4	6.0	6.1	9.8
Father JPN	1.9	2.1	2.0	3.8
Grandparents JPN	4.8	5.1	5.1	6.8
Subtotal	99.9	99.7	99.8	100.0
Missing	.1	.3	.2	.0
Total	100.0	100.0	100.0	100.0

Table 6

Percents for Years of High School Japanese on the Four Tests

Years of High School Japanese	LIST *N*=939	GRAM *N*=1294	RECOG *N*=1124	ESSAY *N*=234
0.0	3.5	9.8	8.0	4.7
1.0	3.5	4.7	4.3	3.8
2.0	20.3	22.4	19.3	15.4
3.0	29.1	27.2	27.8	20.9
4.0	38.8	31.9	36.0	47.5
5.0	2.8	2.3	2.7	4.3
6.0 or more	1.8	1.6	1.8	3.4
Subtotal	99.9	99.7	99.8	100.0
Missing	.1	.3	.2	.0
Total	100.0	100.0	100.0	100.0

Table 7

Percents for Years of Other Japanese Programs on the Four Tests

Years in Other Japanese Schools	LIST *N*=939	GRAM *N*=1294	RECOG *N*=1124	ESSAY *N*=234
0.0	65.3	61.2	60.7	59.8
1.0	11.1	12.4	12.0	9.9
2.0	8.9	10.3	10.8	13.2
3.0	2.0	2.5	2.6	.9
4.0	2.1	2.4	2.5	2.1
5.0	2.8	2.6	2.9	3.4
6.0	2.0	2.1	2.0	2.6
7.0	1.7	2.0	2.0	1.3
8.0	1.6	1.7	1.7	3.8
9.0	1.2	1.0	1.2	1.3
10.0 or more	1.1	1.2	1.2	1.7
Subtotal	99.8	99.3	99.6	100.0
Missing	.2	.7	.4	.0
Total	100.0	100.0	100.0	100.0

Table 8

Percents for Years Living in Japan on the Four Tests

Years in Japan	LIST *N*=939	GRAM *N*=1294	RECOG *N*=1124	ESSAY *N*=234
0.0	96.0	92.3	92.0	92.3
1.0	1.1	1.8	2.0	2.6
2.0	.3	1.1	1.1	.4
3.0	.2	.9	.6	.4
4.0	.2	.3	.4	.4
5.0	1.0	1.1	1.1	1.7
6.0	.2	.2	.5	.4
7.0	.3	.6	.7	.0
8.0	.1	.2	.1	.4
9.0	.0	.1	.1	.0
10.0 or more	.5	1.2	1.5	1.2
Subtotal	99.9	99.6	99.7	100.0
Missing	.1	.4	.3	.0
Total	100.0	100.0	100.0	100.0

Table 9

Percents for Ultimate Course Placement on the Four Tests

Course Placement	LIST *N*=939	GRAM *N*=1294	RECOG *N*=1124	ESSAY *N*=234
101	1.1	2.0	.7	.0
101/100	6.1	8.9	5.4	2.6
100	43.7	40.3	40.4	23.9
100/102	22.0	18.5	20.1	25.6
102	9.3	9.0	9.5	14.1
102/201	4.9	4.4	4.8	5.1
201	2.2	2.6	2.8	3.0
201/202	1.4	2.2	2.6	1.7
202	1.6	1.7	2.0	4.3
202 or 301	1.4	1.6	1.9	4.3
300-level	.7	1.2	1.3	2.6
300-level or 401	.1	.2	.2	.4
300-level bilingual	2.7	2.9	3.3	7.7
400-level	.4	1.6	1.8	1.3
Overqualified	.1	.2	.2	.4
Missing	2.3	2.7	3.0	3.0
Total	100.0	100.0	100.0	100.0

Table 10
Three-parameter IRT Analysis of the Items in the Original Listening Test (14 Items, 888 Students)

Item No.	Flag	Discrimination Parameter	Difficulty Parameter	Guessing Parameter	Residual
1		.70	-1.93	.15	.40
2	P	.83	-3.00	.15	1.22
3		.68	.20	.15	.30
4		.67	1.22	.14	.57
5		.73	- .38	.15	.53
6		.73	- .77	.15	.29
7		.52	.50	.15	.83
8		.87	2.51	.12	.91
9		.87	.19	.14	.39
10		.96	- .24	.15	.32
11		1.50	1.19	.10	1.54
12		1.58	.68	.12	1.00
14		.83	.52	.14	.32
15		.94	.85	.15	.20

Table 11

The Three-parameter IRT Analysis of the Items in the Revised Listening Test (13 Items, 888 Students)

Item No.	Discrimination Parameter	Difficulty Parameter	Guessing Parameter	Residual
1	.69	-1.95	.15	.42
6	.73	- .77	.15	.41
5	.72	- .39	.15	.57
10	.95	- .24	.15	.34
9	.87	.19	.14	.38
3	.68	.19	.15	.28
7	.52	.50	.15	.84
14	.83	.53	.14	.29
12	1.60	.68	.12	.99
15	.93	.85	.15	.18
11	1.50	1.20	.10	1.51
4	.67	1.22	.14	.57
8	.87	2.51	.12	.89

Table 12
Three-parameter IRT Analysis of the Items in the Original Grammar Test (70 items, 1284 students)

Item No.	Flag	Discrimination Parameter	Difficulty Parameter	Guessing Parameter	Residual
16	R	1.40	1.64	.34	2.95
17		.90	.13	.10	.85
18		.85	.21	.11	.78
19		1.14	-.36	.10	.38
20		.95	.37	.10	.72
21		.67	-.52	.10	.83
22		.89	-.18	.10	.44
23		.98	.35	.09	1.08
24		.80	-.88	.10	.53
25		.78	.12	.10	.83
26		1.15	-.66	.10	.70
27		1.05	.43	.11	.72
28		1.20	1.21	.10	.67
29		1.03	.07	.09	.75
30	R	.78	.20	.11	2.06
31		.62	-.40	.10	.91
32		.91	-.09	.10	1.12
33		.97	.00	.10	.83
34		1.00	-.37	.10	.71
35		1.09	.83	.09	.91
36		1.05	1.85	.08	1.24
37	R	.54	.90	.13	2.68
38		1.34	.40	.08	1.09
39		1.06	1.56	.14	1.63
40		1.20	.41	.09	.83
41		.93	.39	.11	.80
42		1.60	.13	.09	.64
43		.86	-.75	.11	1.19
44		1.43	.42	.09	.85
45		1.10	1.13	.09	1.42
46		.83	1.26	.10	.56
47		1.27	.32	.10	.70
48		1.34	.64	.10	.59
49	R	1.23	1.58	.07	2.11
50		.65	1.75	.12	1.97
51		1.60	1.44	.09	.68
52		1.03	1.03	.09	1.02
53	R	.51	.58	.12	2.13
54		.65	.54	.10	.96

55	1.28	- .18	.10	.79
56	1.35	1.02	.09	1.18
57	1.71	.94	.07	1.64
58	1.12	1.16	.08	1.28
59	1.03	.52	.12	1.94
60	1.58	1.32	.08	.85
61	1.89	1.16	.07	1.35
62	1.72	.90	.08	.86
63	1.54	1.57	.09	.66
64	1.28	1.68	.10	1.11
65	.74	1.20	.11	.95
66	.93	.39	.11	1.01
67	1.19	1.56	.10	.89
68	.83	.54	.11	1.74
69	1.47	1.63	.07	1.45
70	1.61	.39	.10	.77
71	1.70	.95	.09	.89
72	1.10	.27	.09	.74
73	1.03	1.30	.09	.82
74	1.19	.57	.09	.80
75	2.02	1.19	.06	1.81
76	.99	.77	.10	1.26
77	1.55	1.50	.09	.74
78	1.77	2.16	.06	1.41
79	1.05	1.36	.09	.83
80	1.73	.44	.07	1.35
81	.93	.69	.10	.72
82	1.69	2.30	.09	1.12
83R	.80	.83	.12	2.03
84	1.92	1.16	.07	1.31
85	1.23	2.08	.15	1.58

Table 13

Three-parameter IRT Analysis of the Items in the Revised Grammar Test (50 items, 1284 students)

Item No.	Discrimination Parameter	Difficulty Parameter	Guessing Parameter	Residual
24	.80	- .90	.09	.48
43	.84	- .78	.09	.94
26	1.14	- .69	.09	.53
21	.67	- .54	.09	.72
31	.63	- .42	.09	1.00
34	.98	- .38	.09	.62
19	1.13	- .37	.09	.17
22	.85	- .20	.09	.67
55	1.26	- .19	.09	.73
32	.90	- .10	.09	.98
33	.95	- .01	.09	.88
29	1.03	.06	.08	.59
42	1.57	.15	.08	.50
72	1.09	.28	.08	.74
23	.97	.34	.08	.84
47	1.25	.34	.10	1.13
38	1.34	.41	.08	.77
40	1.21	.42	.08	.63
27	1.04	.43	.10	.80
44	1.47	.43	.08	.68
70	1.72	.44	.10	.86
80	1.76	.47	.07	1.00
74	1.21	.58	.08	.88
48	1.35	.67	.09	.91
81	.93	.71	.09	.84
35	1.12	.86	.09	.79
62	1.77	.91	.08	.70
57	1.79	.95	.06	1.35
71	1.79	.97	.08	.69
56	1.42	1.01	.08	1.03
52	1.03	1.04	.08	.85
45	1.12	1.12	.08	1.31
58	1.14	1.18	.07	1.02
61	2.00	1.18	.06	1.07
84	2.00	1.18	.07	1.05
75	2.17	1.20	.06	1.49
28	1.22	1.23	.10	.67
73	1.05	1.30	.09	.88
60	1.66	1.33	.07	.84
79	1.10	1.35	.09	.79
51	1.75	1.42	.09	.56
77	1.70	1.49	.09	.72
67	1.28	1.54	.10	.92
63	1.65	1.55	.08	.78

69	1.58	1.60	.06	1.21
64	1.37	1.67	.10	1.10
36	1.11	1.80	.08	1.07
85	1.39	2.02	.14	1.50
78	1.95	2.09	.06	1.23
82	1.89	2.22	.08	.97

Table 14

Three-parameter IRT Analysis of the Items in the Original Recognition Test (50 Items, 1120 Students)

Item #	Flag	Discrimination Parameter	Difficulty Parameter	Guessing Parameter	Residual
86		.63	-2.26	.26	1.82
87		.70	-1.62	.26	1.69
88	R	1.17	-2.55	.25	3.95
89		.85	-2.17	.26	1.30
90	R	1.17	-2.54	.25	2.45
91		.75	- .66	.26	1.14
92		1.17	-2.16	.25	1.92
93		1.13	-1.58	.24	1.68
94	R	1.53	-1.65	.24	2.02
95	R	.98	-2.46	.25	6.09
96		2.23	- .73	.23	.84
97	R	1.39	- .47	.20	2.59
98		1.26	.36	.19	1.60
99	R	1.60	.14	.17	2.42
100		2.21	- .37	.21	1.06
101	R	2.12	- .40	.19	2.46
102		1.68	.57	.16	1.82
103		2.50	- .58	.21	.95
104		1.82	- .15	.20	1.38
105		1.47	.41	.19	1.46
106	R	1.67	.45	.16	2.15
107	R	1.57	.38	.16	2.32
108		1.53	1.01	.17	1.46
109	R	2.50	1.02	.11	2.61
110		2.06	.69	.14	1.79
111		2.02	.15	.19	.89
112		1.84	.81	.15	1.69
113	R	2.00	1.53	.11	2.14
114		1.71	.54	.17	1.41
115	R	1.73	1.35	.13	2.06
116	R	1.93	1.36	.13	2.04
117		2.14	.72	.15	1.62
118		1.82	1.26	.13	1.97
119		2.17	.58	.15	1.69
120		1.78	.64	.16	1.82
121	R	1.61	1.65	.12	2.24
122	R	2.41	1.71	.10	2.35
123	R	2.39	1.43	.11	2.31
124	R	2.08	1.82	.11	2.46
125	R	2.38	1.72	.11	2.34
126	R	2.23	1.15	.13	2.07

127R	2.50	1.23	.11	2.36
128	1.90	.95	.15	1.62
129R	2.16	1.95	.11	2.12
130R	2.26	1.73	.11	2.39
131R	2.23	1.24	.11	2.52
132R	2.46	1.39	.11	2.21
133	1.97	1.81	.12	1.99
134	1.88	1.23	.13	1.95
135R	2.11	1.73	.10	2.74

Table 15

Three-parameter IRT Analysis of the Items in the Revised Recognition Test (30 Items, 1120 Students)

Item No.	Discrimination Parameter	Difficulty Parameter	Guessing Parameter	Residual
86	.54	-2.83	.07	1.66
89	.73	-2.62	.07	1.03
92	.93	-2.61	.07	1.56
87	.58	-2.15	.07	1.65
94	1.12	-1.99	.07	1.09
93	.88	-1.96	.07	1.09
91	.60	-1.16	.07	.97
96	1.59	- .92	.07	.64
103	1.80	- .71	.07	.43
100	1.58	- .54	.07	.52
104	1.41	- .33	.07	.76
111	1.60	- .01	.07	.54
98	1.00	.17	.07	.92
105	1.13	.22	.07	.65
114	1.40	.39	.06	.55
102	1.34	.44	.06	1.16
119	1.83	.49	.06	.95
120	1.47	.50	.06	.86
110	1.71	.58	.06	.90
117	1.73	.62	.06	.64
112	1.54	.70	.07	.82
108	1.17	.80	.06	.85
128	1.44	.81	.06	.83
126	1.80	1.06	.06	.99
134	1.60	1.10	.06	.94
118	1.49	1.17	.06	.95
115	1.43	1.22	.06	.97
116	1.45	1.25	.06	1.13
113	1.66	1.44	.05	1.03
133	1.33	1.76	.06	.80

Table 16

Interrater Correlation Coefficients Between Pairs of Raters

	CONT	ORGAN	VOCAB	LANG	MECH	TOTAL
Rater 1/Rater 2	.831	.814	.859	.826	.875	.904
Rater 2/Rater 3	.886	.836	.861	.820	.848	.902
Rater 1/Rater 3	.811	.770	.859	.794	.859	.881
Average (using Fisher *z* trans.)	.845	.810	.860	.815	.860	.895
Two Raters Combined	.916	.895	.925	.898	.925	.945
Three Raters Combined	.942	.927	.949	.930	.949	.962

Table 17

Descriptive Statistics for Raters on the Essay Test

(N = 234 in all cases)

Rater	*M*	*SD*	*Min*	*Max*
1	71.68	10.85	40	98
2	68.76	11.35	45	100
3	69.73	12.23	36	100

Table 18

Descriptive Statistics for Subscales on the Essay Test

(N = 234 in all cases)

Subscale	*M*	*SD*	Min	Max
Content	14.07	2.53	8.0	20.0
Organization	13.80	2.43	7.7	19.7
Vocabulary	13.94	2.40	8.0	20.0
Language Use	14.07	2.15	8.3	19.7
Mechanics	14.18	1.95	7.7	19.7

Table 19

Descriptive Statistics for Raters by Subscales on the Essay Test

(N = 234 in all cases)

Rater	Subscale				
1	Content	14.59	2.44	8	20
1	Organization	14.13	2.40	7	20
1	Vocabulary	14.00	2.48	8	20
1	Language Use	14.47	2.14	9	20
1	Mechanics	14.49	2.12	7	19
2	Content	13.52	2.56	9	20
2	Organization	13.48	2.55	9	20
2	Vocabulary	13.78	2.39	9	20
2	Language Use	13.78	2.30	9	20
2	Mechanics	14.21	2.03	7	20
3	Content	14.10	3.02	7	20
3	Organization	13.79	2.85	7	20
3	Vocabulary	14.03	2.68	7	20
3	Language Use	13.97	2.46	7	20
3	Mechanics	13.84	2.01	7	20

Table 20

Repeated-Measures ANOVA with Raters and Subscales Repeated on the Essay Test

Source	*SS*	*df*	*MS*	*F*	Eta^2
Rater	205.66	2	102.83	35.56*	.0497
Error (Rater)	1347.54	466	2.89		.3257
Subscale	60.71	4	15.18	10.05*	.0147
Error (Subscale)	1408.09	932	1.511		.3404
Rater x Subscale	97.73	8	12.22	22.39*	.0236
Error (Rater x Subscale)	1017.07	1864	.546		.2459
Total	4136.80	3276			1.000

* $p < .005$

Table 21

Descriptive Statistics for Topics on the Essay Test (N = 234; 78 for each topic)

Topic	*M*	*SD*	*Min*	*Max*
(a)	69.61	12.20	44	96
(b)	73.41	10.21	45	99
(c)	67.15	9.86	40	88
Total	70.06	11.06	40	99

Table 22

Repeated-Measures ANOVA with Raters and Subscales Repeated on the Essay Test

Source		*df*			Eta^2
Between Topics	1551.91	2	775.96	6.65*	.0544
Within Topics	26962.66	231	116.72		.9456
Total	28514.57	233			1.0000

* $p < .005$

Table 23a
Descriptive Statistics and Reliability for the Original Versions of All

Tests and Subscales

TEST or Subscale	*N*	*Possible*					*Reliat*	S
LISTENING	939	14	7.77	3.31	.0	14.0	.797	1.49
GRAMMAR	1294	70	26.62	15.86	.0	70.0	.955	3.36
RECOGNITION	1124	50	22.63	10.79	2.0	49.0	.951	2.39
Content	234	20	14.06	2.53	8.0	20.0	.942	0.61
Organization	234	20	13.80	2.43	7.6	19.7	.927	0.66
Vocabulary	234	20	13.93	2.40	8.0	20.0	.949	0.54
Language Use	234	20	14.07	2.15	8.3	19.7	.930	0.57
Mechanics	234	20	14.18	1.95	7.7	19.7	.949	0.44
ESSAY TOTAL	234	100	70.05	11.06	40.4	99.0	.962	2.16

Table 23b
Descriptive Statistics and Reliability for the Revised Versions of All

Tests and Subcales

TEST or Subscale	*N*	Poss					*Reliat*	S
LISTENING	939	13	6.79	3.30	.0	13.0	.801	1.47
GRAMMAR	1294	50	18.43	11.93	.0	50.0	.948	2.72
RECOGNITION	1124	30	15.23	7.00	.0	30.0	.922	1.95
Content	234	20	14.06	2.53	8.0	20.0	.942	0.61
Organization	234	20	13.80	2.43	7.6	19.7	.927	0.66
Vocabulary	234	20	13.93	2.40	8.0	20.0	.949	0.54
Language Use	234	20	14.07	2.15	8.3	19.7	.930	0.57
Mechanics	234	20	14.18	1.95	7.7	19.7	.949	0.44
ESSAY TOTAL	234	100	70.05	11.06	40.4	99.0	.962	2.16

Table 24a

Correlation Coefficients Among the Original Versions of the Tests and Subscales*

TEST or Subscale**	1	2	3	4	5	6	7	8	9
1. LISTENING	1.000	.754	.577	.658	.673	.689	.681	.618	.689
2. GRAMMAR	.864	1.000	.674	.679	.688	.707	.722	.626	.711
3. RECOGNITION	.663	.707	1.000	.613	.626	.638	.637	.749	.672
4. Content	.759	.716	.648	1.000	.977	.972	.945	.807	.980
5. Organization	.783	.731	.667	1.000	1.000	.973	.958	.823	.986
6. Vocabulary	.792	.743	.672	1.000	1.000	1.000	.966	.831	.987
7. Language Use	.791	.766	.677	1.000	1.000	1.000	1.000	.822	.976
8. Mechanics	.711	.658	.788	.854	.877	.876	.875	1.000	.882
9. ESSAY TOTAL	.787	.742	.703	1.000	1.000	1.000	1.000	.923	1.000

* All correlations significant at the $p < .01$ level (1-tailed)

** n=938 for grammar and listening correlation; n=933 for recognition & listening; n=1123 for recognition and grammar; n=234 for all other correlations involving the essay test or its subscales

Table 24b

Correlation Coefficients Among the Revised Versions of the Tests and Subscales*

TEST or Subscale**	1	2	3	4	5	6	7	8	9
1. LISTENING	1.000	.755	.598	.656	.670	.686	.678	.616	.686
2. GRAMMAR	.866	1.000	.681	.626	.690	.712	.726	.629	.713
3. RECOGNITION	.696	.728	1.000	.678	.639	.646	.644	.754	.682
4. Content	.755	.662	.728	1.000	.977	.972	.945	.807	.980
5. Organization	.778	.736	.691	1.000	1.000	.973	.958	.823	.986
6. Vocabulary	.787	.751	.691	1.000	1.000	1.000	.966	.831	.987
7. Language Use	.786	.773	.695	1.000	1.000	1.000	1.000	.822	.976
8. Mechanics	.707	.663	.806	.854	.877	.876	.875	1.000	.882
9. ESSAY TOTAL	.781	.747	.724	1.000	1.000	1.000	1.000	.923	1.000

* All correlations significant at the $p < .01$ level (1-tailed)

** n=938 for grammar and listening correlation; n=933 for recognition & listening; n=1123 for recognition and grammar; n=234 for all other correlations involving the essay test or its subscales

Table 25a
Results of Principal Components Analysis - Original Versions

TEST or Subscale	**Factor 1**	h^2
LISTENING	.795	*.632*
GRAMMAR	.822	*.676*
RECOGNITION	.774	*.599*
Content	.946	*.895*
Organization	.954	*.910*
Vocabulary	.962	*.925*
Language Use	.955	*.912*
Mechanics	.885	*.783*
Proportion of Variance	*.792*	*Total = .792*

Table 25b

Results of the Two Factor Analysis After Varimax Rotation - Original Versions

TEST or Subscale	Factor 1	Factor 2	h^2
LISTENING	.382	***.799***	*.784*
GRAMMAR	.390	***.832***	*.844*
RECOGNITION	.357	***.798***	*.764*
Content	***.909***	.376	*.968*
Organization	***.906***	.392	*.975*
Vocabulary	***.895***	.418	*.976*
Language Use	***.878***	.429	*.955*
Mechanics	***.733***	.499	*.786*
Proportion of Variance	*.523*	*.359*	*Total = .882*

Table 25c

Results of the Three Factor Analysis After Varimax Rotation - Original Versions

TEST or Subscale	Factor 1	Factor 2	Factor 3	h^2
LISTENING	.367	***.843***	.227	*.897*
GRAMMAR	.364	***.791***	.356	*.885*
RECOGNITION	.284	.389	***.855***	*.963*
Content	***.891***	.335	.251	*.969*
Organization	***.887***	.344	.267	*.976*
Vocabulary	***.876***	.367	.278	*.979*
Language Use	***.858***	.378	.280	*.957*
Mechanics	***.682***	.209	***.623***	*.897*
Proportion of Variance	*.487*	*.255*	*.198*	*Total = .941*

Table 25d

Results of the Two Factor Analysis After Varimax Rotation - Original Versions

TEST or Subscale	Factor 1	Factor 2	Factor 3	Factor 4	h^2
LISTENING	.362	.262	***.845***	.289	*.997*
GRAMMAR	.383	.335	.409	***.740***	*.974*
RECOGNITION	.289	***.851***	.219	.332	*.966*
Content	***.895***	.252	.236	.225	*.971*
Organization	***.891***	.269	.252	.219	*.978*
Vocabulary	***.879***	.281	.268	.235	*.979*
Language Use	***.863***	.279	.254	.271	*.961*
Mechanics	***.675***	***.641***	.254	.002	*.931*
Proportion of Variance	*.491*	*.201*	*.156*	*.121*	*Total = .969*

Table 26a
Results of Principal Components Analysis - Revised Versions

TEST or Subscale	**Factor 1**	h^2
LISTENING	.788	*.601*
GRAMMAR	.818	*.669*
RECOGNITION	.777	*.604*
Content	.947	*.897*
Organization	.956	*.914*
Vocabulary	.964	*.929*
Language Use	.956	*.914*
Mechanics	.886	*.785*
Proportion of Variance	*.789*	*Total = .789*

Table 26b

Results of the Two Factor Analysis After Varimax Rotation - Revised Versions

TEST or Subscale	Factor 1	Factor 2	h^2
LISTENING	.371	***.800***	*.778*
GRAMMAR	.389	***.826***	*.834*
RECOGNITION	.377	***.775***	*.743*
Content	***.906***	.381	*.966*
Organization	***.902***	.401	*.974*
Vocabulary	***.891***	.427	*.976*
Language Use	***.874***	.437	*.955*
Mechanics	***.724***	.513	*.787*
Proportion of Variance	*.518*	*.358*	*Total = .877*

Table 26c

Results of the Three Factor Analysis After Varimax Rotation - Revised Versions

TEST or Subscale	Factor 1	Factor 2	Factor 3	h^2
LISTENING	.362	***.845***	.219	*.893*
GRAMMAR	.365	***.779***	.359	*.869*
RECOGNITION	.296	.359	***.863***	*.961*
Content	***.888***	.332	.264	*.968*
Organization	***.883***	.344	.281	*.977*
Vocabulary	***.871***	.370	.289	*.979*
Language Use	***.854***	.382	.288	*.958*
Mechanics	***.668***	.230	***.620***	*.884*
Proportion of Variance	*.482*	*.252*	*.203*	

936

Table 26d

Results of the Two Factor Analysis After Varimax Rotation - Revised Versions

TEST or Subscale	Factor 1	Factor 2	Factor 3	Factor 4	h^2
LISTENING	.358	.252	***.848***	.295	*.998*
GRAMMAR	.382	.322	.375	***.768***	*.980*
RECOGNITION	.299	***.856***	.204	.316	*.964*
Content	***.891***	.263	.236	.226	*.970*
Organization	***.886***	.280	.250	.228	*.978*
Vocabulary	***.874***	.288	.265	.250	*.980*
Language Use	***.859***	.283	.252	.285	*.963*
Mechanics	***.663***	***.638***	.260	.040	*.916*
Proportion of Variance	*.485*	*.202*	*.153*	*.128*	*Total = .968*

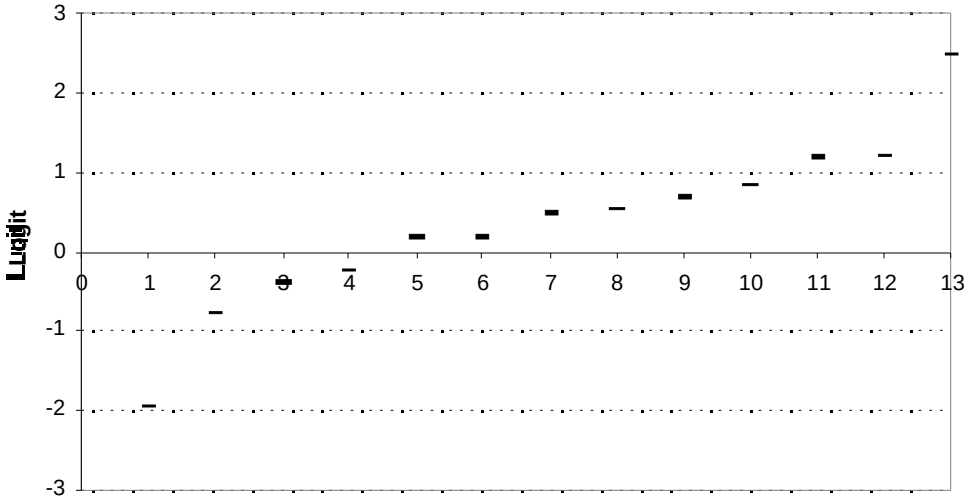

Figure 1. Difficulty estimates for 13 items in the revised listening test

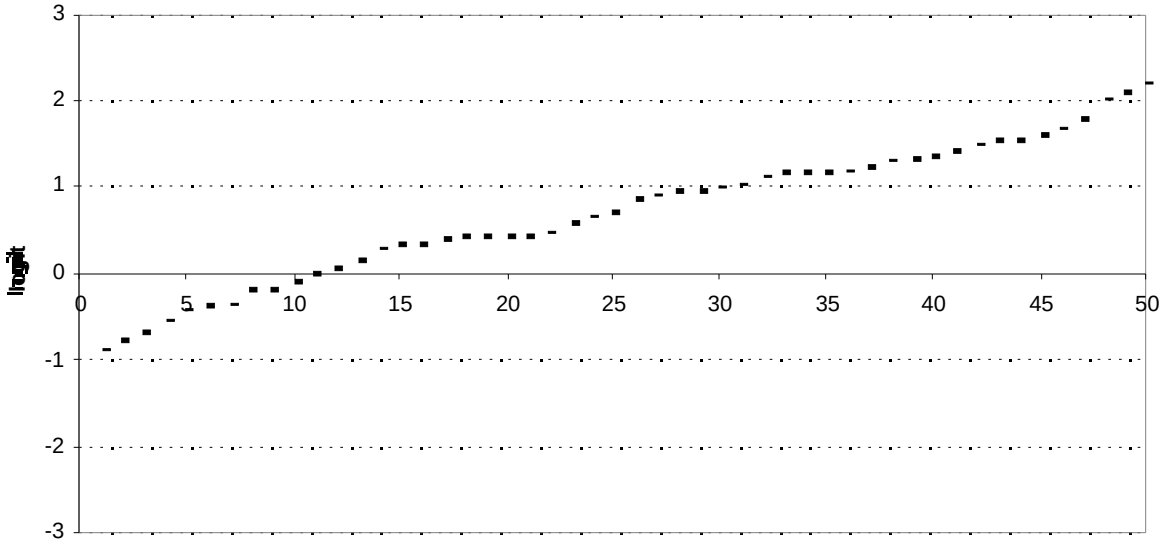

Figure 2. Difficulty estimates for 50 items in the revised grammar test

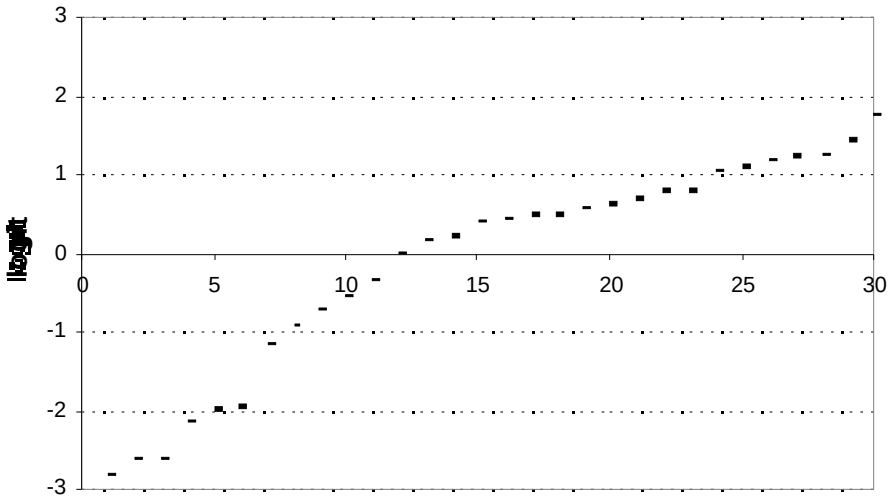

Figure 3. Difficulty estimates for 30 items in the revised recognition test

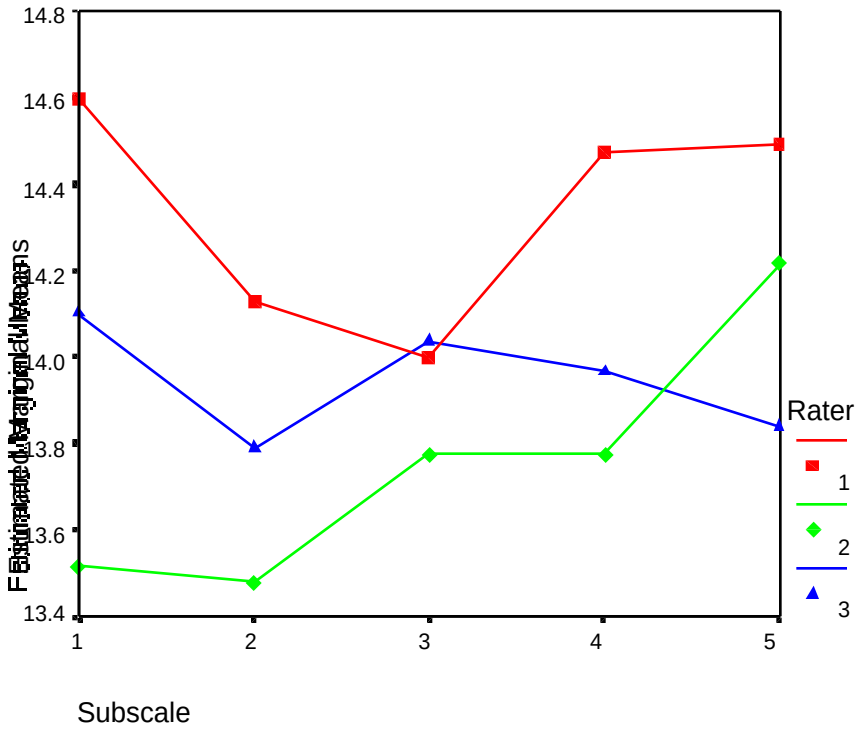

Figure 4. Interaction of raters by subscales

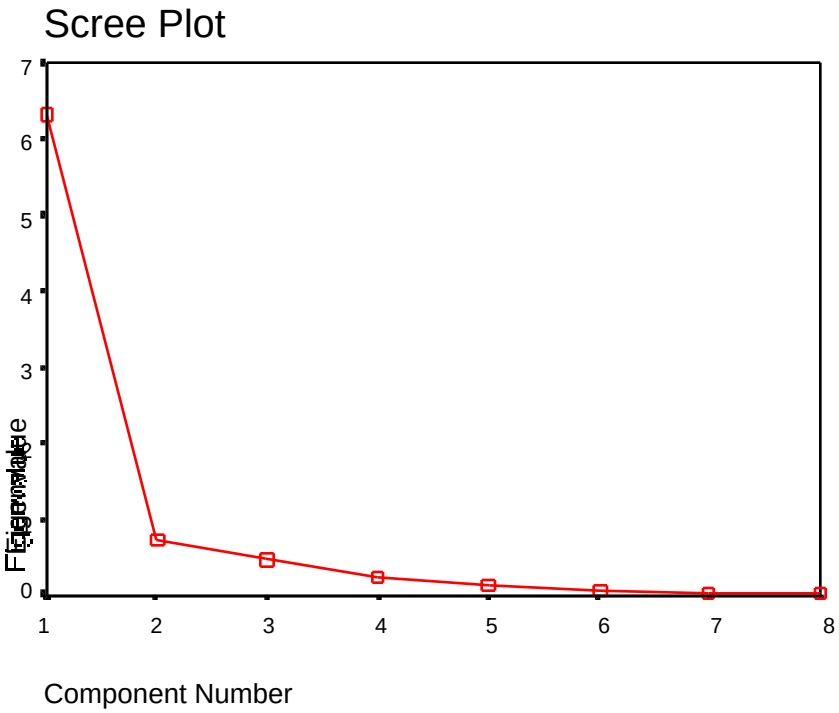

Figure 5. Scree plot for factor analyses of all original JPT tests and subscales

ENDNOTES

[1] This research was funded by the Educational Improvement Fund at the University of Hawai‘i. For those interested, a summary version of this report will soon appear in Kondo-Brown & Brown, forthcoming.

[2] The listening test originally had 15 items, but one of the items (item 13) had been eliminated at the time of this analysis. Hence, the current version has a total of 14 items.

[3] Note that a total of two ANOVA procedures were performed in this study, and as a result, following the Bonferoni procedure, the .005 decision level was used for the two sets of comparison-wise decisions in this study in order to maintain an overall experiment-wise alpha level of .01.

[4] Note again that this .005 comparison-wise alpha level was needed to maintain the .01 experiment-wise alpha level set for this study (given the fact that two ANOVA procedures were performed).