DOCUMENT RESUME

ED 366 172 FL 021 390

AUTHOR Brown, James Dean; Ross, Jacqueline A.
TITLE Decision Dependability of Subtests, Tests, and the Overall TOEFL Test Battery.
INSTITUTION Educational Testing Service, Princeton, N.J.; Hawaii Univ., Manoa.
PUB DATE Apr 93
NOTE 56p.; Paper presented at the Annual Meeting of the Language Testing Research Colloquium (Cambridge, England, United Kingdom, April 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS English (Second Language); Language Tests; Statistical Analysis; *Test Reliability; Test Theory; *Test Validity
IDENTIFIERS *Test of English as a Foreign Language

ABSTRACT

This study investigates the Test of English as a Foreign Language (TOEFL), in particular the relative contributions to score dependability (analogous to classical theory reliability) of various numbers of items and subtests as well as the decision dependability at different cut points. Research questions that apply to the overall TOEFL battery and to its tests and subtests address classical theory reliability estimates; relative contributions to error variance of persons, items, subtests, and their interactions; dependability for varying numbers of items and subtests; and the effect on score dependability of various cutpoints. The study was based on the item responses of 20,000 test takers from 15 different language backgrounds. Data were collected from the May 1991 administration of the TOEFL at foreign and domestic test centers. Analyses included descriptive statistics, classical theory reliability estimates, generalizability theory, and decision dependability estimates for various cut points. Test dependability analyses indicate that subtests can make substantial contributions to the variance of test scores and thus may affect dependability in important ways. However, these results also make it clear that, in some cases, subtests may have a negligible impact on dependability. Thus, while inclusion of subtests or the expansion of the number of subtest on a test may have a substantial beneficial effect on the dependability of the scores on that test, this relationship cannot be taken as a forgone conclusion. Findings also indicate that on the present TOEFL the lowest dependabilities along the range are still very high. Analyses are appended. (Contains 20 references.) (AA)

AUTHORS: **JAMES DEAN BROWN**

AFFILIATION: UNIVERSITY OF HAWAII AT MANOA

ADDRESS:
DEPARTMENT OF ESL
UNIVERSITY OF HAWAII AT MANOA
1890 EAST-WEST ROAD
HONOLULU, HI 96822
WORK PHONE: 808-956-8610
WORK FAX: 808-956-2802
E-MAIL: brownj@uhunix.uhcc.hawaii.edu

JACQUELINE A. ROSS

AFFILIATION: EDUCATIONAL TESTING SERVICE

ADDRESS:
TOEFL PROGRAM - 36V
EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541
USA
WORK PHONE: 1-609-951-1657
WORK FAX: 1-609-520-1093
E-MAIL: jar5501@ets

TITLE: DECISION DEPENDABILITY OF SUBTESTS, TESTS, AND THE OVERALL TOEFL TEST BATTERY

FORMAT: PAPER

DECISION DEPENDABILITY OF SUBTESTS, TESTS, AND THE OVERALL TOEFL TEST BATTERY

James Dean Brown
University of Hawaii at Manoa

Jacqueline A. Ross
Educational Testing Service

ABSTRACT

The reliability of the TOEFL battery total scores and those for each of the tests involved have repeatedly been shown to be high. In addition, the standard error of measurement has long been used as a means for estimating the average unreliable variance across all scores. The purpose of this large-scale study was to examine the reliability and dependability of the TOEFL test battery in a number of new ways. In the process, we wanted to investigate the relative contributions to score dependability (which is analogous to classical theory reliability) of various numbers of items and subtests as well as the decision dependability at different cut points. To achieve the above goals, four research questions were formulated. These research questions apply not only to the overall TOEFL battery, but also to the various tests and subtests that it includes:

1. What are the classical theory reliability estimates?
2. What are the relative contributions to error variance of persons, items, subtests, and their interactions?
3. What is the dependability for varying numbers of items and subtests?
4. What is the effect on score dependability of various cut-points?

The study was based on the item responses of 20,000 test takers from 15 different language backgrounds. The data were collected from the May 1991 administration of the TOEFL at foreign and domestic test centers. The first test in the TOEFL battery is a listening test including three item types: statement items, dialogue-based items, and minitalk items. The second test covers two item types: structure and written expression. The third test consists of vocabulary and reading comprehension items.

The analyses included descriptive statistics, classical theory reliability estimates, generalizability theory, and decision dependability [phi(lambda)] estimates for various cut points. The implications are discussed in terms of the dependability of using various combinations of TOEFL total, test, and subtest scores, as well as the dependability of decisions made at various cut points. Such issues are important because high decision dependability is a precondition for attaining high "systemic" validity.

INTRODUCTION

Scores obtained on the Test of English as a Foreign Language (TOEFL) are frequently used to inform decisions regarding the readiness of nonnative speakers to pursue academic studies in English at colleges and universities in the United States and Canada. As in all measurement, the reliability of the test instrument and the dependability of decisions made on the basis of test performance are of major concern to test developers and test score users. The internal consistency reliability of the TOEFL total and individual test scores has been shown to be high (based on either a classical theory approach or an item response theory approach), and the associated standard error of measurement is published as a means for decision makers to estimate the probable extent of error inherent in the test scores (ETS, 1992, pp. 30-31).

One useful extension of the classical theory approach to estimating the reliability of measurement was provided by with the introduction of generalizability theory by Cronbach, Rajaratnam, and Gleser (1963). In their model, reliability "resolves into a question of the accuracy of generalization, or generalizability" (Cronbach, Gleser, Nanda, & Rajaratnam, 1972, p. 15), i.e., how well one can generalize from one observation to a universe of observations. Generalizability (G) theory views the observed score as if it were the *universe score*, generalizing from the sample to the universe of interest by means of specified estimation procedures (Shavelson & Webb, 1981, pp. 133-137).

As Suen (1990, pp. 41-42) puts it, generalizability theory provides a conceptual framework to assess multiple sources and magnitudes of variation, or measurement error, within the context of a testing situation. In essence, analysis of variance (ANOVA) techniques are used to estimate components of variance associated with the various facets of measurement in a generalizability (G-study). The ability to examine the sources of error in a multifaceted way provides a more comprehensive and differentiated explanation of variance than is possible in classical reliability theory (Shavelson & Webb, 1981, p. 133). This information can be utilized in a decision study (D-study) wherein the results of various measurement designs can be manipulated. Test-design and score-use decisions can then be made that are based on a more accurate estimation of the error inherent in such choices. In turn, the dependability (analogous to reliability in classical theory) of such decisions can also be examined. All of these G-study and D-study techniques are amply demonstrated and exemplified for various statistical designs in Brennan (1983).

Application of G theory to language testing situations is discussed in Bolus, Hinofotis, and Bailey (1982) who further iterate the usefulness of this systematic approach to the study of measurement error. Brown (1984) applied G theory to the study of numbers of items and passages used in measuring engineering English reading ability in EFL situations. Then Brown and Bailey (1984) studied the effect of numbers of raters and scoring categories on the dependability of writing scores. More

recently, Stansfield and Kenyon (1992) applied G theory to the study of the effect of numbers of tests and raters on of oral proficiency interview scores. Brown (1990, 1993) also applied G theory to the problems of estimating score dependability in criterion-referenced language tests.

Purpose

The purpose of this project is to explore two dimensions of the TOEFL that have hitherto received little attention. First, a test development policy issue will be addressed. This issue centers on deciding how many items and subtests to include on the TOEFL for maximum effectiveness. Formulas like the Spearman-Brown prophecy formula can be used to predict the effects on test reliability of different numbers of items. But, such formulas cannot help in determining the optimal combination of numbers of subtests and items that ought to be included on the TOEFL. Fortunately, generalizability theory, discussed above, is particularly well suited to addressing this issue. While this project was primarily designed to investigate the TOEFL test as it exists, it is possible within the generalizability theory framework to also include analyses that allow the results to be generalized to future versions of the TOEFL (e.g., TOEFL 2000) and to other test development projects around the world.

Second, while Educational Testing Service (ETS) has long reported the standard error of measurement (SEM) for the TOEFL to help score users make responsible decisions, there is one issue that continues to be potentially troublesome: in general, tests

are not equally reliable for making decisions at different cut points (for an overview see Feldt & Brennan, 1989, pp. 123-124). Conditional SEM data provided by the TOEFL test analysis reports indicates that the SEM is not currently the highest at the mean of the TOEFL test. However, since the dependability of the scores has been found to be lowest at the mean elsewhere (Brennan, 1984, pp. 312-317), and since the dependability of the TOEFL along the entire range of possible decisions points has not been demonstrated, cut-point dependability seems like an important, yet unresolved, issue. The second general goal of this project, then, is to determine whether differences in dependability exist at different cut points for the total TOEFL scores (or the individual Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary test scores that make up the battery) and to examine the degree to which any such differences may affect the dependability and therefore the validity of score users' decisions.

To achieve the above goals, four research questions were formulated. These research questions apply not only to the overall TOEFL battery, but also to the various tests and sections that it includes:

1. What are the classical theory reliability estimates?
2. What are the relative contributions to error variance of persons, items, subtests, and their interactions?
3. What is the dependability for varying numbers of items and subtests?
4. What is the effect on score dependability of various cut-points?

METHODS

Subjects

The subjects in this study all come from the May 1991 administration of the TOEFL. That administration included a total of 93,960 examinees with 26,371 in the United States and Canada and 67,589 at other test centers around the world. In fall 1992, the International Testing and Training Programs Area at ETS made available a data set (known as the "Generic Data Set"), which was made up of 24,500 item response records from the May 1991 administration of the worldwide TOEFL. For the project reported here, a total of 20,000 students were randomly selected (from the 24,500 records in the generic data set) for convenience in analyzing the results.

Of the 20,000 subjects in this study 59.6 percent were male and 40.4 were female. They were involved in both domestic (26.2%) and foreign (73.8%) administrations of the TOEFL. They reported themselves to be from a total of 144 different countries including Brazil (3.1%), Cyprus (2.8%), France (6.0%), Germany (4.7%), Greece (3.7%), India (4.3%), Indonesia (8.2%), Japan (8.2%), Jordan (1.6%), Republic of Korea (8.1%), Lebanon (1.2%), Malaysia (4.2%), Mexico (1.3%), Pakistan (3.7%), People's Republic of China (5.1%), Saudi Arabia (1.2%), Spain (1.9%), Switzerland (1.1%), Taiwan (2.3%), Thailand (8.3%), Turkey (5.7%), and 123 other countries with one percent or less each (13.3%).

In terms of language background, the subjects reported

themselves as being speakers of Arabic (8.3%), Chinese (8.0%), French (8.0%), German (6.2%), Greek (6.2%), Indonesian (8.2%), Japanese (8.2%), Korean (8.2%), Malay (4.1%), Portuguese (4.1%), Spanish (8.1%), Telugu (4.0%), Thai (8.3%), Turkish (6.1%), and Urdu (4.1%).

Their reasons for taking the TOEFL varied too, as follows: for undergraduate studies (37.0%), for graduate studies (46.2%), for another type of school (2.0%), for a license (1.8%), for a company (8.5%), other (3.5%), and no reason given (1.0%).

Materials

As pointed out in ETS publications (e.g., ETS, 1992, 1993), the TOEFL test battery consists of three separately timed tests in multiple-choice format with four answer options for each test question printed in a test book. All responses are gridded on answer sheets that are later computer scored.

The first test, Listening Comprehension (LC Test), is designed to measure the ability to understand spoken English. The first part (LC1) requires the examinees to listen to a short sentence and to choose the option that is closest to it in meaning. The second part (LC2) consists of short conversations between two people, followed by a spoken question. The examinee decides which option best answers the question. Part 3 (LC3) presents several short talks and extended conversations about a variety of subjects, and requires the examinees to respond to oral questions about what they heard.

The second test, Structure and Written Expression (SWE

Test), is designed to measure the ability to recognize selected points of English structure. In the first part of this test (SWE1), the examinee reads an incomplete sentence and must choose the word or phrase that best completes it. In the second part (SWE2), several words or phrases are underlined in a sentence, and the examinees must choose the underlined segment that is not an acceptable English usage.

The third test, Vocabulary and Reading Comprehension (VRC Test), was designed to test the ability to understand the meaning and use of words as well as the ability to comprehend a variety of reading materials. The first part (VRC1) of this test contains vocabulary items wherein a word or phrase is underlined in a sentence and the examinee must select a word or phrase that could be substituted and still preserve the original meaning of the sentence. In the second part (VRC2), the examinee reads a number of short passages on a variety of academic subjects and must answer questions based on what is stated or implied in the passage.

Procedures

The TOEFL being used here was administered under standard conditions in May 1991. Strict admission procedures were followed, and, during the test, examinees were not allowed to have anything other than the testing materials on their desks. They were not permitted to take notes or make marks of any kind in their test books. Nor were they permitted to work on any section of the test before or after time was called.

After the administration, answer sheets were returned to Educational Testing Service (ETS) for scoring. The raw scores for each Test are the number of questions answered correctly. There is no penalty for guessing. Raw scores are then converted to standardized scales based on the three-parameter item response theory model (*T* scores for the individual tests and *CEEB* scores for the battery as a whole). These scaled scores are reported to the examinees and to institutions that the examimees have selected to receive scores.

Analyses

The analyses in this project began with descriptive statistics and classical theory reliability estimates (split-half adjusted, Guttman, and Cronbach alpha) to provide background and a context for interpreting the generalizability studies (G-studies) and decision studies (D-studies).

[INSERT FIGURE 1 ABOUT HERE]

Five G-studies were conducted based on the overall structure of the TOEFL shown in Figure 1. The first G-study investigated the effects on the Total TOEFL battery scores dependability of numbers of items (items facet) and numbers of test types (subtests facet based on the Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary tests) as shown in Figure 2A. The second, third, and fourth G-studies considered the effects on total test scores for the Listening Comprehension Test (LC Test), Structure and Written Expression Test (SWE Test), and Reading Comprehension and

Vocabulary Test (VRC Test) of numbers of items and subtests (made up of different item types) on those tests as shown in Figures 2B through 2D. The fifth G-study focused on the Reading Comprehension section (VRC2) of the VRC Test. In this case, the effects of numbers of items and subtests (passages P1-P5) was investigated as shown in Figure 2E.

[INSERT FIGURES 2A-2E ABOUT HERE]

All of these G-studies were very similar in structure. In all cases, and analysis of variance (ANOVA) procedures were run using all 20,000 subjects for a persons by items nested within subtests design, or p x (i:s). The result in all cases was a two facet design with items and subtests as the facets. Random effects models were used in the ANOVAs so that the results would be generalizable to the development of the TOEFL 2000 project as well as to other test development projects around the world. However, in some places mixed model ANOVAs with fixed effects for the subtest facet were also used so that the results for the current configuration of the TOEFL could be examined.

In the random effects model, it is assumed that persons, items, and subtests were randomly selected from the universes of all possible persons, items, and subtests. Shavelson and Webb (1981) argued that random effects models are reasonable if one can take an exchangeability perspective:

> Viewed from the exchangeability perspective, the issue of fixed or random effects is not whether one can catalog (etc.) all possible members of a population but whether the members are exchangeable with other potential members. In terms of sampling, if one set of persons and items to which ρ^2

> generalizability coefficient...is generalizable is the set of such persons and items jointly exchangeable with the present sample, it is reasonable to consider the item facet random. The concept of exchangeability, at the minimum, provides reasonable grounds for considering whether a facet is random or fixed.

Thus for those results in this paper that are based on a random effects model, random selection of items and subtests is assumed, while, for those results based on a mixed model (with fixed effects for subtests), no such assumption is made for the subtests facet.

Based on the mean squares obtained in the random effects model ANOVA procedures, variance components were estimated (as will be demonstrated in the **RESULTS** section). Interpreting these variance components helped in understanding the relative contribution of persons to the true score variance, as well as the contributions of items and subtests to the error variance.

Five parallel D-studies followed the G-studies. In these D-studies, the variance components found in the G-studies were used to calculate statistics that can be directly interpreted in making decisions. Two types of error were considered: a) lower-case delta error (δ) for relative decisions (i.e., norm-referenced decisions), and b) upper-case delta error (Δ) for absolute decisions (i.e., criterion-referenced). All relevant D-study statistics are reported in the **RESULTS** section for the combination of items and tests under investigation in this project. In addition, G coefficients (based on lower case delta) are reported for various numbers of items and subtests so that

the reader can directly observe the effect on dependability of these two facets in various combinations of numbers of items and subtests.

The last step in each D-study was to calculate a squared-error loss agreement coefficient known as the phi(lambda), or $\Phi(\lambda)$, at various cut points from 10 percent to 90 percent. These analyses illustrate the effect of various cut points on decision dependability. Phi(lambda) coefficients were calculated for both a random effects model (to provide generalizability of results to other tests) and a mixed model with fixed effects for subtests (to provide estimates for the TOEFL as it existed in this study).

RESULTS

The results of this project will be discussed in the following stages with commensurate section headings: a) descriptive statistics for each of the five generalizability studies will be provided for background; b) classical theory reliability estimates will be presented for later comparison with the G-theory results; c) the variance components for the five G-studies will be presented and compared; d) the five parallel D-study results will be presented along with G coefficients for various numbers of items and subtests; and finally, e) threshold-loss agreement coefficients will be given for different cut points within each of the D-studies.

Descriptive Statistics

The descriptive statistics for the raw scores involved in

each of the five generalizability studies are reported in Table 1. According to the labels across the top of the table, the mean, standard deviation (SD), and number of items (k) are given for the original test and for the G-study sampling. The original test includes the subtests and numbers of items just as they were administered. The G-study sampling results are based on the random samples that were taken from the original test to create balanced subtests (each containing the same number of items) for the generalizability studies.

[INSERT TABLE 1 ABOUT HERE]

The first G-study was on the effects of items and tests on the dependability of Total TOEFL battery scores. Thus descriptive statistics are given for the Total TOEFL and each of the tests which contribute to that total score: Listening Comprehension (LC Test), Structure and Written Expression (SWE Test), and Reading Comprehension and Vocabulary (VRC Test). Notice that the original TOEFL had a total of 146 items and that the original LC, SWE, and VRC tests had 50, 38, and 58 items, respectively. In order to create a balanced design, two of the tests had to be reduced in number of items to match the smallest of the tests. To achieve this, 38 items were randomly selected from the LC and VRC tests to match the existing 38 items in the SWE Test. As a result, in the first G-study, all three Tests were analyzed as 38 item tests with a TOEFL total of 114 items.

The second G-study was focused on the effects of items and subtests on the dependability of LC Test scores. Thus

descriptive statistics are given in Table 1 for the whole LC Test and each of the subtests which contribute to the LC Test scores: LC1, LC2, and LC3. Notice that the original LC Test had a total of 50 items and that the original LC1, LC2, and LC3 sections had 20, 15, and 15 items, respectively. In order to create a balanced design, the longer section had to be reduced in number of items to match the other two sections. To achieve this, 15 items were randomly selected from the LC1 to match the existing 15 items in both the LC2 and LC3 sections. As a result, in the second G-study, all three sections were analyzed as 15 item subtests with an LC Test total of 45 items.

The third G-study was on the effects of items and subtests on the dependability of SWE Test scores. Thus descriptive statistics are given for the whole SWE Test and each of the two sections which contribute to the SWE Test scores: SWE1 and SWE2. Notice that the original SWE Test had a total of 38 items and that the original SWE1 and SWE2 sections had 14 and 24 items, respectively. In order to create a balanced design, 14 items were randomly selected from the SWE2 section to match the existing 14 items in the SWE1 section. As a result, in the third G-study, the two sections were analyzed as 14 item subtests with a SWE Test total of 28 items.

The fourth G-study was on the effects of items and subtests on the dependability of VRC Test scores. Thus descriptive statistics are given for the whole VRC Test and each of the two sections which contribute to the VRC Test scores: Vocabulary and

Reading Comprehension. Notice that the original VRC Test had a total of 58 items and, since each of the sections had 29 items, it was already balanced. Thus no modifications were necessary in preparing it for the fourth G-study.

The fifth G-study was on the effects of items and passages within the Reading Comprehension section (VRC2) on the dependability of VRC2 section scores. Thus descriptive statistics are given for the whole VRC2 section and the items associated with each of the passages which contributed to the VRC2 section scores: Passages 1 to 5. Notice that the original VRC2 section had a total of 29 items and that the original passages had 7, 5, 7, 6, and 4 items associated with them, respectively. In order to create a balanced design, the passages with larger numbers of items had to have the number of items reduced to match the shortest passage (i.e., Passage 5 with four items). To achieve this, four items were randomly selected from those associated with each of the larger passages. As a result, in the fifth G-study, all five passages were analyzed as four item sections with a VRC2 section total of 20 items.

Classical Theory Reliability

Classical theory reliability estimates are presented in Table 2. For ease of interpretation, Table 2 is organized in the same general manner as Table 1. The first classical theory reliability estimate given is the split-half correlation adjusted by the Spearman-Brown prophecy formula. Then the Guttman reliability is given followed by the Cronbach alpha coefficient.

Notice that the first two estimates are consistently lower than the Cronbach alpha coefficients. Since theory indicates that the first two are more likely to be underestimates, the single best estimate is the Cronbach alpha. These estimates are given for the Original Tests and the G-study Samplings (along with the numbers of items, or *k*) so that the effect of the reductions in test length on classical theory reliability can readily be seen.

[INSERT TABLE 2 ABOUT HERE]

Variance Components

Based on ANOVA procedures (shown in Appendix A), G theory allowed for estimation of the relative contributions of persons, items, and subtests in terms of variance components. For example, for the first G-study of the Total TOEFL Battery, which was a p x (i:s) design (like all of the others), the ANOVA results are shown in Table 3.

[INSERT TABLE 3 ABOUT HERE]

Based on the variance components that make up the estimated mean squares (*EMS*) as shown in Brennan (1983) or Kirk (1968), the variance components for persons as well as for the items and subtests facets were isolated from the observed mean squares (*MS*). The *EMS* shown in Table 3 were used systematically to derive the variance components in a step-by-step manner. First, because the estimated variance component for the interaction of persons and items nested within subtests, or $\hat{\sigma}^2(pi:s)$, is equal to the *MS(pi:s)* for that interaction, .16180465 in this case, that variance component is easy to isolate. Formulaically, this

process can be summarized as follows:

$\hat{\sigma}^2(pi:s) = MS(pi:s)$

Second, because, as is shown in Table 3, it is known that the *EMS* for the ps interaction = $\sigma^2(pi:s) + n_i\sigma^2(ps)$, the estimated variance component for this interaction, $\hat{\sigma}^2(ps)$, could be isolated by subtracting the $MS(pi:s)$ from the $MS(ps)$, and dividing the result by the number of items, n_i, in each subtest [i.e., (.43545427 -.16180465)/38 in this case]. Formulaically:

$\hat{\sigma}^2(ps) = [MS(ps) - MS(pi:s)]/n_i$

Third, fourth, and fifth, using the known mathematical relationships shown in Table 3, the other three variance components in this design could then be calculated by using the following formulas:

$\hat{\sigma}^2(p) = [MS(p) - MS(ps)]/n_i n_s$

$\hat{\sigma}^2(i:s) = [MS(i:s) - MS(pi:s)]/n_p$

$\hat{\sigma}^2(s) = [MS(s) - MS(i:s) - MS(ps) + MS(pi:s)]/n_p n_i$

Note that the calculations in this example were based on MS values that have been rounded to eight places. Because the resulting variance components are often very small values, it was essential that nothing be rounded any more than was necessary until the final result was obtained.

[INSERT TABLE 4 ABOUT HERE]

The variance components for each of the G-studies in this project (all calculated in similar manner) are shown in Table 4. Notice that the five G-studies are labeled across the top as columns and that the sources of variance (p, s, i:s, ps, and

pi:s) are labeled at the left as rows. The totals in the last row represent the sums of the variance components isolated in each study.

D-Study Results and Generalizability Coefficients

Summaries of the statistics found in the five D-studies are presented in Table 5. Notice that each D-study is presented in a separate column as labeled across the top of the table. The rows represent each of the statistics. First, the number of subtests (n_s) is given, then the number of items per subtest (n_i), then the total number of items (when the number of subtests is multiplied times the number of items per subtest). Then the estimated variance components (adjusted for the number of items and subtests in the particular D-study) are given for *p*, *s*, *i:s*, and their interactions. Notice that the variance components for *p* are the same as those reported in Table 4, while the variance components for *s*, *i:s*, and their interactions are different in the two tables because those in Table 5 have been adjusted for the numbers of items or subtests in the particular D-study design (after Brennan, 1983). In the next row, the mean proportion scores ($\bar{X}_p$) are given. These means are simply the average of each persons proportion score, which is calculated by dividing the number of correct responses by the number of items (but not moving the decimal two places to the right as would be done in calculating a percent score).

[INSERT TABLE 5 ABOUT HERE]

Next, statistics are given for a random effects model. The

random effects model estimates allow generalization of the results to other tests as discussed above. The statistics for this model include $\sigma^2(\tau)$, which is just another expression of $\sigma^2(p)$. The upper-case delta error term, $\sigma^2(\delta)$, (for relative decisions, i.e., norm-referenced interpretations) and the lower-case delta error term, $\sigma^2(\Delta)$ (for absolute decisions, i.e., criterion-referenced or domain referenced interpretations) are also given. Then the expected observed score variance, $E\sigma^2(X)$, and error variance associated with the grand mean, $\sigma^2(\bar{X})$, are presented. All of these statistics were used in calculating the generalizability coefficients for lower-case delta (norm-referenced) error, $E\rho^2(\delta)$, in the S/N ratios reported in this table, or in the phi(lambda) coefficients reported in the next section.

The G-coefficients, $E\rho^2(\delta)$ that are presented in Table 5 are analogous in interpretation to reliability coefficients. They are calculated by forming a ratio of the persons variance component for the particular number of subtests and items in the G-study over the same persons variance plus the appropriate error term. Thus G-coefficients for relative decisions would use δ error as follows:

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

Similarly, G-coefficients for absolute decisions would use Δ

error as follows:

$$E\rho^2(\Delta) = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)} = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\Delta)}$$

The last statistic presented in the Random Effects part of Table 5 is the signal to noise ratio (S/N). This statistic can be interpreted as the ratio of systematic variance to random error (Brennan & Kane, 1977), or, as Cronbach and Gleser (1964: 468) put it in an earlier discussion of communications systems, the "signal to noise ratio compares the strength of the transmission to the strength of the interference."

At the bottom of the table, the same statistics are presented for a mixed effects model (with subtests as a fixed effect). These results can only be generalized to the TOEFL battery as it was structured and studied here.

Notice that, as would be expected, the generalizability (or G) coefficients [$E\rho^2(\delta)$] for the mixed model are very similar to the Cronbach alpha values reported in Table 2 for the G-study sampling (.9584, .9077, .8686, .9326, and .8280, respectively). In addition, probably because of differences in numbers of items, these G-coefficients are slightly lower than the corresponding Cronbach alpha values reported in Table 2 for the Original test (.9667, .9178, .9016, .9326, and .8769, respectively).

Naturally, the G-coefficients for the random effects model are more conservative than those for the mixed model because the random effects statistics can be generalized beyond the items and subtests of the current TOEFL to other batteries and tests.

Tables 6 to 10 were created by expanding this random effects G-coefficient information. Each of these tables corresponds to one of the D-studies in this project and gives the coefficients that would arise from different numbers of items and subtests.

[INSERT TABLE 6 ABOUT HERE]

For instance, Table 6 is for the Total TOEFL battery and shows that the G-coefficient for 3 subtests with 38 items each (see the point where the 38th row and third column of coefficients intersect) is .892 which is equivalent (though rounded) to the random effects model G-coefficient of .8916 reported in Table 5. Notice in the bottom left corner of the table that the battery configured with the same 114 items but in one subtest instead of three is estimated to be dependable at .785, with two subtests of 57 (total 114), it is predicted to be .862 and, with three subtests of 38 (as shown at the top of this paragraph), it would be .892. Thus the effects of having the items divided up into smaller and smaller subtests are demonstrated.

Clearly, there is considerable dependability gained from having the TOEFL battery made up of three different subtests rather than of one long, homogeneous test. In other words, there is an increase in dependability due to increases in the number of subtests involved while holding the number of items constant. Moreover, these increases are above and beyond predictions that could be made by using formulas like the Spearman-Brown prophecy formula used in classical theory reliability studies.

Table 6 also allows for considering other potential combinations of numbers of items and subtests as part of the D-study to help in deciding what is the optimal number of items and subtests to include in future versions of this and other tests. For instance, by looking at the point where six subtests intersects with 19 items (also for a total of 114 items), the table reveals that a G-coefficient of .923 is predicted.

However, for actual policy decisions, factors other than dependability must come into play. For instance, 100 tests with seven items each are predicted to be dependable at .99, but such a 700 item test is not practical even though the dependability would be near perfect. Thus these dependability estimates for various numbers of items and subtests are meant to provide one piece of information among the many types of information that must be considered in making test development decisions.

[INSERT TABLE 7 ABOUT HERE]

Turning to Table 7 for the LC Test, notice that a single 45 item test would be dependable at .882, while a similar 45 item test based on three subtests of 15 items each would only be slightly more dependable at .899, and a 45 item test based on five subtests of nine items each would only gain .004 points at .903. Thus the pay off in terms of gains in dependability due to increases in the number of subtests (while items are held constant) appear to be minimal for the LC Test.

[INSERT TABLE 8 ABOUT HERE]

Similarly, in Table 8 for the SWE Test, a 28 item test with

only one subtest would be dependable at .836, while a similar 28 item test based on two subtests of 14 items each would only be slightly more dependable at .851, four subtests of seven items each would only be .859, seven subtest of four items each would only be .862, and fourteen subtests of two items each would be .865. In short, there is not nearly as much to gain by dividing the SWE Test into subtests -- certainly not beyond two subtests -- as there was in the Total TOEFL battery.

[INSERT TABLE 9 ABOUT HERE]

Table 9 for the VRC Test is somewhat different. The table seems to indicate that considerable dependability is gained by splitting the 58 items into two subtests, i.e., the one-subtest, 58-item dependability is .858, while the two-subtests version (of 29 items each) dependability is considerably higher at .893. However, a three-subtests version (of 20 items each) would only increase to .907 even though it is two items longer, and a four-subtests version (of 15 items each) would only increase further to .914. Thus, like the SWE Test results, it appears that the present two-subtest version of the VRC Test may include as many subtests as are necessary and practical.

[INSERT TABLE 10 ABOUT HERE]

Table 10 for the VRC2 section is more like the Total TOEFL in terms of the impact of subtests on dependability. For instance, the one-subtest, 20-item dependability is .650, while the two-subtest version (with 10 items each) is considerably higher at .729, and the four-subtest version (with 5 items each)

climbs to .776. Thus differences in the numbers of passages involved in VRC2 section appear to be relatively important to its overall dependability.

More D-study Results: Phi(lambda) Dependability Coefficients

Threshold loss agreement coefficients focus on the degree to which classifications in clear-cut categories have been consistent. Since it is known that such dependability may vary at different cut points (Brennan 1980, 1984) and since TOEFL is widely used as an admissions tool for admit/no-admit decisions (though at different cut points), one of the research questions in this study was the degree to which the dependability of TOEFL changes over the range of possible cut points.

[INSERT TABLE 11 ABOUT HERE]

Table 11 gives the Phi(lambda), or $\Phi(\lambda)$, coefficients for various cut points (in percentage terms). In all cases, these coefficients are based on the p x i:s design and (Δ) error (as suggested by Brennan, 1984) and are therefore more conservative than the (δ) error estimates would have been. Notice that such coefficients are reported for both random effects models and mixed models (with subtests as a fixed effects facet). In each set, the lowest value reported was that for a cut point at the mean. Hence the $\Phi(\lambda)$ values for the cut point at the mean $\Phi(\bar{X})$ are reported below all of the others in each type of model and the mean percentages (upon which the $\Phi(\lambda)$ values are based) are given for reference.

To interpret this table, it is first necessary to decide whether it is results that are generalizable to other tests that are of interest (Random Effects Model), or results that pertain only to the present TOEFL items and subtests that are of interest (Mixed Effects Model). Consider the Mixed Effects Model for the present TOEFL battery as a whole presented in the bottom half of the first column. Notice that $\Phi(\lambda)$ coefficients are presented for decisions made at 10%, 20%, etc. up to 90%. Notice further that the lowest of these is .957 at the 70% cut point. It turns out here and in the other columns that the lowest value will be that closest to the mean. In fact, decisions made at the mean will generally turn out to be the least dependable. Hence, the $\Phi(\lambda)$ at the mean is presented along with that mean in the last two rows of both the upper and lower portions of Table 11.

DISCUSSION

In interpreting the above results, it is important to remember that most of the dependability estimates (i.e., all except those found in the VRC Test analyses in Study Four) are based on fewer items than actually used in the tests because it was necessary to design the various studies so that there would be equal numbers of items on each subtest. Since shorter tests tend to be less reliable, the effect of these reduced numbers of items (if there is any) would be to provide low estimates of dependability. As a result, it is reasonable to interpret the results as conservative underestimates of the true state of affairs. In other words, if the dependability estimates are in

error, they will err on the low side and should not provide overestimates of the dependability of these measures.

The remainder of this discussion will directly address the original research questions posed at the outset of this project. To help organize the discussion, the research questions will be used as headings.

What are the classical theory reliability estimates?

As reported elsewhere in the literature, the Total TOEFL battery and its component tests -- the LC Test, SWE Test, and VRC Test -- proved to be very reliable from a classical theory perspective. The results in Table 2 indicate that these tests in their existing form (labeled Original Test in the table) were reliable at .97, .92, .90, and .93, respectively, using Cronbach alpha. Predictably, the VRC2 section, which was only a portion of the VRC Test, was less reliable at .8769 than the tests and battery considered above because it is considerably shorter than they are. For the sake of comparison, Table 2 also presents the classical theory estimates for the items used in the G-study sampling (done to create balanced designs). These Cronbach alpha estimates later turned out to be comparable to the G-coefficients (for δ error) for the mixed models as would be expected.

What are the relative contributions to error variance of persons, items, subtests, and their interactions?

Examining the variance components shown in Table 4 for the five G-studies in terms of their relative magnitude reveals the relative contributions of persons, subtests, and items nested

within subtests, as well as their interactions. For instance from inspection of the variance components themselves, it is clear that the lion's share of variance in all of these studies is taken up by persons and those interactions involving persons. This is as it should be because the purpose of a norm-referenced test is to differentiate among persons. However, it should be noted that the variance component for the persons by subtests interaction is far smaller than that for the persons by items nested within subtests interaction -- though the persons by subtests interaction is fairly high in Study Five. It is also true in all cases that the variance component due to items nested within subtests is far larger than the component for subtests. Particularly in Study Two (LC Test) and Study Three (SWE Test), the subtests variance component is very small. The subtests component is somewhat larger in Study Four (VRC Test). However, in Study One (Total TOEFL), the variance component for subtests is much more important, amounting to about one-twelfth of the persons component and about one-quarter of the items nested within subtests component. In Study Five (VRC2), the subtests variance component is even more important since it is almost one-fifth as large as the persons component and almost equal in magnitude to the items nested within subtests component. These observations will be further illuminated in the next section.

What is the dependability for varying numbers of items and subtests?

Tables 6 to 10 provided a multitude of direct answers to

this research question. In all cases, the subtests facet was shown to have some influence on the predicted dependability indices as indicated by the fact that in no D-study was the dependability the same for one subtest and more than one subtest with the number of items held constant. In other words, in all cases the dependability was enhanced by having an increased number of subtests even though the number of items was kept the same.

However, a pattern emerged in examining the results across tables which was consistent with the variance component findings in the previous section. The influence of subtests was greatest in the Studies One (Total TOEFL) and Five (VRC2), and to a lesser degree in Study Four (VRC Test). In considering the Total TOEFL results, it might at first glance appear that the affect would be larger here than in the other studies because the length of the subtests themselves were longer at 38 items each than in any of the other studies. However, this reasoning is contradicted by the fact that an even larger effect for subtests was found in the VRC2 results, which was based on four items in each subtest -- the smallest number of items per subtest reported in any of the D-studies in this project.

It should be noted that Studies One and Five were quite different from each other in structure. The relatively large differences in dependability due to subtests in Study One were due to differences between tests (i.e., the LC Test, SWE Test, and VRC Test), while those observed for Study Five were due to

differences between reading passages (i.e., Passages 1 to 5).

What is the effect on score dependability of various cut-points?

The results shown in Table 11 for D-studies One, Two, and Four indicate that, for the existing (i.e., using a Mixed Effects Model) Total TOEFL battery, LC Test and VRC Test, the dependability of decisions is not greatly different at various cut points, and in any case, at the lowest point they are acceptably dependable (at .957, .904, and .928, respectively). The third D-study indicates that the dependability at the mean is more markedly different at .861 from the dependabilities at other cut points. Thus, though .861 is not problematic dependability, it would be most responsible to apply additional caution in interpreting decisions on the SWE test that are at or near the mean (approximately 50 on the standardized scores). [It is also important to note that the .861 found here is probably an underestimate of the existing state of affairs because it is based on two subtests of 14 items while the original subtest was based on 2 subtests of 14 and 24 items.] In short, decisions based on the current TOEFL battery and individual tests of the TOEFL can still be considered dependable even if those decisions are made right at the mean score (of approximately 500 for the battery or approximately 50 for the separate tests).

In the upper portion of Table 11, the Random Effects Model results turned out to be more conservative than the Mixed Effects results, showing both lower dependability in general and a more marked decline in the dependability at and near the mean. Recall

that this difference was expected due to the fact that these results are generalizable to other test development projects.

CONCLUSION

Test Dependability

The effects on dependability of different numbers of subtests and items (based on the random effects model) are shown in Table 4 as variance components and in Tables 6 through 10 as G-coefficients. One pattern that emerged is that the effect of having multiple tests (i.e., the subtest facet) on the Total TOEFL battery seems to have a strong beneficial effect on the dependability of scores for the Total battery. In other words, including component tests like the LC Test, SWE Test, and VRC Test in the Total TOEFL battery has proven to be a sound policy decision from the dependability perspective. In addition, based on Table 6, further policy decisions can be made about the relative merits of adding further items and/or component tests or cutting down on their numbers.

Similarly, the effect of having multiple passages (the subtest facet) in the VRC2 section seems to have a positive advantageous effect on the dependability of scores for the VRC2 section. To some degree, the effect of having both the reading comprehension and vocabulary subtests on the VRC Test also appears to have a beneficial effect on the dependability of this test -- though the strong increases in dependability do not appear to extend beyond two such subtests. In contrast, the individual subtests within the LC Test and SWE Test, while they

do make some difference, appear to have less impact on the dependability of the scores on these tests.

It is possible that in G-studies one, four and five, where the subtests facet did have an important impact, the subtests involved were significantly different from each other and thus contributed to the overall variance on the test above and beyond the contribution made by items. In contrast, in G-studies two and three, the subtests involved may be testing very much the same things.

In terms of developing future versions of the TOEFL (including TOEFL 2000) and other test development projects around the world, recall that the results presented in Tables 4, and 6-10 were for random effects models and that they were therefore generalizable to other versions of the test and other testing projects. In short, the analyses in this project indicate that subtests can make substantial contributions to the variance of test scores and thus may affect dependability in important ways. However, these results also make it clear that, in some cases, subtests may have a negligible impact on dependability. Thus, while inclusion of subtests or the expansion of the number of subtests on a test may have a substantial beneficial effect on the dependability of the scores on that test, this relationship cannot be taken as a forgone conclusion.

Decision Dependability and Validity

The results of this study are also related to the notions of decision dependability and validity. At the beginning of this

paper, concern was expressed about the possibility that test scores may not be equally reliable for making decisions at different cut points in the score range. Since the dependability of a test is often lowest at the mean and since many decisions are made at or near the mean on TOEFL, this was a legitimate concern. Portions of this project were therefore designed to examine the degree to which these differences may affect the dependability and therefore the validity of score users' decisions. The lower portion of Table 11, which reports the $\Phi(\lambda)$ coefficients when a mixed effects model is applied, indicates that on the present TOEFL the lowest dependabilities along the range are still very high. Thus, while it initially seemed like a potential problem for score users, there appears to be no need to worry about differential dependability at different cut points on the existing test. In other words, regardless of the cut point that current TOEFL score users may decide to be valid for their own reasons, the effect on dependability of various cut points is apparently not an issue of great concern. In addition, ETS is currently implementing automated item selection procedures to assemble TOEFL tests which will help to insure that each section will provide high information (or low error variance) at the middle ability range. Naturally, any such validity decisions should be also studied in the actual context(s) in which the decisions are to be made.

In terms of future versions of the TOEFL and other testing projects around the world, the upper portion of Table 11, which

reports the $\Phi(\lambda)$ coefficients when a random effects model is applied, indicates that there may be more variation in dependability estimates across the range of possible decision points. Thus, while such differential dependability is apparently not a problem on the current TOEFL, it is an issue that should continue to concern developers of other tests and future versions of the TOEFL.

Future Research

In the course of conducting this project, a number of questions have occurred to us. They are presented here in the hope that they will be investigated in the future:

1. Would similar results be obtained if these studies were replicated with other TOEFL data sets?
2. Would similar results be obtained if such studies were replicated using other tests as the basis?
3. What could generalizability theory tell us about the effects of raters on the scores of the *Test of Written English*?
4. What could generalizability theory tell us about the effects of items and raters on the scores of the *Test of Spoken English*?
5. What could be learned about the TOEFL battery and other tests by applying classical theory approaches to decision reliability/dependability at different cut points (for an overview of these approaches, see Feldt & Brennan, 1989, pp. 123-124)?

REFERENCES

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

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

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

Brennan, R.L. (1984). Estimating the dependability of the scores. In R.A. Berk (Ed.) *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University Press.

Brennan, R.L. & M.T. Kane. (1977). Signal/noise ratios for domain-referenced tests. *Psychometrika, 42,* 609-625.

Brown, J.D. (1984). A norm-referenced engineering reading test. In A.K. Pugh & J.M. Ulijn (Eds.) *Reading for professional purposes: studies and practices in native and foreign languages*. London: Heinemann Educational Books.

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

Brown, J.D. (1993). A comprehensive criterion-referenced testing project. In D. Douglas & C. Chapelle (Eds.) *A new decade of language testing research* (pp. 163-184). Alexandria, VA: TESOL.

Brown, J.D. & K.M. Bailey. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34,* 4, 21-42.

Cronbach, L.J. & G.C. Gleser (1964). The signal/noise ratio in the comparison of reliability coefficients. *Educational and Psychological Measurement, 24,* 467-480.

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

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

ETS. (1992). *TOEFL test and score manual*. Princeton, NJ: Educational Testing Service.

ETS. (1993). *Bulletin of information for TOEFL, TWE, and TSE.* Princeton, NJ: Educational Testing Service.

Feldt, L.S. & R.L. Brennan. (1989). Reliability. In R.L. Linn (Ed.). *Educational measurement* (3rd ed.). New York: Macmillan.

Hudson, T. & B. Lynch. (1984). A criterion-referenced approach to ESL achievement testing. *Language Testing, 1,* 171-201.

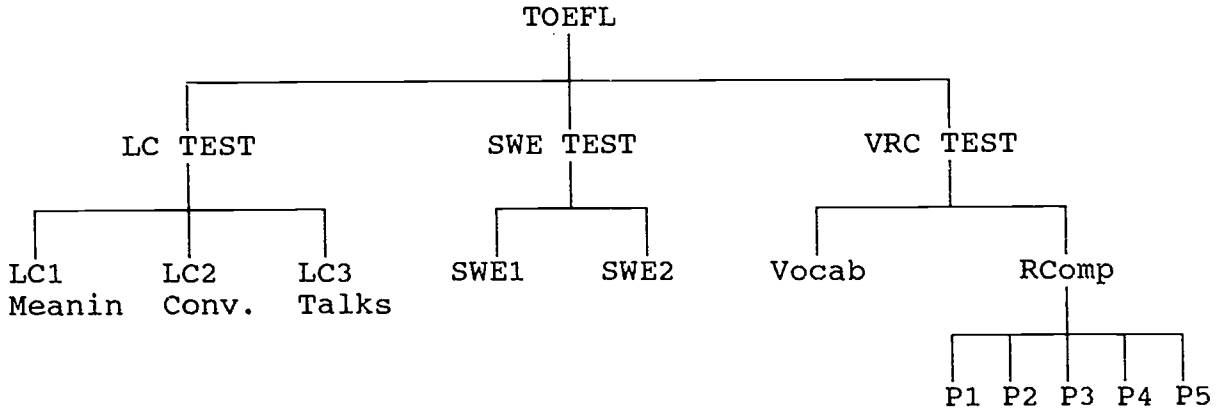
Kirk, R.E. (1968). *Experimental design: Procedures for the behavioral sciences.* Belmont, CA: Brooks/Cole.

Shavelson, R.J. & N.M. Webb (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34,* 133-166.

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

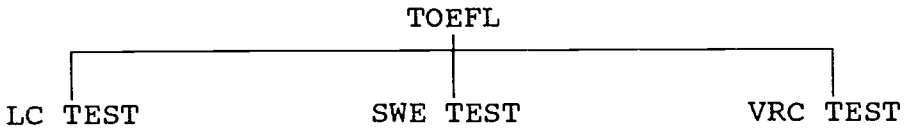
Suen, H.K. (1990). *Principles of test theories.* Hillsdale, NJ: Lawrence Erlbaum.

FIGURE 1: TOEFL STRUCTURE

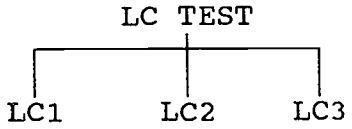


FIGURES 2A-2E: G-STUDY STRUCTURES

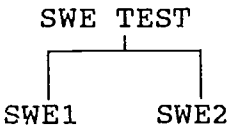
2A. G-STUDY ONE



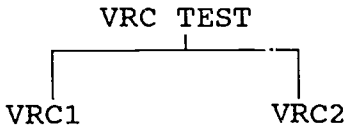
2B. G-STUDY TWO



2C. G-STUDY THREE



2D. G-STUDY FOUR



2E. G-STUDY FIVE

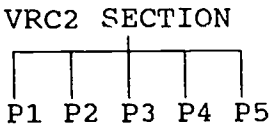


TABLE 1: DESCRIPTIVE STATISTICS

STUDY / BATTERY / TEST / SUBTEST / PASSAGE	ORIGINAL TEST MEAN	ORIGINAL TEST STD	ORIGINAL TEST *k*	G-STUDY SAMPLING MEAN	G-STUDY SAMPLING STD	G-STUDY SAMPLING *k*
STUDY ONE						
TOTAL TOEFL	99.6788	27.2798	146	78.9712	21.3691	114
LC TEST	31.1660	10.5508	50	24.0095	8.1238	38
SWE TEST	27.5541	7.6467	38	27.5541	7.6467	38
VRC TEST	40.9588	11.6609	58	27.4077	7.7959	38
STUDY TWO						
LC TEST	31.1660	10.5508	50	27.7185	9.5038	45
LC1	12.6897	4.6491	20	9.2422	3.5805	15
LC2	9.5017	3.3634	15	9.5017	3.3634	15
LC3	8.9746	3.4902	15	8.9746	3.4902	15
STUDY THREE						
SWE TEST	27.5541	7.6467	38	20.4521	5.6912	28
SWE1	10.5301	2.9676	14	10.5301	2.9676	14
SWE2	17.0240	5.1066	24	9.9220	3.1298	14
STUDY FOUR						
VRC TEST	40.9588	11.6609	58	40.9588	11.6609	58
VRC1	20.5785	6.2482	29	20.5785	6.2482	29
VRC2	20.3804	6.0880	29	20.3804	6.0880	29
STUDY FIVE						
VRC2	20.3804	6.0880	29	13.8102	4.3375	20
PASSAGE 1	5.5312	1.5248	7	3.1075	0.9972	4
PASSAGE 2	3.9047	1.2311	5	3.0894	1.0408	4
PASSAGE 3	4.4653	2.0514	7	2.5816	1.2644	4
PASSAGE 4	4.2770	1.7342	6	2.8295	1.2352	4
PASSAGE 5	2.2022	1.3318	4	2.2022	1.3318	4

TABLE 2: CLASSICAL THEORY RELIABILITY STATISTICS

STUDY / BATTERY / TEST / SUBTEST / PASSAGE	ORIGINAL TEST S-H	ORIGINAL TEST Guttman	ORIGINAL TEST Alpha	ORIGINAL TEST *k*	G-STUDY SAMPLING S-H	G-STUDY SAMPLING Guttman	G-STUDY SAMPLING Alpha	G-STUDY SAMPLING *k*
STUDY ONE								
TOTAL TOEFL	.8927	.8916	.9667	146	.8896	.8881	.9584	114
LC TEST	.8978	.8978	.9178	50	.8789	.8788	.8964	38
SWE TEST	.8752	.8652	.9016	38	.8752	.8652	.9016	38
VRC TEST	.8808	.8806	.9326	58	.8617	.8616	.9033	38
STUDY TWO								
LC TEST	.8978	.8978	.9178	50	.8941	.8930	.9077	45
LC1	.8209	.8209	.8349	20	.7706	.7669	.7845	15
LC2	.7541	.7512	.7618	15	.7541	.7512	.7618	15
LC3	.7457	.7437	.7677	15	.7457	.7437	.7677	15
STUDY THREE								
SWE TEST	.8752	.8652	.9016	38	.8520	.8513	.8686	28
SWE1	.7565	.7478	.7726	14	.7565	.7478	.7726	14
SWE2	.8263	.8113	.8574	24	.7466	.7370	.7723	14
STUDY FOUR								
VRC TEST	.8808	.8806	.9326	58	.8808	.8806	.9326	58
VRC1	.8749	.8745	.8854	29	.8749	.8745	.8854	29
VRC2	.8106	.8028	.8769	29	.8106	.8028	.8769	29
STUDY FIVE								
VRC2	.8106	.8028	.8769	29	.7654	.7552	.8280	20
PASSAGE 1	.6232	.6101	.6124	7	.4735	.4684	.4542	4
PASSAGE 2	.5837	.5696	.5715	5	.5219	.5199	.5061	4
PASSAGE 3	.7205	.7100	.7323	7	.5903	.5903	.5910	4
PASSAGE 4	.7191	.7187	.7181	6	.6149	.6148	.6190	4
PASSAGE 5	.5823	.5822	.5964	4	.5823	.5822	.5964	4

TABLE 3: G-STUDY ONE - TOTAL TOEFL BATTERY

SOURCE	*SS*	*df*	*MS*		*EMS*	VARIANCE COMPONENTS
p	80306.73	19999	4.01553728	=	$\sigma^2(pi:s) + n_i\sigma^2(ps) + n_i n_s\sigma^2(p)$	.03140424
s	4200.90	2	2100.45000000	=	$\sigma^2(pi:s) + n_i\sigma^2(ps) + n_p\sigma^2(i:s) + n_p n_i\sigma^2(s)$	.00247421
i:s	24395.20	111	219.77657658	=	$\sigma^2(pi:s) + n_p\sigma^2(i:s)$	.01098074
ps	17417.30	39998	0.43545427	=	$\sigma^2(pi:s) + n_i\sigma^2(ps)$	.00720131
pi:s	359188.37	2219889	0.16180465	=	$\sigma^2(pi:s)$	.16180465

TABLE 4: VARIANCE COMPONENTS FOR FIVE G-STUDIES

	VARIANCE COMPONENTS FOR				
SOURCE	STUDY ONE: TOTAL TOEFL	STUDY TWO: LC TEST	STUDY THREE: SWE TEST	STUDY FOUR: VRC TEST	STUDY FIVE: VRC2 SECTION
RAW COMPONENTS					
p	.03140424	.04035400	.03517126	.03614178	.03699441
s	.00247421	.00000000*	.00028243	.00060287	.00710587
i:s	.01098074	.01178236	.00924644	.01189282	.00762986
ps	.00720131	.00136198	.00148391	.00327638	.01234403
pi:s	.16180465	.18380686	.15119135	.15613306	.15143681
Total	.21386515	.23750520	.19737539	.20804691	.21551098

*This value was a negative variance component, which was rounded to zero after Brennan (1983: 47-48)

TABLE 5: SUMMARY OF D-STUDY RESULTS

MODEL Statistic	D-STUDY RESULTS FOR STUDY ONE: TOTAL TOEFL	STUDY TWO: LC TEST	STUDY THREE: SWE TEST	STUDY FOUR: VRC TEST	STUDY FIVE: VRC2 SECTION
n_s	3	3	2	2	5
n_i	38	15	14	29	4
$n_i n_s$	114	45	28	58	20
$\sigma^2(p)$	.0314	.0406	.0352	.0361	.0370
$\sigma^2(s)$	.0008	.0000	.0001	.0003	.0014
$\sigma^2(i:s)$	.0001	.0003	.0003	.0002	.0004
$\sigma^2(ps)$	.0024	.0005	.0007	.0016	.0025
$\sigma^2(pi:s)$	.0014	.0041	.0054	.0027	.0076
$\bar{X}_p$	.6927	.6160	.7304	.7062	.6905
RANDOM EFFECTS MODEL					
$\sigma^2(\tau)$	.0314	.0406	.0352	.0361	.0370
$\sigma^2(\delta)$	.0038	.0045	.0061	.0043	.0100
$\sigma^2(\Delta)$	.0047	.0048	.0066	.0048	.0118
$E\sigma^2(X)$	.0352	.0451	.0413	.0405	.0470
$\sigma^2(\bar{X})$	.0009	.0003	.0005	.0005	.0018
$E\rho^2(\delta)$	.8916	.8993	.8513	.8930	.7865
S/N	8.2251	8.9305	5.7249	8.3458	3.6838
MIXED EFFECTS MODEL					
$\sigma^2(\tau)$	.0338	.0410	.0359	.0378	.0395
$\sigma^2(\delta)$	.0014	.0041	.0054	.0027	.0076
$\sigma^2(\Delta)$	.0015	.0043	.0057	.0029	.0080
$E\sigma^2(X)$	.0352	.0451	.0413	.0405	.0470
$\sigma^2(\bar{X})$	.0001	.0003	.0003	.0002	.0004
$E\rho^2(\delta)$	.9597	.9094	.8693	.9335	.8390
S/N	23.8139	10.0375	6.6511	14.0376	5.2112

6: GENERALIZABILITY COEFFICIENTS FOR THE TOTAL TOEFL

ITEMS	SUBTEST TYPES																											
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	30	40	50	60	70	80	90	100
1	.157	.271	.358	.426	.482	.527	.565	.598	.626	.650	.671	.690	.707	.722	.736	.748	.760	.770	.779	.788	.848	.881	.903	.918	.929	.937	.944	.949
2	.263	.416	.517	.588	.641	.681	.714	.740	.762	.781	.797	.811	.823	.833	.842	.851	.858	.865	.871	.877	.914	.934	.947	.955	.961	.966	.970	.973
3	.339	.507	.606	.673	.720	.755	.782	.804	.822	.837	.850	.860	.870	.878	.885	.892	.897	.902	.907	.911	.939	.954	.963	.969	.973	.976	.979	.981
4	.397	.569	.664	.725	.767	.798	.822	.841	.856	.868	.879	.888	.895	.902	.908	.913	.918	.922	.926	.929	.952	.963	.971	.975	.979	.981	.983	.985
5	.443	.614	.704	.760	.799	.826	.847	.864	.877	.888	.897	.905	.912	.917	.923	.927	.931	.935	.938	.941	.960	.969	.975	.979	.982	.984	.986	.988
6	.479	.648	.734	.786	.821	.846	.865	.880	.892	.902	.910	.917	.923	.928	.932	.936	.940	.943	.946	.948	.965	.974	.979	.982	.985	.987	.988	.989
7	.509	.674	.757	.806	.838	.861	.879	.892	.903	.912	.919	.926	.931	.935	.940	.943	.946	.949	.952	.954	.969	.976	.981	.984	.986	.988	.989	.990
8	.534	.696	.775	.821	.851	.873	.889	.902	.912	.920	.926	.932	.937	.941	.945	.948	.951	.954	.956	.958	.972	.979	.983	.986	.988	.989	.990	.991
9	.555	.714	.789	.833	.862	.882	.897	.909	.918	.926	.932	.937	.942	.946	.949	.952	.955	.957	.960	.961	.974	.980	.984	.987	.989	.990	.991	.992
10	.573	.729	.801	.843	.870	.890	.904	.915	.924	.931	.937	.942	.946	.950	.953	.956	.958	.960	.962	.964	.976	.982	.985	.988	.989	.991	.992	.993
11	.589	.741	.811	.851	.878	.896	.909	.920	.928	.935	.940	.945	.949	.953	.956	.958	.961	.963	.965	.966	.977	.983	.986	.989	.990	.991	.992	.993
12	.603	.752	.820	.859	.884	.901	.914	.924	.932	.938	.944	.948	.952	.955	.958	.960	.963	.965	.966	.968	.979	.984	.987	.989	.991	.992	.993	.993
13	.615	.762	.827	.865	.889	.906	.918	.927	.935	.941	.946	.950	.954	.957	.960	.962	.965	.966	.968	.970	.980	.985	.988	.990	.991	.992	.993	.994
14	.626	.770	.834	.870	.893	.909	.921	.931	.938	.944	.948	.953	.956	.959	.962	.964	.966	.968	.970	.971	.980	.985	.988	.990	.992	.993	.993	.994
15	.636	.777	.840	.875	.897	.913	.924	.933	.940	.946	.951	.954	.958	.961	.963	.965	.967	.969	.971	.972	.981	.986	.989	.991	.992	.993	.994	.994
16	.645	.784	.845	.879	.901	.916	.927	.936	.942	.948	.952	.956	.959	.962	.965	.967	.969	.970	.972	.973	.982	.986	.989	.991	.992	.993	.994	.995
17	.653	.790	.849	.883	.904	.919	.929	.938	.944	.949	.954	.958	.961	.963	.966	.968	.970	.971	.973	.974	.983	.987	.989	.991	.992	.993	.994	.995
18	.660	.795	.853	.886	.907	.921	.931	.939	.946	.951	.955	.959	.962	.964	.967	.969	.971	.972	.974	.975	.983	.987	.990	.991	.993	.994	.994	.995
19	.666	.800	.857	.889	.909	.923	.933	.941	.947	.952	.956	.960	.963	.965	.968	.970	.971	.973	.974	.976	.984	.988	.990	.992	.993	.994	.994	.995
20	.673	.804	.860	.891	.911	.925	.935	.943	.949	.954	.958	.961	.964	.966	.969	.970	.972	.974	.975	.976	.984	.988	.990	.992	.993	.994	.995	.995
21	.678	.808	.863	.894	.913	.927	.936	.944	.950	.955	.959	.962	.965	.967	.969	.971	.973	.974	.976	.977	.984	.988	.991	.992	.993	.994	.995	.995
22	.683	.812	.866	.896	.915	.928	.938	.945	.951	.956	.960	.963	.966	.968	.970	.972	.973	.975	.976	.977	.985	.989	.991	.992	.993	.994	.995	.995
23	.688	.815	.869	.898	.917	.930	.939	.946	.952	.957	.960	.964	.966	.969	.971	.972	.974	.975	.977	.978	.985	.989	.991	.993	.994	.994	.995	.995
24	.693	.818	.871	.900	.918	.931	.940	.947	.953	.957	.961	.964	.967	.969	.971	.973	.975	.976	.977	.978	.985	.989	.991	.993	.994	.994	.995	.996
25	.697	.821	.873	.902	.920	.932	.941	.948	.954	.958	.962	.965	.968	.970	.972	.974	.975	.976	.978	.979	.986	.989	.991	.993	.994	.995	.995	.996
26	.701	.824	.875	.903	.921	.933	.942	.949	.955	.959	.963	.966	.968	.970	.972	.974	.975	.977	.978	.979	.986	.989	.992	.993	.994	.995	.995	.996
27	.704	.826	.877	.905	.922	.935	.943	.950	.955	.960	.963	.966	.969	.971	.973	.974	.976	.977	.978	.979	.986	.990	.992	.993	.994	.995	.995	.996
28	.708	.829	.879	.906	.924	.936	.944	.951	.956	.960	.964	.967	.969	.971	.973	.975	.976	.978	.979	.980	.986	.990	.992	.993	.994	.995	.995	.996
29	.711	.831	.881	.908	.925	.936	.945	.952	.957	.961	.964	.967	.970	.972	.974	.975	.977	.978	.979	.980	.987	.990	.992	.993	.994	.995	.995	.996
30	.714	.833	.882	.909	.926	.937	.946	.952	.957	.961	.965	.968	.970	.972	.974	.976	.977	.978	.979	.980	.987	.990	.992	.993	.994	.995	.996	.996
38	.733	.846	.892	.916	.932	.943	.950	.956	.961	.965	.968	.970	.973	.975	.976	.978	.979	.980	.981	.982	.988	.991	.993	.994	.995	.995	.996	.996
40	.736	.848	.893	.918	.933	.944	.951	.957	.962	.965	.968	.971	.973	.975	.977	.978	.979	.980	.982	.982	.988	.991	.993	.994	.995	.996	.996	.996
50	.751	.858	.900	.923	.938	.948	.955	.960	.964	.968	.971	.973	.975	.977	.978	.980	.981	.982	.983	.984	.989	.992	.993	.994	.995	.996	.996	.997
57	.758	.862	.904	.926	.940	.949	.956	.962	.966	.969	.972	.974	.976	.978	.979	.980	.982	.983	.983	.984	.989	.992	.994	.995	.995	.996	.996	.997
60	.760	.864	.905	.927	.941	.950	.957	.962	.966	.969	.972	.974	.976	.978	.979	.981	.982	.983	.984	.984	.990	.992	.994	.995	.996	.996	.997	.997
70	.768	.868	.908	.930	.943	.952	.959	.964	.967	.971	.973	.975	.977	.979	.980	.981	.982	.983	.984	.985	.990	.992	.994	.995	.996	.996	.997	.997
80	.773	.872	.911	.932	.945	.953	.960	.965	.968	.971	.974	.976	.978	.979	.981	.982	.983	.984	.985	.986	.990	.993	.994	.995	.996	.996	.997	.997
90	.777	.875	.913	.933	.946	.954	.961	.965	.969	.972	.975	.977	.978	.980	.981	.982	.983	.984	.985	.986	.991	.993	.994	.995	.996	.996	.997	.997
100	.781	.877	.914	.934	.947	.955	.961	.966	.970	.973	.975	.977	.979	.980	.982	.983	.984	.985	.985	.986	.991	.993	.994	.995	.996	.997	.997	.997
114	.785	.879	.916	.936	.948	.956	.962	.967	.970	.973	.976	.978	.979	.981	.982	.983	.984	.985	.986	.986	.991	.993	.995	.995	.996	.997	.997	.997

TABLE 7: GENERALIZABILITY COEFFICIENTS FOR THE LSC TEST

ITEMS	SUBTEST TYPES																										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	30	40	50	60	70	80	90	100
1	.180	.305	.397	.467	.523	.568	.605	.637	.663	.687	.707	.724	.740	.754	.767	.778	.788	.798	.806	.814	.868	.898	.916	.929	.939	.946	.952	.956
2	.303	.465	.566	.635	.685	.723	.753	.777	.796	.813	.827	.839	.850	.859	.867	.874	.881	.887	.892	.897	.929	.946	.956	.963	.968	.972	.975	.978
3	.393	.564	.660	.721	.764	.795	.819	.838	.854	.866	.877	.886	.894	.901	.907	.912	.917	.921	.925	.928	.951	.963	.970	.975	.978	.981	.983	.985
4	.462	.632	.720	.774	.811	.837	.857	.873	.885	.896	.904	.911	.918	.923	.928	.932	.936	.939	.942	.945	.963	.972	.977	.981	.984	.986	.987	.988
5	.515	.680	.761	.810	.842	.865	.882	.895	.905	.914	.921	.927	.933	.937	.941	.945	.948	.950	.953	.955	.970	.977	.982	.985	.987	.988	.990	.991
6	.559	.717	.792	.835	.864	.884	.899	.910	.919	.927	.933	.938	.943	.947	.950	.953	.956	.958	.960	.962	.974	.981	.984	.987	.989	.990	.991	.992
7	.595	.746	.815	.855	.880	.898	.911	.922	.930	.936	.942	.946	.950	.954	.957	.959	.961	.964	.965	.967	.978	.983	.987	.989	.990	.992	.992	.993
8	.625	.769	.833	.870	.893	.909	.921	.930	.937	.943	.948	.952	.956	.959	.962	.964	.966	.968	.969	.971	.980	.985	.988	.990	.991	.993	.993	.994
9	.651	.788	.848	.882	.903	.918	.929	.937	.944	.949	.953	.957	.960	.963	.965	.968	.969	.971	.973	.974	.982	.987	.989	.991	.992	.993	.994	.995
10	.673	.804	.860	.891	.911	.925	.935	.943	.949	.954	.958	.961	.964	.966	.969	.970	.972	.974	.975	.976	.984	.988	.990	.992	.993	.994	.995	.995
11	.692	.818	.871	.900	.918	.931	.940	.947	.953	.957	.961	.964	.967	.969	.971	.973	.974	.976	.977	.978	.985	.989	.991	.993	.994	.994	.995	.996
12	.709	.829	.879	.907	.924	.936	.945	.951	.956	.960	.964	.967	.969	.971	.973	.975	.976	.978	.979	.980	.986	.990	.992	.993	.994	.995	.995	.996
13	.723	.840	.887	.913	.929	.940	.948	.954	.959	.963	.966	.969	.971	.973	.975	.977	.978	.979	.980	.981	.987	.991	.992	.994	.995	.995	.996	.996
14	.737	.848	.894	.918	.933	.944	.951	.957	.962	.966	.969	.971	.973	.975	.977	.978	.979	.981	.982	.982	.988	.991	.993	.994	.995	.996	.996	.996
15	.749	.856	.899	.923	.937	.947	.954	.960	.964	.968	.970	.973	.975	.977	.978	.979	.981	.982	.983	.983	.989	.992	.993	.994	.995	.996	.996	.997
16	.759	.863	.904	.927	.940	.950	.957	.962	.966	.969	.972	.974	.976	.978	.979	.981	.982	.983	.984	.984	.990	.992	.994	.995	.995	.996	.996	.997
17	.769	.869	.909	.930	.943	.952	.959	.964	.968	.971	.973	.976	.977	.979	.980	.982	.983	.984	.984	.985	.990	.993	.994	.995	.996	.996	.997	.997
18	.778	.875	.913	.933	.946	.955	.961	.966	.969	.972	.975	.977	.979	.980	.981	.982	.983	.984	.985	.986	.991	.993	.994	.995	.996	.996	.997	.997
19	.786	.880	.917	.936	.948	.957	.963	.967	.971	.974	.976	.978	.979	.981	.982	.983	.984	.985	.986	.987	.991	.993	.995	.995	.996	.997	.997	.997
20	.794	.885	.920	.939	.951	.958	.964	.968	.972	.975	.977	.979	.980	.982	.983	.984	.985	.986	.986	.987	.991	.994	.995	.996	.996	.997	.997	.997
21	.800	.889	.923	.941	.952	.960	.966	.970	.973	.976	.978	.980	.981	.982	.984	.985	.986	.986	.987	.988	.992	.994	.995	.996	.996	.997	.997	.998
22	.807	.893	.926	.943	.954	.962	.967	.971	.974	.977	.979	.980	.982	.983	.984	.985	.986	.987	.988	.988	.992	.994	.995	.996	.997	.997	.997	.998
23	.813	.897	.929	.945	.956	.963	.968	.972	.975	.977	.979	.981	.983	.984	.985	.986	.987	.987	.988	.989	.992	.994	.995	.996	.997	.997	.997	.998
24	.818	.900	.931	.947	.957	.964	.969	.973	.976	.978	.980	.982	.983	.984	.985	.986	.987	.988	.988	.989	.993	.994	.996	.996	.997	.997	.998	.998
25	.823	.903	.933	.949	.959	.965	.970	.974	.977	.979	.981	.982	.984	.985	.986	.987	.988	.988	.989	.989	.993	.995	.996	.996	.997	.997	.998	.998
26	.828	.906	.935	.951	.960	.967	.971	.975	.977	.980	.981	.983	.984	.985	.986	.987	.988	.989	.989	.990	.993	.995	.996	.997	.997	.997	.998	.998
27	.832	.908	.937	.952	.961	.968	.972	.975	.978	.980	.982	.983	.985	.986	.987	.988	.988	.989	.990	.990	.993	.995	.996	.997	.997	.997	.998	.998
28	.837	.911	.939	.953	.962	.968	.973	.976	.979	.981	.983	.984	.985	.986	.987	.988	.989	.989	.990	.990	.994	.995	.996	.997	.997	.998	.998	.998
29	.840	.913	.940	.955	.963	.969	.974	.977	.979	.981	.983	.984	.986	.987	.987	.988	.989	.990	.990	.991	.994	.995	.996	.997	.997	.998	.998	.998
30	.844	.915	.942	.956	.964	.970	.974	.977	.980	.982	.983	.985	.986	.987	.988	.989	.989	.990	.990	.991	.994	.995	.996	.997	.997	.998	.998	.998
40	.872	.932	.953	.965	.971	.976	.979	.982	.984	.986	.987	.988	.989	.990	.990	.991	.991	.992	.992	.993	.995	.996	.997	.998	.998	.998	.998	.999
45	.882	.937	.957	.968	.974	.978	.981	.983	.985	.987	.988	.989	.990	.990	.991	.992	.992	.993	.993	.993	.996	.997	.997	.998	.998	.998	.999	.999
50	.889	.942	.960	.970	.976	.980	.983	.985	.986	.988	.989	.990	.991	.991	.992	.992	.993	.993	.994	.994	.996	.997	.998	.998	.998	.998	.999	.999
60	.902	.948	.965	.973	.979	.982	.985	.987	.988	.989	.990	.991	.992	.992	.993	.993	.994	.994	.994	.995	.996	.997	.998	.998	.998	.999	.999	.999
70	.910	.953	.968	.976	.981	.984	.986	.988	.989	.990	.991	.992	.992	.993	.993	.994	.994	.995	.995	.995	.997	.998	.998	.998	.998	.999	.999	.999
80	.917	.957	.971	.978	.982	.985	.987	.989	.990	.991	.992	.993	.993	.994	.994	.994	.995	.995	.995	.996	.997	.998	.998	.998	.998	.999	.999	.999
90	.923	.960	.973	.979	.983	.986	.988	.990	.991	.992	.992	.993	.994	.994	.994	.995	.995	.995	.996	.996	.997	.998	.998	.999	.999	.999	.999	.999
100	.927	.962	.974	.981	.984	.987	.989	.990	.991	.992	.993	.993	.994	.994	.995	.995	.995	.996	.996	.996	.997	.998	.998	.999	.999	.999	.999	.999

TABLE 8: GENERALIZABILITY COEFFICIENTS FOR THE SWE TEST

ITEMS	SUBTEST TYPES																										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	30	40	50	60	70	80	90	100
1	.187	.315	.409	.480	.535	.580	.617	.648	.675	.697	.717	.734	.750	.763	.776	.787	.797	.806	.814	.822	.874	.902	.920	.933	.942	.949	.954	.958
2	.313	.477	.578	.646	.695	.732	.762	.785	.804	.820	.834	.846	.856	.865	.873	.880	.886	.891	.897	.901	.932	.948	.958	.965	.970	.973	.976	.979
3	.404	.576	.670	.731	.772	.803	.826	.844	.859	.871	.882	.891	.898	.905	.910	.916	.920	.924	.928	.931	.953	.964	.971	.976	.979	.982	.984	.985
4	.472	.642	.729	.782	.817	.843	.862	.877	.890	.900	.908	.915	.921	.926	.931	.935	.938	.942	.944	.947	.964	.973	.978	.982	.984	.986	.988	.989
5	.526	.689	.769	.816	.847	.869	.886	.899	.909	.917	.924	.930	.935	.939	.943	.947	.950	.952	.955	.957	.971	.978	.982	.985	.987	.989	.990	.991
6	.569	.725	.798	.841	.868	.888	.902	.913	.922	.929	.935	.941	.945	.949	.952	.955	.957	.960	.962	.963	.975	.981	.985	.988	.989	.991	.992	.992
7	.604	.753	.821	.859	.884	.901	.914	.924	.932	.938	.944	.948	.952	.955	.958	.961	.963	.965	.967	.968	.979	.984	.987	.989	.991	.992	.993	.993
8	.633	.775	.838	.873	.896	.912	.924	.932	.940	.945	.950	.954	.957	.960	.963	.965	.967	.969	.970	.972	.981	.986	.989	.990	.992	.993	.994	.994
9	.658	.794	.852	.885	.906	.920	.931	.939	.945	.951	.955	.958	.962	.964	.967	.969	.970	.972	.973	.975	.983	.987	.990	.991	.993	.994	.994	.995
10	.679	.809	.864	.894	.914	.927	.937	.944	.950	.955	.959	.962	.965	.967	.969	.971	.973	.974	.976	.977	.985	.988	.991	.992	.993	.994	.995	.995
11	.698	.822	.874	.902	.920	.933	.942	.949	.954	.958	.962	.965	.968	.970	.972	.974	.975	.977	.978	.979	.986	.989	.991	.993	.994	.995	.995	.996
12	.714	.833	.882	.909	.926	.937	.946	.952	.957	.961	.965	.968	.970	.972	.974	.976	.977	.978	.979	.980	.987	.990	.992	.993	.994	.995	.996	.996
13	.728	.843	.889	.915	.931	.941	.949	.955	.960	.964	.967	.970	.972	.974	.976	.977	.979	.980	.981	.982	.988	.991	.993	.994	.995	.995	.996	.996
14	.741	.851	.896	.920	.935	.945	.952	.958	.963	.966	.969	.972	.974	.976	.977	.979	.980	.981	.982	.983	.988	.991	.993	.994	.995	.996	.996	.997
15	.753	.859	.901	.924	.938	.948	.955	.961	.965	.968	.971	.973	.975	.977	.979	.980	.981	.982	.983	.984	.989	.992	.993	.995	.995	.996	.996	.997
16	.763	.865	.906	.928	.941	.951	.957	.963	.967	.970	.973	.975	.977	.978	.980	.981	.982	.983	.984	.985	.990	.992	.994	.995	.996	.996	.997	.997
17	.772	.871	.910	.931	.944	.953	.960	.964	.968	.971	.974	.976	.978	.979	.981	.982	.983	.984	.985	.985	.990	.993	.994	.995	.996	.996	.997	.997
18	.781	.877	.914	.934	.947	.955	.961	.966	.970	.973	.975	.977	.979	.980	.982	.983	.984	.985	.985	.986	.991	.993	.994	.995	.996	.996	.997	.997
19	.788	.882	.918	.937	.949	.957	.963	.968	.971	.974	.976	.978	.980	.981	.982	.983	.984	.985	.986	.987	.991	.993	.995	.996	.996	.997	.997	.997
20	.795	.886	.921	.940	.951	.959	.965	.969	.972	.975	.977	.979	.981	.982	.983	.984	.985	.986	.987	.987	.992	.994	.995	.996	.996	.997	.997	.997
21	.802	.890	.924	.942	.953	.960	.966	.970	.973	.976	.978	.980	.981	.983	.984	.985	.986	.986	.987	.988	.992	.994	.995	.996	.996	.997	.997	.998
22	.808	.894	.927	.944	.955	.962	.967	.971	.974	.977	.979	.981	.982	.983	.984	.985	.986	.987	.988	.988	.992	.994	.995	.996	.997	.997	.997	.998
23	.814	.897	.929	.946	.956	.963	.968	.972	.975	.978	.980	.981	.983	.984	.985	.986	.987	.987	.988	.989	.992	.994	.995	.996	.997	.997	.997	.998
24	.819	.900	.931	.948	.958	.964	.969	.973	.976	.978	.980	.982	.983	.984	.985	.986	.987	.988	.988	.989	.993	.994	.996	.996	.997	.997	.998	.998
25	.824	.903	.933	.949	.959	.966	.970	.974	.977	.979	.981	.982	.984	.985	.986	.987	.988	.988	.989	.989	.993	.995	.996	.996	.997	.997	.998	.998
26	.828	.906	.935	.951	.960	.967	.971	.975	.977	.980	.981	.983	.984	.985	.986	.987	.988	.989	.989	.990	.993	.995	.996	.997	.997	.997	.998	.998
27	.832	.909	.937	.952	.961	.968	.972	.975	.978	.980	.982	.983	.985	.986	.987	.988	.988	.989	.990	.990	.993	.995	.996	.997	.997	.997	.998	.998
28	.836	.911	.939	.953	.962	.968	.973	.976	.979	.981	.983	.984	.985	.986	.987	.988	.989	.989	.990	.990	.994	.995	.996	.997	.997	.998	.998	.998
29	.840	.913	.940	.955	.963	.969	.974	.977	.979	.981	.983	.984	.986	.987	.987	.988	.989	.990	.990	.991	.994	.995	.996	.997	.997	.998	.998	.998
30	.844	.915	.942	.956	.964	.970	.974	.977	.980	.982	.983	.985	.986	.987	.988	.989	.989	.990	.990	.991	.994	.995	.996	.997	.997	.998	.998	.998
40	.870	.930	.952	.964	.971	.976	.979	.982	.984	.985	.987	.988	.989	.989	.990	.991	.991	.992	.992	.993	.995	.996	.997	.998	.998	.998	.998	.999
50	.886	.940	.959	.969	.975	.979	.982	.984	.986	.987	.988	.989	.990	.991	.992	.992	.993	.993	.993	.994	.996	.997	.997	.998	.998	.998	.999	.999
60	.898	.946	.963	.972	.978	.981	.984	.986	.988	.989	.990	.991	.991	.992	.992	.993	.993	.994	.994	.994	.996	.997	.998	.998	.998	.999	.999	.999
70	.906	.951	.967	.975	.980	.983	.985	.987	.989	.990	.991	.991	.992	.993	.993	.994	.994	.994	.995	.995	.997	.997	.998	.998	.999	.999	.999	.999
80	.912	.954	.969	.977	.981	.984	.986	.988	.989	.990	.991	.992	.993	.993	.994	.994	.994	.995	.995	.995	.997	.998	.998	.998	.999	.999	.999	.999
90	.917	.957	.971	.978	.982	.985	.987	.989	.990	.991	.992	.993	.993	.994	.994	.994	.995	.995	.995	.996	.997	.998	.998	.999	.999	.999	.999	.999
100	.922	.959	.972	.979	.983	.986	.988	.989	.991	.992	.992	.993	.993	.994	.994	.995	.995	.995	.996	.996	.997	.998	.998	.999	.999	.999	.999	.999

TABLE 9: GENERALIZABILITY COEFFICIENTS FOR THE RCV TEST

ITEMS	SUBTEST TYPES																											
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	30	40	50	60	70	80	90	100
1	.185	.312	.405	.476	.531	.576	.613	.645	.671	.694	.714	.731	.747	.760	.773	.784	.794	.803	.812	.819	.872	.901	.919	.932	.941	.948	.953	.958
2	.308	.471	.571	.640	.690	.727	.757	.780	.800	.816	.830	.842	.852	.862	.870	.877	.883	.889	.894	.899	.930	.947	.957	.964	.969	.973	.976	.978
3	.395	.566	.662	.723	.766	.797	.821	.839	.855	.867	.878	.887	.895	.901	.907	.913	.917	.922	.925	.929	.951	.963	.970	.975	.979	.981	.983	.985
4	.461	.631	.719	.774	.810	.837	.857	.872	.885	.895	.904	.911	.917	.923	.928	.932	.936	.939	.942	.945	.962	.972	.977	.981	.984	.986	.987	.988
5	.512	.677	.759	.807	.840	.863	.880	.893	.904	.913	.920	.926	.932	.936	.940	.944	.947	.950	.952	.954	.969	.977	.981	.984	.987	.988	.990	.991
6	.552	.712	.787	.831	.860	.881	.896	.908	.917	.925	.931	.937	.941	.945	.949	.952	.954	.957	.959	.961	.974	.980	.984	.987	.989	.990	.991	.992
7	.586	.739	.809	.850	.876	.894	.908	.919	.927	.934	.940	.944	.948	.952	.955	.958	.960	.962	.964	.966	.977	.983	.986	.988	.990	.991	.992	.993
8	.613	.760	.826	.864	.888	.905	.917	.927	.935	.941	.946	.950	.954	.957	.960	.962	.964	.966	.968	.969	.979	.984	.988	.990	.991	.992	.993	.994
9	.637	.778	.840	.875	.898	.913	.925	.933	.940	.946	.951	.955	.958	.961	.963	.966	.968	.969	.971	.972	.981	.986	.989	.991	.992	.993	.994	.994
10	.657	.793	.852	.884	.905	.920	.931	.939	.945	.950	.955	.958	.961	.964	.966	.968	.970	.972	.973	.975	.983	.987	.990	.991	.993	.994	.994	.995
11	.674	.805	.861	.892	.912	.925	.935	.943	.949	.954	.958	.961	.964	.967	.969	.971	.972	.974	.975	.976	.984	.988	.990	.992	.993	.994	.995	.995
12	.689	.816	.869	.899	.917	.930	.940	.947	.952	.957	.961	.964	.966	.969	.971	.973	.974	.976	.977	.978	.985	.989	.991	.993	.994	.994	.995	.996
13	.703	.825	.876	.904	.922	.934	.943	.950	.955	.959	.963	.966	.968	.971	.973	.974	.976	.977	.978	.979	.986	.990	.992	.993	.994	.995	.995	.996
14	.715	.834	.883	.909	.926	.938	.946	.952	.958	.962	.965	.968	.970	.972	.974	.976	.977	.978	.979	.980	.987	.990	.992	.993	.994	.995	.996	.996
15	.725	.841	.888	.914	.930	.941	.949	.955	.960	.964	.967	.969	.972	.974	.975	.977	.978	.979	.980	.981	.988	.991	.992	.994	.995	.995	.996	.996
16	.735	.847	.893	.917	.933	.943	.951	.957	.961	.965	.968	.971	.973	.975	.977	.978	.979	.980	.981	.982	.988	.991	.993	.994	.995	.996	.996	.996
17	.744	.853	.897	.921	.935	.946	.953	.959	.963	.967	.970	.972	.974	.976	.978	.979	.980	.981	.982	.983	.989	.991	.993	.994	.995	.996	.996	.997
18	.752	.858	.901	.924	.938	.948	.955	.960	.965	.968	.971	.973	.975	.977	.978	.980	.981	.982	.983	.984	.989	.992	.993	.995	.995	.996	.996	.997
19	.759	.863	.904	.926	.940	.950	.957	.962	.966	.969	.972	.974	.976	.978	.979	.981	.982	.983	.984	.984	.990	.992	.994	.995	.995	.996	.996	.997
20	.765	.867	.907	.929	.942	.951	.958	.963	.967	.970	.973	.975	.977	.979	.980	.981	.982	.983	.984	.985	.990	.992	.994	.995	.996	.996	.997	.997
21	.771	.871	.910	.931	.944	.953	.959	.964	.968	.971	.974	.976	.978	.979	.981	.982	.983	.984	.985	.985	.990	.993	.994	.995	.996	.996	.997	.997
22	.777	.875	.913	.933	.946	.954	.961	.965	.969	.972	.975	.977	.978	.980	.981	.982	.983	.984	.985	.986	.991	.993	.994	.995	.996	.996	.997	.997
23	.782	.878	.915	.935	.947	.956	.962	.966	.970	.973	.975	.977	.979	.980	.982	.983	.984	.985	.986	.986	.991	.993	.994	.995	.996	.997	.997	.997
24	.787	.881	.917	.937	.949	.957	.963	.967	.971	.974	.976	.978	.980	.981	.982	.983	.984	.985	.986	.987	.991	.993	.995	.996	.996	.997	.997	.997
25	.791	.884	.919	.938	.950	.958	.964	.968	.972	.974	.977	.979	.980	.982	.983	.984	.985	.986	.986	.987	.991	.993	.995	.996	.996	.997	.997	.997
26	.796	.886	.921	.940	.951	.959	.965	.969	.972	.975	.977	.979	.981	.982	.983	.984	.985	.986	.987	.987	.992	.994	.995	.996	.996	.997	.997	.997
27	.800	.889	.923	.941	.952	.960	.965	.970	.973	.976	.978	.980	.981	.982	.984	.985	.985	.986	.987	.988	.992	.994	.995	.996	.996	.997	.997	.997
28	.803	.891	.925	.942	.953	.961	.966	.970	.974	.976	.978	.980	.982	.983	.984	.985	.986	.987	.987	.988	.992	.994	.995	.996	.997	.997	.997	.998
29	.807	.893	.926	.943	.954	.962	.967	.971	.974	.977	.979	.980	.982	.983	.984	.985	.986	.987	.988	.988	.992	.994	.995	.996	.997	.997	.997	.998
30	.810	.895	.927	.945	.955	.962	.968	.972	.975	.977	.979	.981	.982	.984	.985	.986	.986	.987	.988	.988	.992	.994	.995	.996	.997	.997	.997	.998
40	.834	.910	.938	.953	.962	.968	.972	.976	.978	.981	.982	.984	.985	.986	.987	.988	.988	.989	.990	.990	.993	.995	.996	.997	.997	.998	.998	.998
50	.850	.919	.944	.958	.966	.971	.975	.978	.981	.983	.984	.985	.987	.988	.988	.989	.990	.990	.991	.991	.994	.996	.996	.997	.997	.998	.998	.998
58	.858	.924	.948	.960	.968	.973	.977	.980	.982	.984	.985	.986	.987	.988	.989	.990	.990	.991	.991	.992	.995	.996	.997	.997	.998	.998	.998	.998
60	.860	.925	.949	.961	.968	.974	.977	.980	.982	.984	.985	.987	.988	.989	.989	.990	.991	.991	.992	.992	.995	.996	.997	.997	.998	.998	.998	.998
70	.868	.929	.952	.963	.970	.975	.979	.981	.983	.985	.986	.987	.988	.989	.990	.991	.991	.992	.992	.992	.995	.996	.997	.997	.998	.998	.998	.998
80	.874	.933	.954	.965	.972	.976	.980	.982	.984	.986	.987	.988	.989	.990	.990	.991	.992	.992	.992	.993	.995	.996	.997	.998	.998	.998	.998	.999
90	.878	.935	.956	.966	.973	.977	.981	.983	.985	.986	.988	.989	.989	.990	.991	.991	.992	.992	.993	.993	.995	.997	.997	.998	.998	.998	.998	.999
100	.882	.937	.957	.968	.974	.978	.981	.984	.985	.987	.988	.989	.990	.991	.991	.992	.992	.993	.993	.993	.996	.997	.997	.998	.998	.998	.999	.999

TABLE 10: GENERALIZABILITY COEFFICIENTS FOR THE RCOMP SUBTEST

ITEMS	SUBTEST TYPES																										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	30	40	50	60	70	80	90	100
1	.184	.311	.404	.475	.530	.575	.613	.644	.670	.693	.713	.730	.746	.760	.772	.783	.793	.803	.811	.819	.871	.900	.919	.931	.941	.948	.953	.958
2	.296	.457	.558	.627	.677	.716	.746	.771	.791	.808	.822	.834	.845	.855	.863	.870	.877	.883	.889	.894	.926	.944	.955	.962	.967	.971	.974	.977
3	.371	.541	.639	.702	.746	.779	.805	.825	.841	.855	.866	.876	.884	.892	.898	.904	.909	.914	.918	.922	.946	.959	.967	.972	.976	.979	.981	.983
4	.424	.596	.689	.747	.787	.816	.838	.855	.869	.881	.890	.898	.905	.912	.917	.922	.926	.930	.933	.936	.957	.967	.974	.978	.981	.983	.985	.987
5	.465	.634	.722	.776	.813	.839	.859	.874	.886	.897	.905	.912	.919	.924	.929	.933	.937	.940	.943	.946	.963	.972	.977	.981	.984	.986	.987	.989
6	.496	.663	.747	.797	.831	.855	.873	.887	.899	.908	.915	.922	.928	.932	.937	.940	.944	.947	.949	.952	.967	.975	.980	.983	.986	.987	.989	.990
7	.521	.685	.766	.813	.845	.867	.884	.897	.907	.916	.923	.929	.934	.938	.942	.946	.949	.951	.954	.956	.970	.978	.982	.985	.987	.989	.990	.991
8	.542	.703	.780	.826	.855	.877	.892	.904	.914	.922	.929	.934	.939	.943	.947	.950	.953	.955	.957	.959	.973	.979	.983	.986	.988	.990	.991	.992
9	.559	.717	.792	.835	.864	.884	.899	.910	.919	.927	.933	.938	.943	.947	.950	.953	.956	.958	.960	.962	.974	.981	.984	.987	.989	.990	.991	.992
10	.574	.729	.801	.843	.871	.890	.904	.915	.924	.931	.937	.942	.946	.950	.953	.956	.958	.960	.962	.964	.976	.982	.985	.988	.989	.991	.992	.993
11	.586	.739	.810	.850	.876	.895	.908	.919	.927	.934	.940	.944	.949	.952	.955	.958	.960	.962	.964	.966	.977	.983	.986	.988	.990	.991	.992	.993
12	.597	.748	.816	.856	.881	.899	.912	.922	.930	.937	.942	.947	.951	.954	.957	.960	.962	.964	.966	.967	.978	.983	.987	.989	.990	.992	.993	.993
13	.607	.755	.822	.860	.885	.902	.915	.925	.933	.939	.944	.949	.952	.956	.959	.961	.963	.965	.967	.969	.979	.984	.987	.989	.991	.992	.993	.994
14	.615	.762	.827	.865	.889	.906	.918	.927	.935	.941	.946	.950	.954	.957	.960	.962	.964	.966	.968	.970	.980	.985	.988	.990	.991	.992	.993	.994
15	.622	.767	.832	.868	.892	.908	.920	.930	.937	.943	.948	.952	.955	.958	.961	.963	.966	.967	.969	.971	.980	.985	.988	.990	.991	.992	.993	.994
16	.629	.772	.836	.872	.895	.911	.922	.931	.939	.944	.949	.953	.957	.960	.962	.964	.966	.968	.970	.971	.981	.985	.988	.990	.992	.993	.993	.994
17	.635	.777	.839	.874	.897	.913	.924	.933	.940	.946	.950	.954	.958	.961	.963	.965	.967	.969	.971	.972	.981	.986	.989	.991	.992	.993	.994	.994
18	.641	.781	.842	.877	.899	.914	.926	.934	.941	.947	.951	.955	.959	.961	.964	.966	.968	.970	.971	.973	.982	.986	.989	.991	.992	.993	.994	.994
19	.646	.785	.845	.879	.901	.916	.927	.936	.942	.948	.952	.956	.959	.962	.965	.967	.969	.970	.972	.973	.982	.986	.989	.991	.992	.993	.994	.995
20	.650	.788	.848	.881	.903	.918	.929	.937	.944	.949	.953	.957	.960	.963	.965	.967	.969	.971	.972	.974	.982	.987	.989	.991	.992	.993	.994	.995
21	.654	.791	.850	.883	.904	.919	.930	.938	.945	.950	.954	.958	.961	.964	.966	.968	.970	.971	.973	.974	.983	.987	.990	.991	.993	.993	.994	.995
22	.658	.794	.852	.885	.906	.920	.931	.939	.945	.951	.955	.958	.962	.964	.967	.969	.970	.972	.973	.975	.983	.987	.990	.991	.993	.994	.994	.995
23	.662	.796	.854	.887	.907	.921	.932	.940	.946	.951	.956	.959	.962	.965	.967	.969	.971	.972	.974	.975	.983	.987	.990	.992	.993	.994	.994	.995
24	.665	.799	.856	.888	.908	.922	.933	.941	.947	.952	.956	.960	.963	.965	.967	.969	.971	.973	.974	.975	.983	.988	.990	.992	.993	.994	.994	.995
25	.668	.801	.858	.889	.910	.923	.934	.941	.948	.953	.957	.960	.963	.966	.968	.970	.972	.973	.974	.976	.984	.988	.990	.992	.993	.994	.995	.995
26	.671	.803	.859	.891	.911	.924	.934	.942	.948	.953	.957	.961	.964	.966	.968	.970	.972	.973	.975	.976	.984	.988	.990	.992	.993	.994	.995	.995
27	.673	.805	.861	.892	.912	.925	.935	.943	.949	.954	.958	.961	.964	.966	.969	.971	.972	.974	.975	.976	.984	.988	.990	.992	.993	.994	.995	.995
28	.676	.806	.862	.893	.912	.926	.936	.943	.949	.954	.958	.962	.964	.967	.969	.971	.973	.974	.975	.977	.984	.988	.990	.992	.993	.994	.995	.995
29	.678	.808	.863	.894	.913	.927	.936	.944	.950	.955	.959	.962	.965	.967	.969	.971	.973	.974	.976	.977	.984	.988	.991	.992	.993	.994	.995	.995
30	.680	.810	.865	.895	.914	.927	.937	.944	.950	.955	.959	.962	.965	.968	.970	.971	.973	.975	.976	.977	.985	.988	.991	.992	.993	.994	.995	.995
40	.696	.821	.873	.902	.920	.932	.941	.948	.954	.958	.962	.965	.968	.970	.972	.973	.975	.976	.978	.979	.986	.989	.991	.993	.994	.995	.995	.996
50	.706	.828	.878	.906	.923	.935	.944	.951	.956	.960	.964	.967	.969	.971	.973	.975	.976	.977	.979	.980	.986	.990	.992	.993	.994	.995	.995	.996
60	.713	.833	.882	.909	.926	.937	.946	.952	.957	.961	.965	.968	.970	.972	.974	.975	.977	.978	.979	.980	.987	.990	.992	.993	.994	.995	.996	.996
70	.718	.836	.884	.911	.927	.939	.947	.953	.958	.962	.966	.968	.971	.973	.975	.976	.977	.979	.980	.981	.987	.990	.992	.994	.994	.995	.996	.996
80	.722	.839	.886	.912	.929	.940	.948	.954	.959	.963	.966	.969	.971	.973	.975	.977	.978	.979	.980	.981	.987	.990	.992	.994	.995	.995	.996	.996
90	.725	.841	.888	.913	.930	.941	.949	.955	.960	.963	.967	.969	.972	.974	.975	.977	.978	.979	.980	.981	.988	.991	.992	.994	.995	.995	.996	.996
100	.727	.842	.889	.914	.930	.941	.949	.955	.960	.964	.967	.970	.972	.974	.976	.977	.978	.980	.981	.982	.988	.991	.993	.994	.995	.995	.996	.996

TABLE 11: SUMMARY OF PHI(LAMBDA) RESULTS

MODEL Cut Point	PHI(LAMBDA) RESULTS FOR STUDY ONE: TOTAL TOEFL	STUDY TWO: LC TEST	STUDY THREE: SWE TEST	STUDY FOUR: VRC TEST	STUDY FIVE: VRC2 SECTION
RANDOM EFFECTS MODEL					
90%	.939	.962	.906	.938	.870
80%	.899	.939	.857	.902	.799
70%	.866	.908	.843	.881	.749
60%	.892	.894	.887	.907	.786
50%	.934	.918	.930	.942	.858
40%	.961	.948	.956	.964	.910
30%	.975	.967	.971	.976	.941
20%	.983	.978	.980	.984	.959
10%	.988	.985	.985	.988	.970
$\Phi(\bar{X})$	.865	.894	.840	.880	.748
$\bar{X}$ =	69%	62%	73%	71%	69%
MIXED EFFECTS MODEL					
90%	.981	.965	.918	.963	.913
80%	.968	.945	.876	.941	.865
70%	.957	.917	.864	.928	.831
60%	.965	.904	.902	.944	.856
50%	.979	.926	.939	.965	.905
40%	.987	.953	.962	.978	.939
30%	.992	.970	.975	.986	.960
20%	.995	.980	.982	.990	.972
10%	.996	.986	.987	.993	.980
$\Phi(\bar{X})$	.957	.904	.861	.928	.831
$\bar{X}$ =	69%	62%	73%	71%	69%

APPENDIX A: G-STUDY RESULTS [p x (i:s) DESIGNS]

STUDY ONE - TOTAL TOEFL BATTERY

SOURCE	SS	df	MS	EVC
p	80306.73	19999	4.0155	0.03140424
s	4200.90	2	2100.4500	0.00247421
i:s	24395.20	111	219.7766	0.01098074
ps	17417.30	39998	0.4355	0.00720131
pi:s	359188.37	2219889	0.1618	0.16180465

STUDY TWO - LC TEST

SOURCE	SS	df	MS	EVC
p	40581.31	19999	2.0292	0.04055400
s	183.80	2	91.9020	0.00000000
i:s	9904.90	42	235.8310	0.01178236
ps	8169.05	39998	0.2042	0.00136198
pi:s	154390.04	839958	0.1838	0.18380686

STUDY THREE - SWE TEST

SOURCE	SS	df	MS	EVC
p	23134.07	19999	1.1568	0.03517126
s	264.18	1	264.1800	0.00028243
i:s	4812.08	26	185.0800	0.00924644
ps	3439.15	19999	0.1720	0.00148391
pi:s	78615.57	519974	0.1512	0.15119135

STUDY FOUR - VRC TEST

SOURCE	SS	df	MS	EVC
p	46945.08	19999	2.3474	0.03614178
s	587.77	1	587.7720	0.00060287
i:s	13328.70	56	238.0125	0.01189282
ps	5022.71	19999	0.2511	0.00327638
pi:s	174860.28	1119944	0.1561	0.15613306

STUDY FIVE - VRC2 SECTION

SOURCE	SS	df	MS	EVC
p	18813.08	19999	0.9407	0.03699441
s	2885.07	4	721.2675	0.00710587
i:s	2291.23	15	152.7487	0.00762986
ps	16064.23	79996	0.2008	0.01234403
pi:s	45428.77	299985	0.1514	0.15143681