DOCUMENT RESUME

ED 400 327 TM 025 732

AUTHOR Schedl, Mary; And Others
TITLE An Analysis of the Dimensionality of TOEFL Reading Comprehension Items. TOEFL Research Reports, 53.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-95-27
PUB DATE Mar 96
NOTE 36p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adults; *English (Second Language); Factor Analysis; Factor Structure; *Language Tests; *Limited English Speaking; *Reading Comprehension; Reading Tests; Second Language Learning; Test Construction; Test Items; *Thinking Skills
IDENTIFIERS *Dimensionality (Tests); NOHARM Computer Program; *Test of English as a Foreign Language; Test Specifications

ABSTRACT

The issue of what exactly is measured by different types of reading items has been a matter of interest in the field of reading research for many years. Language teaching and testing specialists have raised the question of whether a reading test for foreign students wishing to enter a university in the United States should include questions testing abilities beyond linguistic and very general discourse competencies, or indeed whether it is possible to separate these language competencies from other competencies. This study investigates the dimensionality of the Test of English as a Foreign Language (TOEFL) reading test, based on the specifications in use as of April 1991. Of particular interest was whether four item types identified in the test specification as "reasoning items" could be shown to measure, in addition to general reading ability, any abilities not measured by the other item types in the test. Two techniques, Stout's procedure and NOHARM analyses, were used to investigate the hypothesized two-factor model. In both cases the data failed to fit the model, indicating that TOEFL "reasoning items" cannot be shown to measure a unique construct. However, the followup exploratory analyses indicated that all 10 test forms used in the study violated the assumption of essential unidimensionality, and all of the forms appeared to fit a two-factor model where the second factor may be related to passage content or position. (Contains 4 tables, 7 figures, and 20 references.) (Author/SLD)

TOEFL®

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 53
MARCH 1996

An Analysis of the Dimensionality of TOEFL Reading Comprehension Items

Mary Schedl
Ann Gordon
Patricia A. Carey
K. Linda Tang

ETS®
Educational Testing Service

An Analysis of the Dimensionality of TOEFL Reading Comprehension Items

Mary Schedl
Ann Gordon
Patricia A. Carey
K. Linda Tang

Educational Testing Service
Princeton, New Jersey

RR-95-27

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

Abstract

The issue of exactly what is measured by different types of reading items has been a matter of interest in the field of reading research for many years. Language teaching and testing specialists have raised the question of whether a reading test for foreign students wishing to enter university in the United States should include questions testing abilities beyond linguistic and very general discourse competencies, or indeed whether it is possible to separate these language competencies from other competencies. The purpose of this study was to investigate the dimensionality of the TOEFL® reading test, based on the specifications in use as of April 1991. Of particular interest was whether four item types identified in the test specifications as "reasoning items" could be shown to measure, in addition to general reading ability, any abilities not measured by the other item types in the TOEFL reading test. Two techniques, Stout's procedure and NOHARM analyses, were employed to investigate the hypothesized two-factor model. In both cases the data failed to fit the model, indicating that TOEFL "reasoning items" cannot be shown to measure a unique construct. However, the follow-up exploratory analyses indicated that all 10 test forms used in the study violated the assumption of essential unidimensionality, and all of the forms appeared to fit a two-factor model where the second factor may be related to passage content or position.

The Test of English as a Foreign Language (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS®) and the College Board assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. The Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. Many projects require the cooperation of other institutions, however, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that data confidentiality will be protected.

Current (1995-96) members of the TOEFL Research Committee are:

Paul Angelis	Southern Illinois University at Carbondale
Carol Chapelle	Iowa State University
Fred Davidson	University of Illinois at Urbana-Champaign
Thom Hudson	University of Hawaii
Linda Schinke-Llano	Millikin University
John Upshur (Chair)	Concordia University

Acknowledgments

The authors wish to express their gratitude to the following current and former ETS staff members who contributed to the study:

Steve Laue and Judy DeChamplain for helping with the statistical analysis.

Andre DeChamplain for technical assistance with the analysis procedures.

Susan Chyn, Dan Eignor, and Philip Oltman for helpful reviews of earlier drafts of the report.

Ming Mei Wang for valuable guidance in structuring the analyses and interpreting the results.

Mary Anne Nieciecki and Eugenia Tye for secretarial assistance.

Table of Contents

List of Tables

List of Figures

Introduction

The issue of exactly what is measured by different types of reading items has been a matter of interest in the field of reading research for many years. In terms of the TOEFL® test, the question has been raised as to whether a reading test for foreign students wishing to enter university in the United States should include questions testing abilities beyond linguistic and very general discourse competencies, or indeed whether it is possible to separate these from other competencies involved in reading.

A number of investigations related to the dimensionality of the TOEFL test as a whole have been conducted. Dunbar (1982) used confirmatory factor analysis to examine the factor structure of the TOEFL test for each of seven language groups. In general, the data showed the dominance of one general factor while, at the same time, indicating the importance of various factors associated with different sections of the test. Oltman, Stricker, and Barrows (1988) investigated whether there are systematic differences among item types in patterns of responses, and whether these differences are associated with native language. They concluded that test interpretation varies with individual examinees' English proficiency and that dimensionality of the TOEFL test and of competence in English also depend on examinees' English proficiency. Hale, Stansfield, Rock, Hicks, Butler, and Oller (1988) examined the relationship of multiple-choice cloze items to different parts of the TOEFL test. Results suggested that skills associated with grammar, vocabulary, and reading comprehension are highly interrelated, as assessed by the TOEFL test and by the multiple-choice cloze items. Hale, Rock, and Jirele (1989) investigated two hypotheses regarding discrepancies in the number of factors contributing to TOEFL performance for each of several major language groups and found that the use of different factor-analytic methodologies in earlier studies contributed to inconsistencies. McKinley and Way (1992) explored the feasibility of an item response theory-based method of modeling examinee performance on secondary dimensions present in the test. As part of a feasibility study in preparation for the introduction of a revised TOEFL test in 1995, Schedl, Thomas, and Way (1995) conducted an analysis of the two subparts of Section 3, Vocabulary and Reading Comprehension, and found that departures from essential unidimensionality for the two subsections were primarily due to end-of-test effects that might be related to the timing of the test. While previous research has shown multidimensionality for the test as a whole, none has addressed the constructs implied by the specifications for reading comprehension items.

In contrast to these previous studies, which focused on the TOEFL test as a whole, or on a comparison of the Vocabulary and Reading Comprehension subparts of Section 3, the purpose of the study reported here was to investigate the dimensionality of the Reading Comprehension subpart alone, which at the time of the study consisted of

sets of items based on reading passages[1]. Examinee performance on different item types included in the TOEFL specifications for these passage sets was evaluated based on operational test specifications in use since April 1991. Performance on items classified as "reasoning items" was compared to performance on all other reading items in the reading comprehension sets to determine whether such items could be shown to measure a unique ability in addition to general reading ability.

Relevant Literature

Many lists of reading skills have been assembled over the years, primarily by first-language reading specialists, with the assumption that reading comprehension includes a number of distinct subskills or abilities. While different experts identify somewhat different skills or factors that are involved in the reading process, skills such as the ability to understand the meaning of words and structures in the text and to understand literal, explicitly stated information in a text are generally thought to be "lower level" skills. "Higher level" reading skills, on the other hand, include the ability to extrapolate from a text to situations outside the text, to recognize an author's tone, and to draw analogies and inferences (Barrett, 1968; Davies and Widdowson, 1974; Davis, 1968; Grabe, 1986). Recent studies, however, have failed to find evidence of separate subskills. Lunzer, Waite, and Dolan (1979) rejected the hypothesis that comprehension involves distinct skills or subskills. They found that a single factor accounted for most of the total score variance in their study, which was designed to answer the question of whether or not comprehension is a unitary ability, based on the reading test performance of native speakers between the ages of 10 and 11 years. They also rejected a second hypothesis that two levels of comprehension may exist when they found no group of pupils whose performance on higher level tasks varied significantly from what it would be predicted to be on the basis of their performance on lower level tasks.

Although the Lunzer, Waite, and Dolan study is sometimes cited as evidence that identifiable reading comprehension subskills do not exist, it is far from conclusive evidence of this. Alderson and Lukmani (1989) attempted to replicate the findings with an English as a Second Language (ESL) first-year university population using subskills assumed to be measured in a reading test used at the University of Bombay. They found considerable agreement among judges in categorizing 14 out of 41 reading comprehension questions as measuring lower, middle, or higher order skills, although there was disagreement even for these 14 items as to which specific skills were being tested by each. Of these 14 items, five were categorized as measuring lower order skills, four as measuring middle order skills, and five as measuring higher order skills. The authors found that, contrary to their expectations, scores for items identified as lower level correlated better with the total score for this sample than with scores for items identified as higher level for the 100 students in their sample. They speculated that

[1]With the introduction of the revised TOEFL test in July 1995, the Vocabulary subpart of Section 3 was eliminated and the number of items based on Reading Comprehension passages was increased.

lower level questions might measure *language* ability while higher level questions might involve "cognitive skills, logic, reasoning ability, and so on (p. 268)."[2]

On the other hand, items that measure higher level skills would also require lower level skills (perhaps even have lower level skills as prerequisites) so that the total scores would most likely reflect a greater weight for lower level ability than for higher level ability in any case.

Using an analysis of TOEFL examinee performance on items designed to test different subskills required in reading, this study investigated the possibility that higher level reading comprehension items measure a unique trait in addition to the general reading ability measured by all items in the reading test. TOEFL Reading Comprehension test specifications in use since April 1991 were used as a framework for categorizing reading comprehension test questions as higher level reading comprehension item types, loosely grouped together in the test specifications as "reasoning items."[3] These included four types of items testing: (1) analogy, (2) extrapolation, (3) organization and logic, and (4) author's purpose/attitude. It was recommended that these items be distinguished from other types in the test specifications (primarily consisting of items testing vocabulary, syntax, and explicitly stated information) because the items identified as reasoning items seemed to require more complex reasoning processes. Item types specified for inclusion in the TOEFL reading test, and the TOEFL test population, are similar to, although not identical to, those in the Alderson and Lukmani study.

Objectives of the Study

The study had two objectives. The major objective was to provide information related to the dimensionality of the TOEFL reading comprehension section to the TOEFL program for use in choosing and weighting test specifications for the TOEFL reading test. If examinee performance were shown to differ on reasoning items from examinee performance on other items, indicating that the reasoning items were measuring an additional trait, it would be necessary to consider the extent to which it is desirable to measure this trait. Content specifications for reasoning items would need to be weighted appropriately as a group of items contributing a separate dimension to the measurement rather than as individual item types contributing variety to the overall assessment.

---

[2]The term "language ability" was not defined by the authors but by its juxtaposition with the specific abilities listed here as being tested by higher level questions. It would seem to refer to basic linguistic competencies related to syntax and lexicon.

[3]Although the terms "lower level" and "higher level" are common in the literature, the term "reasoning items" was used for the TOEFL test specifications because "higher level" has a connotation of "more difficult," and the use of the term "reasoning" does not imply a specific level of difficulty.

Another possible outcome of a finding that reasoning items in the reading section measure a somewhat different underlying construct than the "nonreasoning" reading comprehension items would be that the use of standard IRT procedures in TOEFL item calibration and equating might need to be reconsidered, since the IRT procedures used at present are based on the assumption that all items in the reading section measure a unidimensional construct.

A secondary objective was to provide information to the TOEFL program for consideration in design decisions related to the development of a new TOEFL test. The TOEFL program would be able to use information regarding dimensionality in language assessment as a first step in considering dimensionality issues in trial designs. It was thought to be possible that some of the academic "tasks" identified as part of a possible academic domain for a new test would involve more explicit reasoning than that involved in current TOEFL tasks, or might involve multidimensional aspects of language competence. The results of the current study could be informative to the TOEFL program in this regard.

Methods

The research question investigated was whether the reasoning items measure a unique ability in addition to that being assessed by the remaining items associated with passages in the reading test. Two techniques were employed to investigate this research question: Stout's procedure for assessing essential unidimensionality (Stout, 1987; Nandakumar, 1991) and McDonald's nonlinear factor analysis procedure (McDonald, 1967, 1982, 1983) as implemented in the computer program NOHARM II (Fraser, 1983; Fraser & McDonald, 1988). Stout's procedure is based on a nonparametric item response theory model and takes a hypothesis-testing approach to assessing dimensionality. NOHARM II, on the other hand, is based on the three-parameter normal ogive model and a factor-analytical approach. Both procedures were used, because it was thought that the results of the study would be more convincing if the two procedures led to the same conclusions.

Stout's procedure tests the following hypothesis: H_o: $d_E = 1$ versus H_1: $d_E > 1$, where d_E denotes the essential dimensionality underlying the responses to a set of items. Stout's procedure assumes the average covariance of item responses over item pairs for a given θ (ability) is small in magnitude for all values of θ if H_o is true. In order to compute Stout's test statistic, a given test form is divided into three subtests: a partitioning subtest (PT), an assessment subtest 1 (AT1), and an assessment subtest 2 (AT2). The partitioning subtest assigns examinees to different subgroups according to their subtest scores, such that within a subgroup the examinees' abilities will be approximately equal. AT1 includes those items in a test form that are assumed to measure a different ability dimension. For this subtest, the average covariance of examinees' responses to item pairs within a given ability subgroup defined by the partitioning subtest is expected to be large if H_o is false. AT2 is used to adjust examinee variability bias and item difficulty bias. Examinee variability bias occurs when the

covariance of an examinee's responses is small because AT1 is too short. Item difficulty bias occurs when the AT1 items are overly homogeneous with respect to item difficulty.

In the present study, the reasoning items were defined a priori as AT1. AT2 included those reading items that have an item difficulty distribution similar to that of AT1. The rest of the reading items were used in the partitioning subtest. The procedure then assigned examinees to different subgroups according to their partitioning subtest scores. The variance estimates of AT1 and AT2 for each subgroup were computed, and the value of test statistic T was computed based on these variance estimates. The hypothesis H_o: $d_E = 1$ was to be rejected if the value of T was greater than Z_α, where Z_α is the upper $100(1 - \alpha)$ percentile of the standard normal distribution. To follow general practice, $\alpha = 0.05$ was used in the study. In subsequent exploratory analyses, the items in AT1 were not pre-identified, as they were in the confirmatory analyses, but were identified by the program through a principle factor analysis of the interitem tetrachoric correlations.

Confirmatory multidimensional calibrations were done using the computer program NOHARM II, which uses harmonic analysis to approximate a multidimensional normal ogive model. The normal ogive model can be stated as:

$$P_i(\underline{\theta}_j) = c_i + (1 - c_i)\, N[f_{0i} + \underline{f}_{1i}'\underline{\theta}_j].$$

where $P_i(\underline{\theta}_j)$ is the probability of a correct response to item i by examinee j, N represents the normal distribution function, f_{0i} is a threshold value for item i, $\underline{f}_{1i}$ is a $k \times 1$ discrimination vector for item i, c_i is the lower asymptote parameter for item i, $\underline{\theta}_j$ is a $k \times 1$ vector of abilities for examinee j, and there are k dimensions. The item discrimination vector can be reparameterized to factor loadings in the factor analysis model. The confirmatory solution was obtained with all of the items hypothesized to measure an overall proficiency dimension and reasoning items hypothesized to measure an additional proficiency dimension, that is, $k = 2$ was used. The relative fit of this solution, along with that of alternative one-factor and two-factor exploratory solutions, was evaluated by the magnitude of the root mean square of the residual item covariances. In the two-factor exploratory solution, items that might be associated with a second factor were not specified in advance but were identified subsequently by rotation of the factor loadings.

Several studies have been conducted to evaluate the Type I error rate and power of Stout's T statistics. Nandakumar (1991) reported that Stout's T statistic had empirical Type I error rates at or close to the nominal Type I error rate (the error that occurs when a null hypothesis is rejected even when it is true; $\alpha = 0.05$ in this study) when the data set contained 30 to 40 items, five of which tested a second "minor" ability dimension, and sample sizes ranged from about 1,000 to 1,800 examinees. De Champlain (1992) found that the power of Stout's T statistic with regard to rejecting the null hypothesis of unidimensionality was close to 1 when the number of examinees was greater than 500, the test length was 15 to 45 items, and the ratio of the number of items belonging to each of the two dimensions was 4:1. The results of these studies provided guidelines for the design of the current study.

Three recent TOEFL operational forms with 10 of their embedded pretests were used for the study. That is, there were three unique operational forms replicated in two to four pretest forms, with the operational items common across a group of unique pretests. Each total form (i.e., one of the three operational forms plus one of the 10 pretests) originally had 44 reading items, among which at least six were reasoning items. The number of examinees taking these items was greater than 1,000. Each total form consisted of seven reading passage sets, five of which were common across a set of forms, and two were unique to each form. To avoid the confounding of possible effects that may be related to the timing of the test with dimensionality effects (Oltman, Stricker, & Barrows, 1988; Schedl, Thomas, & Way, 1995), the items in the final, or seventh, passage in each of the TOEFL forms were eliminated from the analyses, leaving the total number of items per modified total form to be analyzed to be 38 or 39. The study design provided 10 replications (i.e., 10 total form combinations) and reasonable power to detect whether the reasoning items measure an ability different from that being assessed by the remaining items in the reading test.

Results

Tests of Essential Unidimensionality

Table 1 presents, for each of the 10 test forms, the results of the tests of essential unidimensionality as obtained from Stout's procedure. These analyses included both the operational and pretest items from the reading comprehension subpart of Section 3, excluding the items from the last passage of the test. As mentioned earlier, the operational items from the last passage of the test were excluded from the analyses in an attempt to avoid contamination from any possible end-of-test effects that may be related to speededness. Table 1 contains the form designation, number of examinees, number of items, and Stout's T statistics and associated probabilities from the confirmatory and exploratory analyses for each of the forms. For the confirmatory analyses, the reasoning items were assigned to assessment subtest 1 (AT1); and for the exploratory analyses, the items in AT1 were identified by the program.

As can be seen from Table 1, all of the T statistics from the confirmatory analyses were nonsignificant (i.e., all of the probabilities were greater than 0.05), indicating that the reasoning items do not comprise a second dimension. However, the significant T statistics from the exploratory analyses indicate that the test of essential unidimensionality was rejected for each of the forms; that is, one or more secondary proficiencies, in addition to overall reading proficiency, are being measured by some of the items in the test. Inspection of the output from the exploratory Stout's analyses revealed that for all of the forms, AT1 was generally comprised of items from passage sets 5 and 6. While these statistical results failed to support the hypothesis that the reasoning items measure a different underlying construct than is measured by the other items in the reading test, the results do suggest that another subset of items may contribute to a minor secondary dimension.

Multidimensional Calibrations

Table 2 presents the results of the multidimensional NOHARM analyses for each of the test forms. The table contains the root mean square residuals (RMSR) from two sets of confirmatory analyses, one with a specific reasoning factor and another with a specific passage content or passage position factor. In the latter analyses, the items from passage sets 5 and 6 were specified as belonging to a second factor. Also presented are the RMSRs from exploratory one-factor and two-factor analyses, as well as the factor correlations from the exploratory two-factor analyses. In the two-factor exploratory solution, no constraints on the factor pattern were imposed, so items that might be associated with a second factor were not pre-identified.

The RMSRs from the NOHARM analyses with a specific reasoning factor were all comparable to the corresponding RMSRs for the one-factor analyses, suggesting the two-factor reasoning model provides no practical improvement in fit over a one-factor model. In contrast, both the two-factor exploratory analyses and the two-factor confirmatory analyses with a specific passage content or passage position factor produced relatively lower RMSRs than both the specific reasoning factor model and the one-factor model. These results suggest that the data do not appear to support the hypothesis that the reasoning items measure an essentially different underlying construct than is measured by the other items in the TOEFL reading comprehension subpart. While a second, passage content or passage position-related proficiency, dimension may be present, the factor correlations in the last column of Table 2 indicate that the unique contribution of this dimension to examinees' scores is relatively minor.

Tables 3a, 3b, and 3c list examples of the oblique rotated factor loadings from the exploratory two-factor NOHARM II analyses for the items in Forms 1C, 2A, and 3C, respectively. The tables list the items within each passage in the order in which they appeared in the test. Each set is also labeled sequentially from 1 to 6 in its tested order. As can be seen in Tables 3a - 3c, each passage set consists of from four to eight items. Inspection of the factor loadings in columns two and three for factors 1 and 2, respectively, reveal that for all three forms, the items in sets 5 and 6 appear to load more heavily on the second factor. That is, the magnitude of the factor loadings in the last column (factor 2) for most of the items in sets 5 and 6 is greater than the magnitude of the factor loadings in the second column (factor 1) for these items. These results suggest a possible passage content or passage position effect for these test forms; that is, there may be something about the content in passages 5 and 6 or their position in the test to cause what appears to be a minor secondary factor.

These results are also illustrated in Figures 1-6 for Forms 1C, 2A, and 3C, respectively. Figures 1-3 present bivariate plots of the factor loadings from exploratory two-factor NOHARM analyses. The data points represented by a square correspond to the reasoning items; and those represented by a plus correspond to the remaining items in the test. For each of the forms, it can be seen that the reasoning items appear to be scattered randomly among the rest of the items in the test.

Figures 4-6 again present bivariate plots of the factor loadings from the exploratory two-factor analyses for Forms 1C, 2A, and 3C. For these figures, the data points represented by a square correspond to items in set 5 and those represented by a diamond correspond to items in set 6. In Figures 4 and 5, for Forms 1C and 2A, respectively, the items in sets 5 and 6 load more heavily on factor 2 and clearly form two distinct clusters in the lower right portion of the plot, apart from the remaining items in the test. As in Figures 4 and 5, most of the items in factor 2 for Form 3C fall in the lower right portion of the plot in Figure 6. These are the items in sets 5 and 6 in Form 3C. For Forms 1C and 2A, the items in set 6 load more heavily on factor 2 than do those in set 5, as illustrated in Figures 4 and 5. In Form 3C, however, the items in set 5 load more heavily on factor 2 than do those in set 6, as illustrated in Figure 6. Based on these observations, the second factor is not likely to be entirely related to position.

Additional exploratory NOHARM II analyses were also conducted for several of the test forms to explore the possibility of more than two dimensions contributing to the measurement of examinees' proficiency. Again, the results of these analyses suggest that the two-factor model is adequate for describing the empirical data. Figure 7 displays the RMSRs for two of the forms, Forms 1A and 3C. The y axis represents the RMSR, in thousandths; and the x axis represents exploratory factor models up to the sixth factor model. The steep decrease in magnitude of the root mean square residuals from the one-factor to the two-factor exploratory model, and the relatively small decreases in the magnitudes of the RMSR in going to higher order solutions, suggests that improvements in fit provided by the higher order solutions are of much less practical consequence.

The initial findings of this study suggest that if a second factor is involved in the reading comprehension section, it appears not to be associated specifically with the reasoning items. Rather, the oblique rotations of the two-factor exploratory NOHARM II solutions and the follow-up confirmatory two-factor analyses both suggest that the specific factor might be related to passage content or passage position. In addition, Stout's dimensionality tests employing the automatic procedure to form the AT1 consistently rejected the hypothesis of essential unidimensionality.

Summary Statistics

Table 4 presents the item difficulty summary statistics for each passage set in test Forms 1A through 3D. The table lists the average IRT b-parameter estimates and average delta and their standard deviations for each set. Because 1A through 1D, 2A and 2B, and 3A through 3D are subtest versions of three final forms (that is, each group of tests contains the same final form but different pretests), and IRT calibrations were performed on the aggregated data for the operational items in each of the forms, the b-parameter estimates are identical for the final form items in each group of forms. The deltas for the final form items in each group of test forms are also identical. It can be seen from Table 4 that difficulty does not account for the second dimension. The most difficult passage sets are not generally sets 5 and 6, the sets that appear to load on the second dimension. In terms of delta, the classical difficulty index, set 6 was the most difficult (had the highest average delta) in only two of the test forms (1C and 2A); and the IRT b-parameter was highest for set 6 in only four of the forms (1C, 2A, 3B, and 3D). For most of the forms tested, the most difficult set

was within the first four sets. These results indicate the second factor is not related to difficulty but possibly to something about the content or position of the passages.

Discussion

Reasoning Items Do Not Comprise a Second Dimension

The major objective of this study was to provide information related to the dimensionality of TOEFL reading comprehension items, specifically in terms of the current distinction in test specifications between four types of items identified as reasoning items, that is, items testing (1) analogy, (2) extrapolation, (3) organization and logic, and (4) author's purpose/attitude, and reading comprehension items of all other types. Results of the analyses indicate that such a distinction is not useful for the purpose of weighting test specifications. These four item types could be weighted individually, as other item types in the reading comprehension test specifications are, in terms of the contributions they are believed to provide individually to overall measurement. A distinction between reasoning and nonreasoning items may of course still be a useful part of content specifications for other reasons.

The study's secondary objective was to provide information relevant to design decisions for the research and development of a new TOEFL test. It is possible that some of the tasks identified as part of the possible academic language domain for a new TOEFL might involve cognitive or other abilities that are related to reading comprehension but that go beyond language abilities alone. This issue is clearly related to the ongoing discussion among reading specialists about the existence of higher and lower level reading skills. In failing to substantiate the hypothesis that items identified in test specifications as reasoning items measure an additional ability different from the rest of the reading items in the TOEFL reading test, this study could be seen as lending some support to the conclusions of Lunzer, Waite, and Dolan (1979) that comprehension does not involve separate subskills, and that separate levels of comprehension are not evident for higher and lower level tasks.

However, while these results provide an initial piece of evidence about the unidimensionality of the current reading comprehension test in terms of the current item types, their relevance for a cognitive theory of reading ability should not be assumed. The fact that the items identified as reasoning items in this study did not form a separate factor or dimension from items identified as nonreasoning items does not mean that such a distinction is necessarily irrelevant, or that separate subskills do not exist in reading comprehension. Conceptually distinct subskills may exist, even though psychometric modeling does not indicate that separate latent abilities are needed to characterize the performance differences on the test for the TOEFL test-taking population. Latent ability may be a composite of conceptually distinct subskills. Reasoning and nonreasoning subskills may exist separately, as they do conceptually, but be so highly correlated in this data that they do not define separate factors or dimensions. There may also be other subskills in reading that are not represented in this data set.

Passage Content or Passage Position Effects May Comprise a Second Dimension.

Although the hypothesized confirmatory two-factor model with reasoning items defining the second dimension did not fit the data in this study, the exploratory two-factor analyses suggest the existence of a minor secondary factor, possibly related to passage content or passage position effects. All 10 of the tests violated the assumption of essential unidimensionality, none of the forms fit an exploratory one-factor model, and all of them appeared to fit a two-factor model in which the magnitude of the loadings on the second factor was greater for the items in the last two passages included in the analyses. Additionally, results indicated that the second factor is unrelated to difficulty of the passage sets.

Conclusions

The results of this study failed to show that reasoning items measure a separate construct in addition to that being measured by the remaining items in the reading subpart of the TOEFL test. However, the results of the exploratory analyses suggest the existence of a minor secondary factor, possibly related to passage content or passage position effects. Areas for further research might include the examination of passage content as it relates to dimensionality. For example, passages loading on factors 1 and 2 in this study could be examined for content differences that might form the basis for hypotheses related to the two factors. Additional passages with similar content characteristics could be identified for testing such hypotheses, which could then be confirmed or rejected based on the extent to which the appearance of the two factors coincided with predictions based on the content analyses. Another area for further research would be the investigation of passage position effects that might be related to the timing of the test or to examinee fatigue. Since this study analyzed longer operational test forms with embedded pretests, it would be valuable to investigate some TOEFL reading tests that do not have embedded pretest passages to see whether similar types of factor analytic results are demonstrated. The results of such research could provide valuable information related to the content, development, and administration of the current test. Such research could also provide information relevant to design decisions for a new TOEFL test.

Table 1

Stout's Tests of Essential Unidimensionality

Test Form	Number of Examinees	Number of items	Stout's *T* Confirmatory	Probability	Stout's *T* Exploratory	Probability
1A	1642	39	0.665	0.253	4.660	0.000
1B	1818	39	-0.307	0.621	5.535	0.000
1C	1827	39	-1.843	0.967	4.226	0.000
1D	1580	39	0.547	0.292	3.367	0.000
2A	1238	38	-1.517	0.935	4.751	0.000
2B	1032	38	-0.312	0.622	4.193	0.000
3A	1076	39	0.016	0.493	3.530	0.000
3B	1068	39	-1.523	0.936	2.016	0.022
3C	1047	39	0.685	0.247	4.307	0.000
3D	1069	39	1.475	0.070	3.658	0.000

Table 2

Multidimensional NOHARM Analyses

Test Form	RMSR Confirmatory Reasoning	RMSR Confirmatory Passages	RMSR Exploratory 1-Factor Model	RMSR Exploratory 2-Factor Model	r_{12}
1A	0.00764	0.00650	0.00767	0.00630	0.74
1B	0.00722	0.00586	0.00724	0.00567	0.74
1C	0.00681	0.00582	0.00684	0.00564	0.77
1D	0.00741	0.00645	0.00751	0.00619	0.78
2A	0.00760	0.00636	0.00759	0.00620	0.70
2B	0.00777	0.00663	0.00782	0.00636	0.71
3A	0.00828	0.00783	0.00837	0.00737	0.76
3B	0.00714	0.00680	0.00718	0.00651	0.73
3C	0.00960	0.00833	0.00971	0.00805	0.62
3D	0.00866	0.00801	0.00877	0.00745	0.70

Table 3a

Factor Loadings for Form 1C
Exploratory Two-Factor Run

Set	Factor1	Factor2
1	0.556	-0.156
1	0.528	0.154
1	0.451	0.093
1	0.609	-0.013
1	0.504	0.224
1	0.683	0.019
2	0.492	0.052
2	0.505	0.156
2	1.142	-0.373
2	0.652	-0.010
2	0.670	0.053
2	0.937	-0.213
2	0.607	0.441
2	0.849	-0.124
3	0.920	-0.242
3	0.405	0.134
3	0.628	0.038
3	0.713	0.037
3	0.919	-0.155
3	0.524	0.017
3	0.491	0.186
4	0.542	0.168
4	0.653	0.274
4	0.454	0.200
4	0.385	0.324
4	0.454	0.307
4	0.490	0.121
5	0.139	0.507
5	0.222	0.505
5	0.193	0.642
5	0.220	0.503
5	0.239	0.611
5	0.235	0.568
6	-0.015	0.836
6	-0.224	1.106
6	-0.380	1.175
6	-0.192	0.990
6	0.018	0.947
6	0.363	0.589

Table 3b

Factor Loadings for Form 2A
Exploratory Two-Factor Run

Set	Factor1	Factor2
1	0.841	-0.344
1	0.597	0.020
1	0.834	-0.133
1	0.799	-0.218
1	0.821	-0.031
1	0.659	0.089
2	0.652	0.103
2	0.490	0.210
2	0.670	0.013
2	0.579	0.009
3	0.506	0.032
3	0.582	0.062
3	0.445	0.093
3	0.723	0.063
3	0.759	-0.092
3	0.929	0.038
3	0.639	0.108
4	0.735	0.306
4	0.770	0.105
4	0.458	0.153
4	0.585	0.055
4	0.657	0.093
4	0.308	0.365
4	0.087	0.117
4	0.313	0.288
5	0.117	0.443
5	0.156	0.634
5	0.411	0.299
5	0.114	0.471
5	0.233	0.166
5	0.044	0.620
5	0.344	0.613
6	-0.115	0.929
6	0.069	0.736
6	-0.129	0.791
6	-0.134	1.017
6	0.169	0.832
6	-0.033	0.710

Table 3c

Factor Loadings for Form 3C
Exploratory Two-Factor Run

Set	Factor1	Factor2
1	0.955	-0.392
1	0.639	0.077
1	0.817	-0.163
1	0.791	-0.197
1	0.769	-0.404
1	0.634	0.032
2	0.506	-0.081
2	0.730	-0.066
2	0.388	0.098
2	0.640	0.105
2	0.617	0.026
2	0.447	0.046
2	0.446	0.060
3	0.563	0.193
3	0.416	0.159
3	0.684	0.053
3	0.690	0.129
3	0.534	0.290
3	0.287	-0.021
3	0.499	0.200
3	0.794	0.193
4	0.241	0.312
4	0.397	0.343
4	0.142	0.343
4	0.252	0.371
4	0.203	0.126
4	0.238	0.472
4	0.512	0.443
5	-0.133	0.801
5	-0.077	0.665
5	-0.250	0.988
5	0.255	0.383
5	-0.130	0.940
5	0.225	0.687
6	-0.061	0.635
6	0.022	0.589
6	-0.015	0.597
6	-0.034	0.459
6	-0.005	0.728

Table 4

Summary Statistics for Sets

		Set 1		Set 2		Set 3		Set 4		Set 5		Set 6	
		B	Delta	B	Delta	B	Delta	B	Delta	B	Delta	B	Delta
FORMS 1A - 1D													
1A	MEAN	0.05	12.2	0.46	13.0	0.38	12.9	-0.29	11.8	0.16	12.1	0.21	12.3
	SD	0.71	1.2	0.35	0.6	0.41	1.0	0.44	1.2	0.24	0.6	0.17	0.5
1B	MEAN	0.05	12.2	0.51	13.0	0.00	12.3	-0.29	11.8	0.16	12.1	0.21	12.3
	SD	0.71	1.2	0.39	0.7	0.46	0.8	0.44	1.2	0.24	0.6	0.17	0.5
1C	MEAN	0.05	12.2	-0.21	11.8	-0.07	11.8	-0.29	11.8	0.16	12.1	0.21	12.3
	SD	0.71	1.2	0.66	1.5	0.66	1.2	0.44	1.2	0.24	0.6	0.17	0.5
1D	MEAN	0.05	12.2	0.14	12.6	0.37	12.7	-0.29	11.8	0.16	12.1	0.21	12.3
	SD	0.71	1.2	0.47	1.0	0.39	1.0	0.44	1.2	0.24	0.6	0.17	0.5
FORMS 2A & 2B													
2A	MEAN	-0.03	12.0	0.36	12.6	0.02	12.2	0.45	13.2	0.47	12.7	0.63	13.2
	SD	0.41	0.8	0.57	1.0	0.77	1.4	0.72	0.8	0.32	0.5	0.35	0.7
2B	MEAN	-0.03	12.0	0.36	12.6	0.02	12.2	0.54	12.8	0.82	13.5	0.63	13.2
	SD	0.41	0.8	0.57	1.0	0.77	1.4	0.73	1.0	0.67	1.5	0.35	0.7
FORMS 3A - 3D													
3A	MEAN	-0.10	12.1	0.22	12.5	0.93	13.5	0.56	13.2	0.27	12.8	0.58	13.0
	SD	0.50	1.0	0.53	1.0	0.83	0.7	0.57	0.8	0.34	0.7	0.58	0.7
3B	MEAN	-0.10	12.1	0.22	12.5	0.36	13.1	0.44	13.0	0.27	12.8	0.58	13.0
	SD	0.50	1.0	0.53	1.0	0.78	1.1	1.11	1.5	0.34	0.7	0.58	0.7
3C	MEAN	-0.10	12.1	0.22	12.5	0.51	13.4	0.73	13.5	0.27	12.8	0.58	13.0
	SD	0.50	1.0	0.53	1.0	0.39	0.7	0.53	0.8	0.34	0.7	0.58	0.7
3D	MEAN	-0.10	12.1	0.22	12.5	0.61	13.5	0.44	12.8	0.27	12.8	0.58	13.0
	SD	0.50	1.0	0.53	1.0	0.35	0.7	0.59	1.2	0.34	0.7	0.58	0.7

Figure 1

Exploratory 2-Factor Analysis
Form 1C Reasoning Items

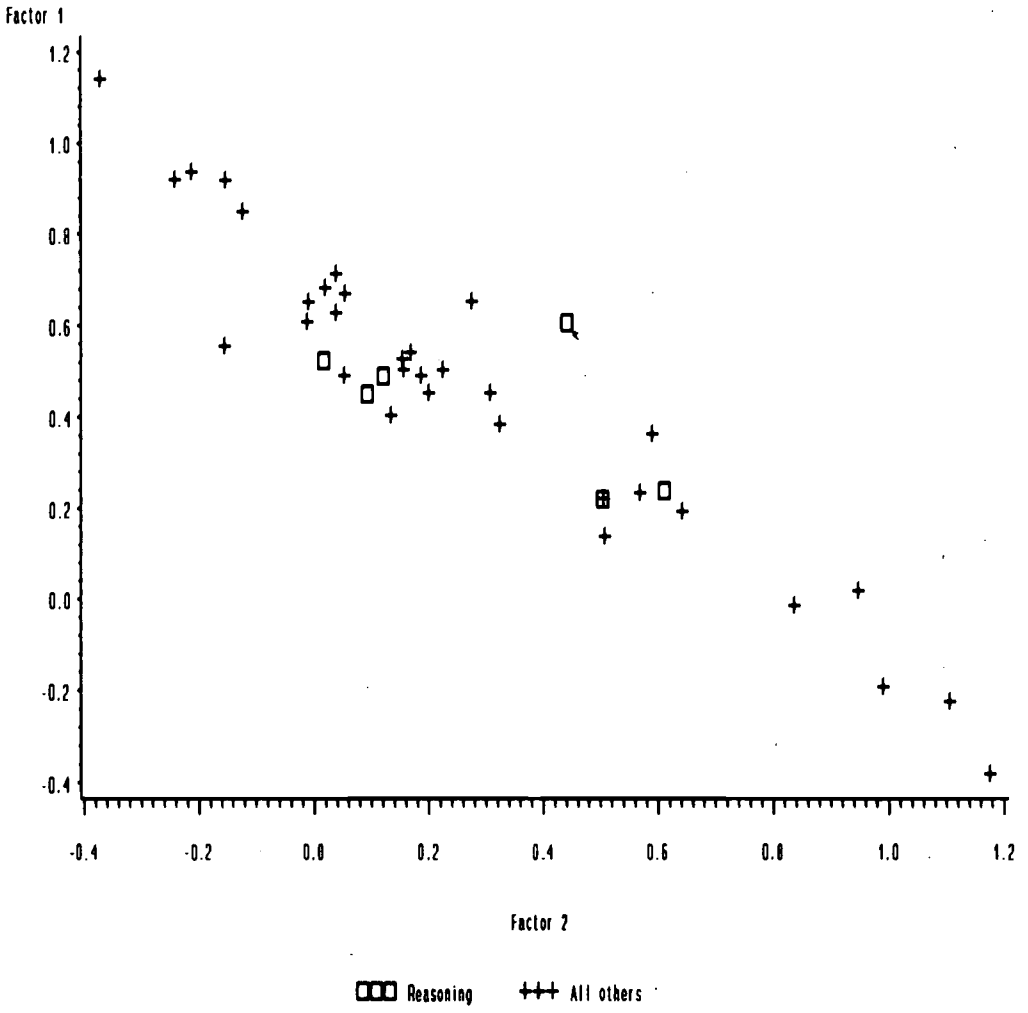


Figure 2

Exploratory 2-Factor Analysis
Form 2A Reasoning Items

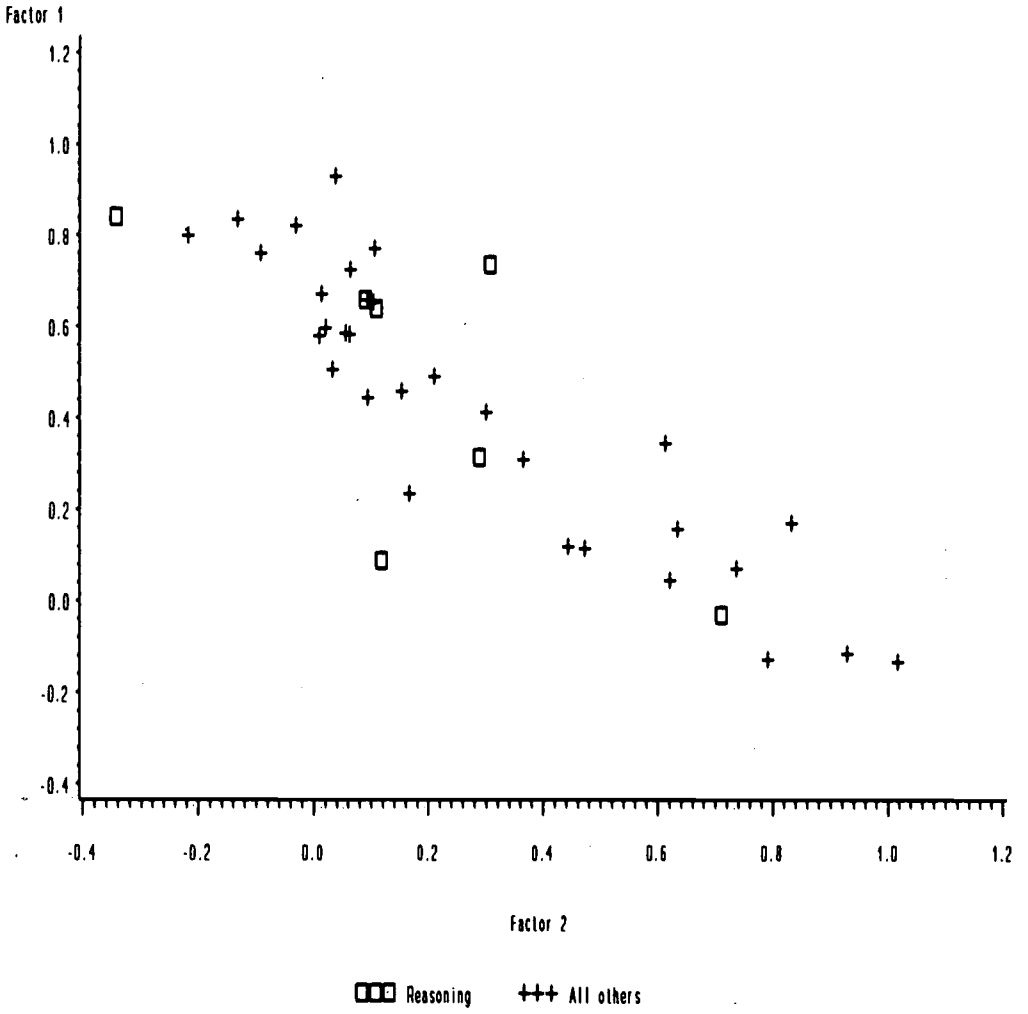


Figure 3
Exploratory 2-Factor Analysis
Form 3C Reasoning Items

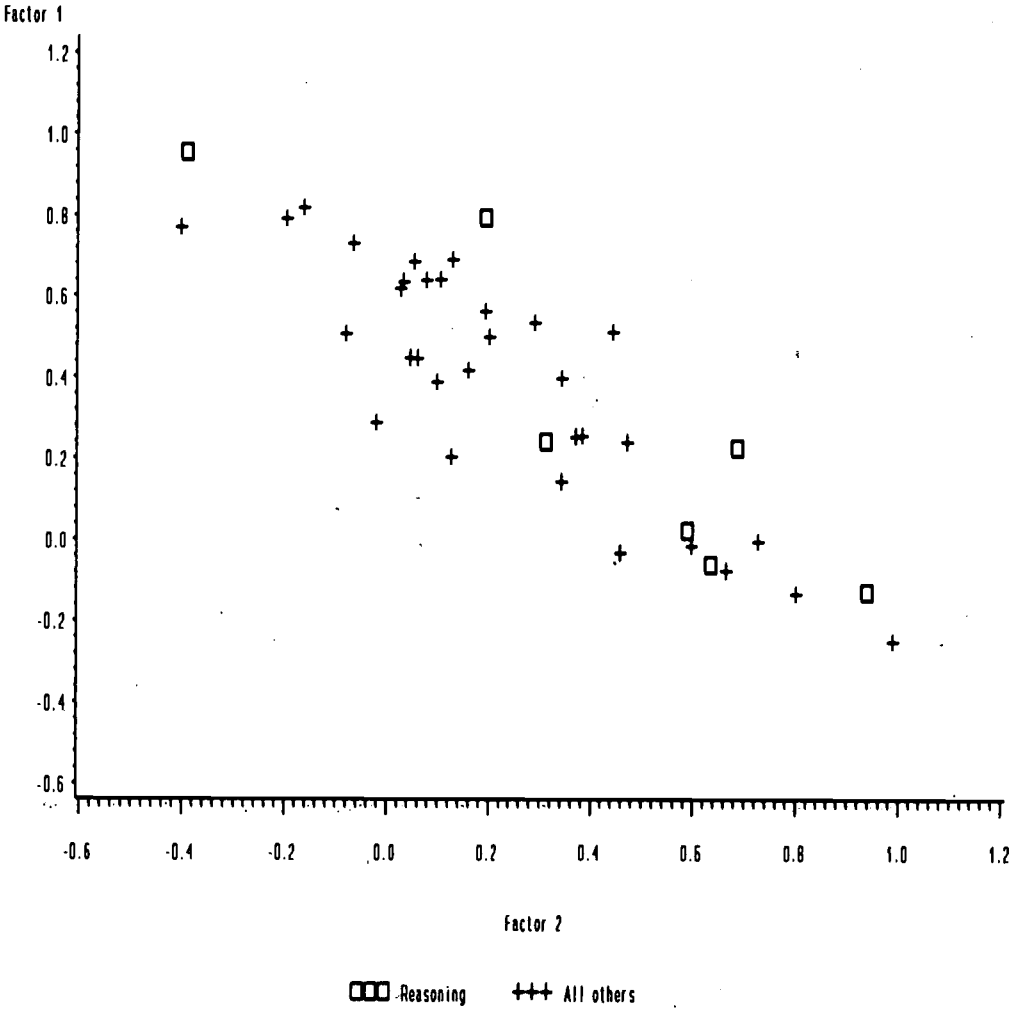


Figure 4

Exploratory 2-Factor Analysis
Form 1C Sets 5 and 6

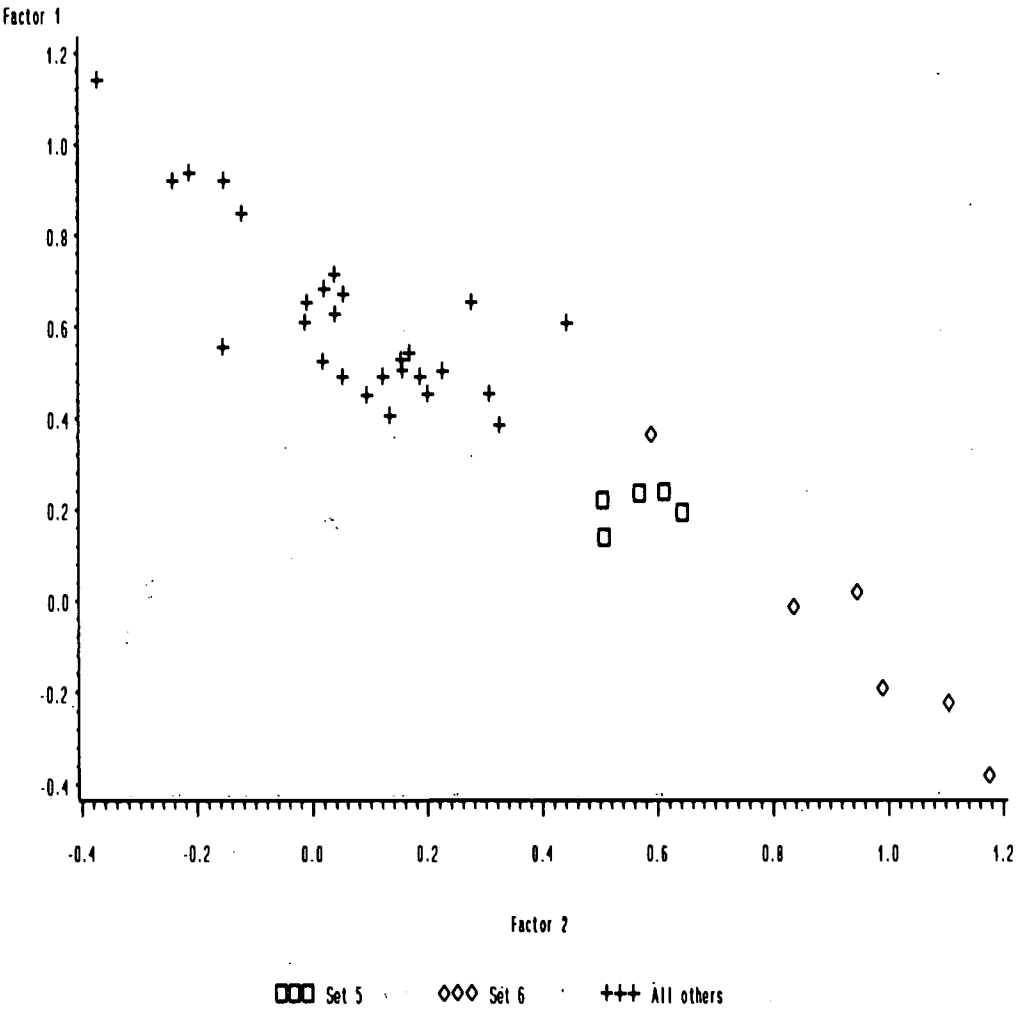


Figure 5

Exploratory 2-Factor Analysis
Form 2A Sets 5 and 6

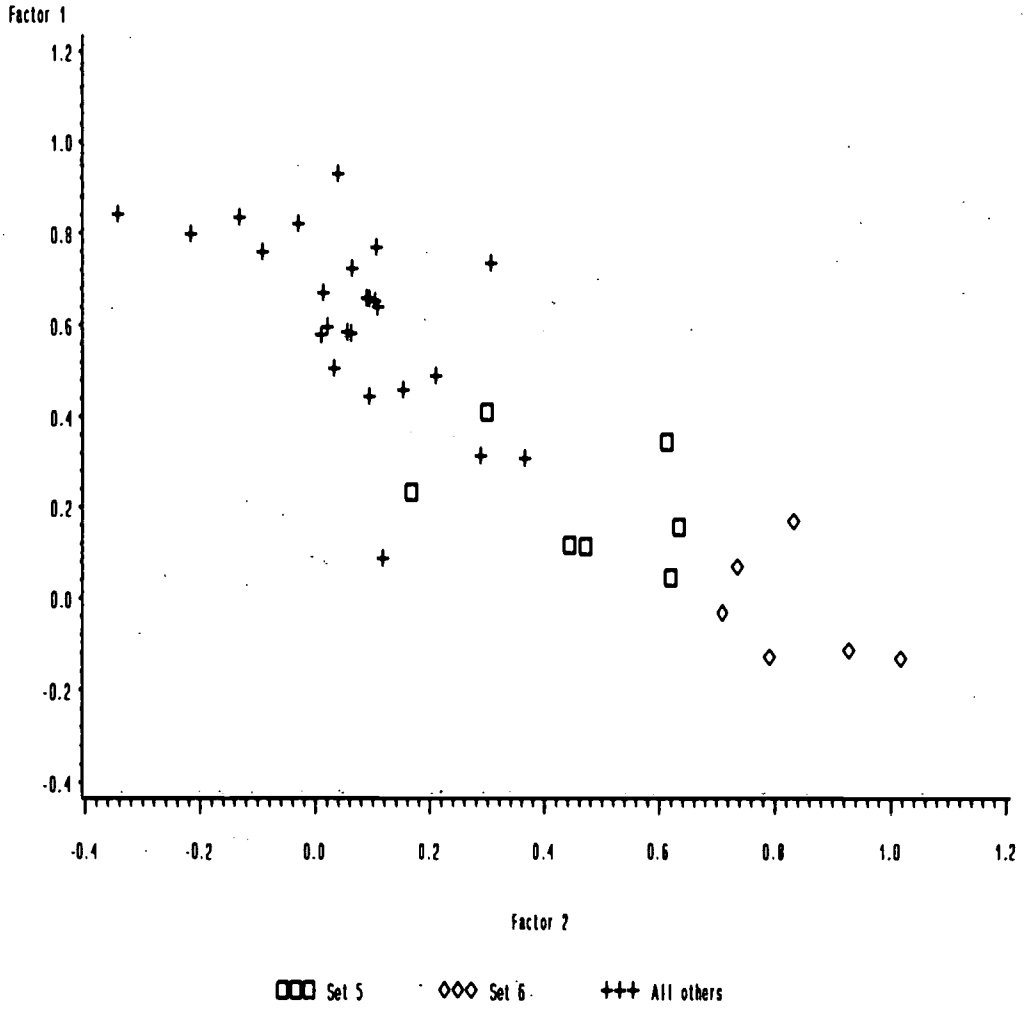


Figure 6

Exploratory 2-Factor Analysis
Form 3C Sets 5 and 6

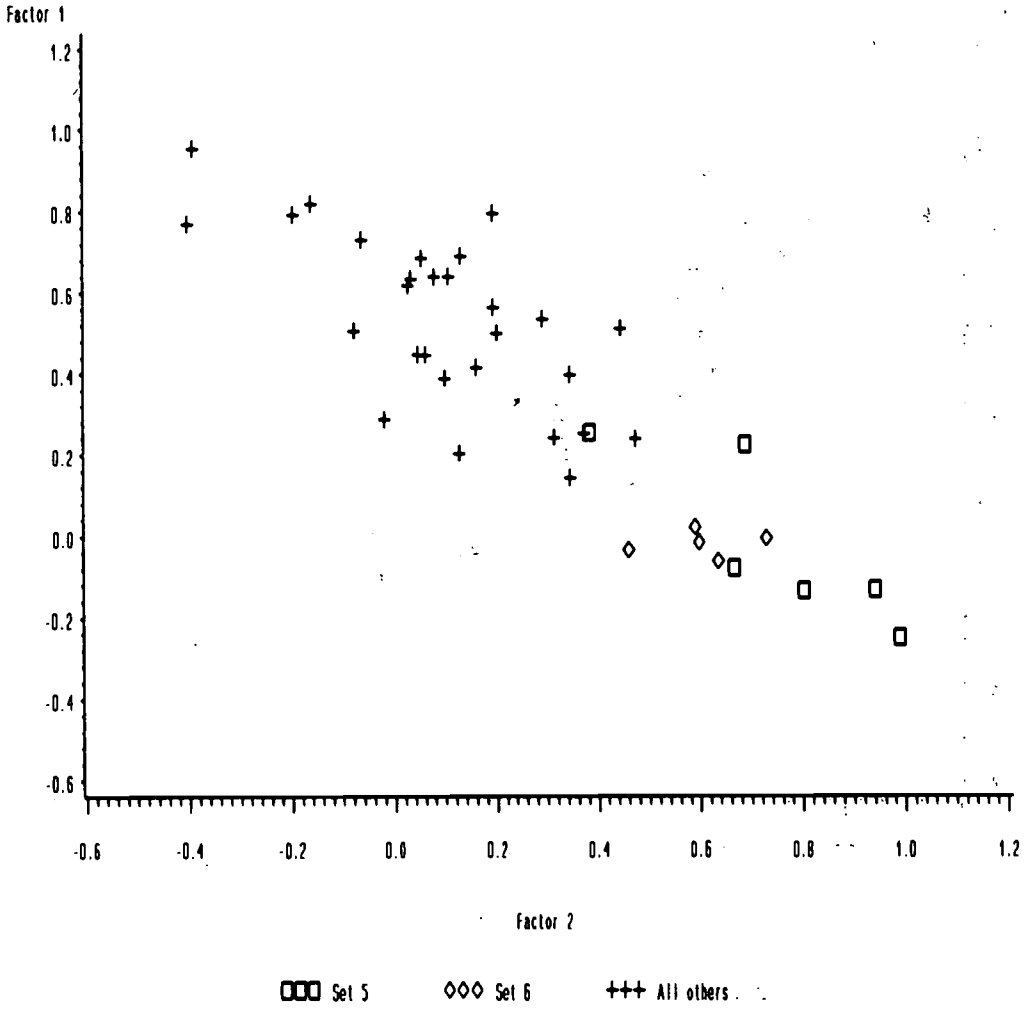
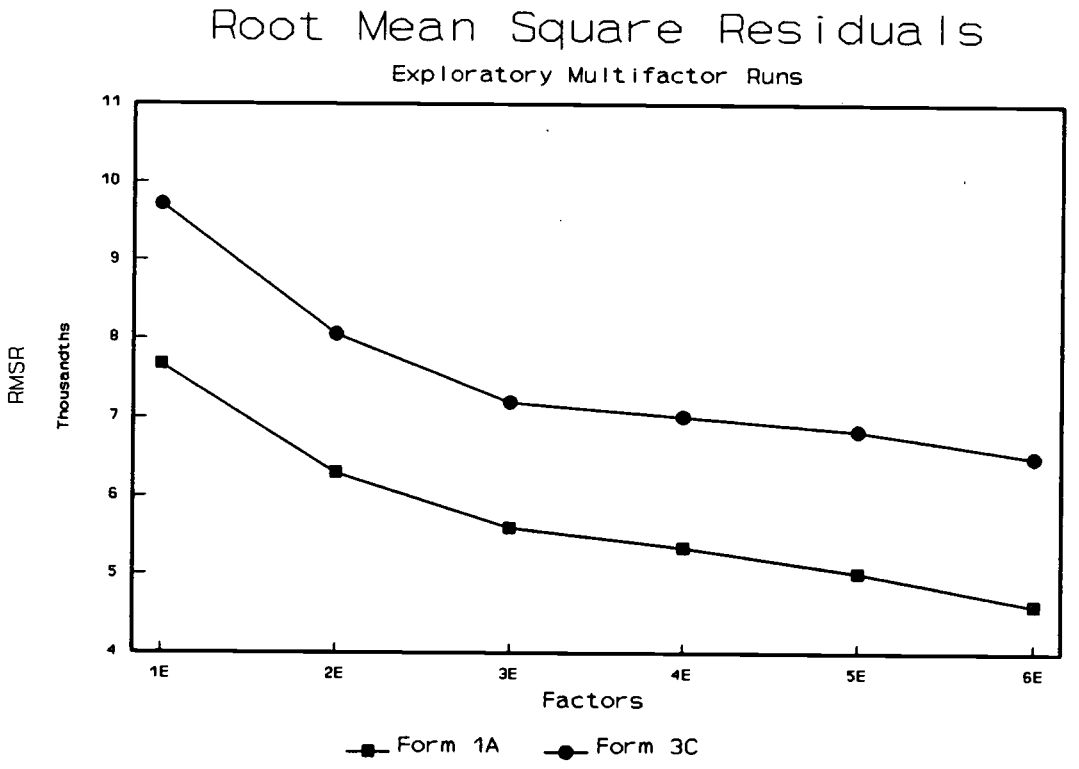


Figure 7


Root Mean Square Residuals
Exploratory Multifactor Runs
RMSR
Thousandths
11
10
9
8
7
6
5
4
1E
2E
3E
4E
5E
6E
Factors
Form 1A
Form 3C

References

Alderson, J., & Lukmani, Y. (1989). Cognition and reading: cognitive levels as embodied in test questions. Reading in a Foreign Language, 5(2), 353-270.

Barrett, T. C. (1968). What is reading? In T. Clymer (Ed.), Innovation and change in reading instruction. 67th Yearbook of the National Society for the Study of Education, University of Chicago Press.

Davies, A., & Widdowson, H. (1974). The teaching of reading and writing. In J. P. B. Allen and S. P. Corder (Eds.), Techniques in applied linguistics, (Vol. 3). Oxford University Press.

Davis, F. B. (1968). Research in comprehension in reading. Reading Research Quarterly, 3, 499-545.

De Champlain, A. F. (1992). Assessing test dimensionality using two approximate chi-square statistics. Unpublished Dissertation, University of Ottawa, Canada.

Dunbar, S. B. (1982, April). Construct validity and the internal structure of a foreign language test for several native language groups. Paper presented at the annual meeting of the American Educational Research Association, New York.

Fraser, C. (1983). NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, Australia: The University of New England, Center for Behavioral Studies.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. Multivariate Behavioral Research, 23, 267-269.

Grabe, W. (1986). The transition from theory to practice in teaching reading. In F. Dubin, D. Eskey, W. Grabe (Eds.), Teaching second language reading for academic purposes. Reading, MA: Addison-Wesley.

Hale, G. A., Rock, D. A., & Jirele, T. (1989). Confirmatory factor analysis of the Test of English as a Foreign Language (TOEFL Research Report No. 32). Princeton, NJ: Educational Testing Service.

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W. (1988). Multiple-choice cloze items and the Test of English as a Foreign Language (TOEFL Research Report 26). Princeton, NJ: Educational Testing Service.

Lunzer, E., Waite, M., & Dolan, T. (1979). Comprehension and comprehension tests. In E. Lunzer and K. Gardner (Eds.), The effective use of reading. Heinemann Educational Books.

McDonald, R. P. (1967). Nonlinear factor analysis. Psychometric Monographs (No. 15).

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. Applied Psychological Measurement, 6, 379-396.

McDonald, R. P. (1983). Exploratory and confirmatory factor analysis. In H. Wainer & S. Messick (Eds.), Principles of modern psychological measurement. Hillsdale, NJ: Lawrence Erlbaum.

McKinley, R., & Way, W. (1992). The feasibility of modeling secondary TOEFL ability dimensions using multidimensional IRT models (TOEFL Technical Report TR-5). Princeton, NJ: Educational Testing Service.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. Journal of Educational Measurement, 28, 99-118.

Oltman, P. K., Stricker, L. J., & Barrows, T. (1988). Native language, English proficiency, and the structure of the Test of English as a Foreign Language (TOEFL Research Report No. 27). Princeton, NJ: Educational Testing Service.

Schedl, M., Thomas, N., & Way, W. (1995). An investigation of proposed revisions to section 3 of the TOEFL test (TOEFL Research Report No. 47). Princeton, NJ: Educational Testing Service.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Cover Printed on Recycled Paper

57906-09070 • Y36M.5 • 275711 • Printed in U.S.A.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").