DOCUMENT RESUME

ED 385 559 TM 023 981

AUTHOR	Golub-Smith, Marna; And Others
TITLE	Topic and Topic Type Comparability on the Test of Written English.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-93-10; TOEFL-RR-42
PUB DATE	Mar 93
NOTE	48p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Charts; Comparative Analysis; *English (Second Language); Essay Tests; Graphs; *Scoring; Writing (Composition); *Writing Evaluation
IDENTIFIERS	*Essay Topics; Explicitness; Test of English as a Foreign Language; *Test of Written English; Writing Prompts

ABSTRACT

The Test of Written English (TWE), administered with certain designated examinations of the Test of English as a Foreign Language (TOEFL), consists of a single essay prompt to which examinees have 30 minutes to respond. Questions have been raised about the comparability of different TWE prompts. This study was designed to elicit essays for prompts that differed both in subject matter (topic) and in level of explicitness with which the essay task was presented (topic type). Eight different prompts were spiraled worldwide at the October 1989 TOEFL administration, with each prompt eliciting approximately 10,000 essays. Results of the analyses indicated that there were small differences among the prompts. The most notable differences were obtained among the scores for topics using the most explicit comparison. Across all the prompts, the chart-graph topic with the explicit comparison statement produced the highest mean scores. Because it was the first study to focus on the comparability of prompts in a major testing program, the authors had difficulty making definitive statements about the meaningfulness of the obtained differences. Such differences may warrant further consideration by the TOEFL program. Two appendixes present prompts and scoring guidelines. Fifteen tables present analysis results. (Contains 17 references.) (Author/SLD)

TOEFL®

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 42
March 1993

Topic and Topic Type Comparability on the Test of Written English

Marna Golub-Smith
Clyde Reese
Karin Steinhaus

ETS
Educational Testing Service

Topic and Topic Type Comparability on the Test of Written English

Marna Golub-Smith
Clyde Reese
Karin Steinhaus

Educational Testing Service
Princeton, New Jersey

RR-93-10

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

Abstract

The Test of Written English (TWE), administered with certain designated TOEFL® examinations, consists of a single essay prompt to which examinees have 30 minutes to respond. It was introduced in 1986 to provide TOEFL score users with a direct measure of examinees' writing ability. Preliminary studies had indicated that the two different kinds of prompts: *prose compare, contrast, and take a position,* and *describe or interpret a chart or graph,* elicited comparable writing performance. However, questions were subsequently raised with respect to continued comparability of different TWE® prompts administered under operational conditions.

The present study was designed to elicit essays for prompts that differed in both subject matter (Topic) and in the level of explicitness with which the essay task was presented (Topic Type). Eight different prompts were spiraled worldwide at the October 1989 TOEFL administration, with each prompt eliciting approximately 10,000 essays.

The results of the analyses indicated that there were small differences among the prompts. The most notable differences were obtained among the scores for topics using the explicit comparison. Across all the prompts, the chart-graph with the explicit comparison statement produced the highest mean scores.

Because it was the first study of its kind to focus on the comparability of prompts in a major testing program, the authors had difficulty making definitive statements regarding the meaningfulness of the obtained differences. While many of the differences in means observed in this study were so small as to be of no practical significance, differences observed across prompts in the numbers of examinees at each score level were not. Such differences may warrant further consideration by the TOEFL program.

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, which was formed through the cooperative effort of more than thirty organizations, public and private, that were 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, and 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 and 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. Currently 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. However, many projects require the cooperation of other institutions, 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 the confidentiality of data will be protected.

Current (1991-92) members of the TOEFL Research Committee are:

James Dean Brown	University of Hawaii
Patricia Dunkel (Chair)	Pennsylvania State University
William Grabe	Northern Arizona University
Kyle Perkins	Southern Illinois University at Carbondale
Elizabeth C. Traugott	Stanford University
John Upshur	Concordia University

Acknowledgments

The authors thank Patricia Carey and Stephen Laue for programming and data analysis support, and Nancy Petersen, Walter Way, Gordon Hale, and Hunter Breland for reviewing earlier drafts of this manuscript.

Table of Contents

List of Appendices

List of Figures

List of Tables

Introduction

The Test of Written English (TWE) is the essay component of the Test of English as a Foreign Language (TOEFL), the multiple-choice test used by more than 3,000 institutions to evaluate the English proficiency of applicants whose native language is not English. As a direct, productive skills test, the TWE complements TOEFL's Section II (Structure and Written Expression), an indirect test of an examinee's knowledge of important structural and grammatical points of standard written English. The TWE uses a holistic score to provide information about the examinee's ability to generate and organize ideas on paper, to support those ideas with examples or evidence, and to use the conventions of standard written English. Introduced in July 1986, the TWE was developed in response to requests from TOEFL score users for a direct measure of examinees' writing ability. It is currently a required component in five administrations each year of the regular TOEFL examination.

The 30-minute test requires examinees to write an essay in response to a single prompt. The two kinds of prompts that were being administered at the time this study began were *(1) compare-contrast and take a position* and *(2) interpret a chart or a graph*. For security reasons, several different prompts may be administered at any given TWE administration.

During the first few years of its operation, the TWE was considered experimental. To provide evidence that the test is a reliable and valid measure of English writing proficiency, it has been important to undertake significant research on technical matters related to the psychometric quality of the TWE. In their long-term research agenda for the TWE, Stansfield and Ross (1988) state that the greatest priority should be given to the issue of the comparability of scores obtained on different prompts. Since no statistical method now exists that controls for difficulty when only one prompt is administered to each examinee, comparisons of the difficulties of different kinds of prompts would demonstrate the extent to which prompt difficulty is being controlled through careful development and scoring practices.

The initial prompt formats that were proposed followed the recommendations of the Bridgeman and Carlson (1983) survey of the academic writing tasks and skills required of students at major universities. The basic design of the TWE was developed and validated in the study done by Carlson, Bridgeman, Camp, and Waanders (1985), in which scoring methods and the performance of different formats were investigated.

That study found that these format were comparable in terms of difficulty. However, the formats may have been comparable because essay readers of the original topics had to take a predominantly norm-referenced approach, relying only on sample "range finder" papers to illustrate the six score levels. Today's TWE essay readers refer to a criterion-referenced scoring guide to help them focus on characteristics that are intended to ensure uniform evaluation of the essays. The use of a scoring guide helps prevent readers from being influenced by the relative quality of the essays they read. Without a scoring guide, an essay could receive a higher score if it were read

along with many poorly written papers or it could receive a lower score if it were read along with many well-written papers.

Early questions about prompt development

TWE prompts are currently developed by ETS staff and by the TWE Core Reader Group which is an external group of writing specialists who serve as consultants to the program. After the prompts are pretested, essays elicited by each prompt are reviewed by the TWE Core Reader Group and are either approved for administration or rejected. As the group has gained experience with examinees' responses to individual prompts, it has questioned a number of issues related to the comparability of the two kinds of prompts and the effects of variations of wording within similar prompts.

The two kinds of prompts involve different kinds of tasks. A compare-contrast prose prompt sets up a contrast and requires the examinee to compare some relatively abstract concepts and to construct an essay based on individual background--personal experience or opinion. A chart-graph prompt presents graphic information on which an essay will be based, yet the content of a TWE chart or graph must not require the examinee to have any particular background knowledge in order to discuss the issues involved. The format of the chart or graph also must be simple enough to be interpreted by those examinees unaccustomed to dealing with graphics.

Although the Carlson and Bridgeman (1983) study indicated that an examinee was likely to receive a similar writing score for essays written on either kind of prompt, it remained clear that the writing tasks differ significantly. The group felt that it was especially important to investigate any impact the differences might have because the TWE was being administered as a single-item test with the two kinds of prompts being used interchangeably.

There was concern that the scoring guide might not be performing equally well on essays elicited by the two kinds of prompts. To help ensure that the scoring guide could be applied equally effectively to both kinds of prompts, the TWE Core Reader Group gradually began to develop chart-graph prompts with texts that strongly resembled the texts of *the compare, contrast and take a position* prompts. Nevertheless, as the group began to read essays that were written in response to both kinds of pretested prompts, they began to speculate that the responses to the chart-graph prompts might differ qualitatively from those elicited by prose prompts. This led the group to further question the comparability of these two kinds of prompts.

TWE Core Readers and Reading Management staff have observed that it is possible for examinees with low English proficiency to use language that is provided in a chart or graph and string it together into sentences, so that relatively few very low-level essays are produced in response to chart-graph prompts. At the same time, the process of describing the graphic seems to limit the creativity of the better writers, thus resulting in fewer very high-level essays than might be expected from compare-contrast prompts.

TWE Core Readers suspected that the way a prompt is worded may have an effect on the way examinees respond in their essays. Over time, the group also became aware that variations in wording might affect responses to similar prompts. The group considers it desirable to produce prompts that are brief but that are also very specific about the task that examinees are expected to perform in their essays. The goals of brevity and specificity often conflict, however, and the group wanted evidence to help determine which principles of item development that are most appropriate for the TOEFL population.

To help determine optimal levels of specificity, the TWE Core Reader Group had already begun to develop prompts with alternate versions. One version was longer and included a *compare* directive explicitly stated in the prompt; the shorter version had a comparison implied by the general content and organization of the prompt. The group hoped that pretest results would indicate whether variations in wording would produce predictable variations in examinee writing.

The nature of the subject matter of a prompt may have a strong effect on how well individual examinees are able to respond to the prompt, regardless of their general level of English skills. The nature of the subject matter was another variable that the group wished to investigate. The TWE Core Reader Group speculated that decisions about the appropriate length and level of explicitness of any prompt might depend, to a certain extent, on the topic of the prompt. Thus, prompts could be very similar in their wording and yet might elicit very different writing, depending on the subject matter involved. The group was interested in finding out how much real difference variations in subject matter might make on an examinee's TWE score.

The effectiveness of the TWE program might be diminished if TWE prompts resemble each other too closely. The concern for maintaining consistency has been complicated, however, by an equally serious concern about cumulative effects that similarities among prompts might have on examinee performance. It was becoming increasingly difficult for the group to generate fresh prompts of the approved types. Group members felt their best ideas were used in the earliest prompts, and their capacity to generate an adequate supply of interesting new prompts seemed in danger of diminishing. The group was also concerned about the likelihood of a "backwash effect"--the possibility that writing instruction (especially in TWE preparation courses) would be reduced to practicing only two kinds of prompts. Therefore, in addition to wanting to know more about the impact of <u>slight variation</u> in wording in any given prompt, the group also wanted some guidelines for determining characteristics of the <u>maximum acceptable variation</u> among the formats and wording of TWE prompts to ensure consistent scoring of essays elicited by clearly different kinds of prompts.

<u>Expectations about the value of the present study</u>

As work on the TWE moved away from the small scale of the validation study and headed toward large-scale operational testing conditions, questions about the equivalency of prompts continued to arise. Therefore, it would be valuable to try to replicate parts of the Carlson and Bridgeman (1983) study to see

whether examinee performance on different prompts would still prove comparable if the researchers took into consideration practical lessons that had been learned during the initial experience with the TWE testing program.

The possibility of differences in prompt difficulty was suggested by the mean scores for the July 1986 and July 1987 administrations of the TWE. The July 1986 administration consisted of one compare-contrast prompt administered worldwide and the July 1987 administration consisted of one chart-graph prompt administered worldwide. For July 1986, the mean TWE score was 3.30 with a standard deviation of 1.01 ($n = 10,413$). For July 1987, the mean was 3.51 with a standard deviation of 0.95 ($n = 10,980$). The difference was significant ($p < .0001$). This indicates that the average performance of examinees was somewhat better on the chart-graph prompt. The mean TOEFL scores for these administrations were fairly comparable: 505.9 for July 1986 and 505.4 for July 1987. Because of the difference between TWE means, and because the TWE Core Reader Group did not feel confident that compare-contrast and chart-graph prompts were eliciting comparable writing samples, the chart-graph format was temporarily discontinued pending further investigation.

In addition to possible differences in the difficulty among individual topics and between the two formats, the group felt that differences in difficulty might also result from variations within each of the formats. For example, for the compare-contrast prose prompt, two types have been used in the TWE: those in which the writing task is explicitly stated in the prompt and those in which the task is implicitly set up by the subject matter and structure of the prompt.

An explicit compare-contrast subtype directly asks examinees to compare and contrast two ideas and take a position and support the position. An implicit compare-contrast prompt asks examinees to take a position and support the position, but does not directly ask the examinee to compare and contrast. In fact, an implicit compare-contrast prompt type does not explicitly require the examinee to discuss two different ideas or concepts. Often the examinee is asked to weigh the positive and negative features of one idea or concept.

There have also been variations in the chart-graph topic format. The chart-graph prompts first designed for use in the TWE might be called "simple" chart-graph prompts because they direct the writer to interpret information provided in the chart or graph, without asking for personal input or choices. A different chart-graph prompt, which evolved with development efforts for the TWE, contains elements of a compare-contrast prompt; examinees are asked to take and support a position using data in the chart or graph. In a sense, this new chart-graph prompt is similar to an explicit or implicit compare-contrast subtype, except that the examinee is provided with information to use in support of a particular position.

Because different formats, different topics, and different topic types can present examinees with tasks of differing levels of difficulty, the issue of comparability is important. How is it possible to ensure that scores obtained at one administration of the TWE are comparable to scores obtained at other administrations, and that scores obtained at the same administration are comparable across the different prompts administered in different regions of

the world? Different formats, topics, and topic types might elicit different writing performances from the same examinee or may promote successful performance for one examinee while impeding successful performance for another. While Carlson et al. (1985) found comparability both within and across format, Freedman and Calfee (1983) found significant differences between scores for different formats and different topics within a given format. Given these disparate findings, research in this area takes on greater importance.

In recent years, the problem of topic comparability has received increased attention (Brossell, 1986; Ruth and Murphy, 1988). Phillips (1987) has stated that topics can be pre-equated by field testing topics and maintaining reader training standards, scoring rubrics, and test specifications. Legg (1987) has recommended a similar set of requirements in order to ensure score reliability. TWE prompt development and scoring procedures attempt to adhere to practices similar to those discussed by Phillips and Legg. For this reason, the likelihood of TWE topics being comparable is increased. However, the interaction between examinee writing ability and format, topic, and topic type is complex. Performance may be influenced by such factors as the examinee's familiarity with the topic, the complexity of the writing task, and the examinee's access to vocabulary and other expository information (Hout, 1990).

As Cooper (1984) has observed, "different topics often require different skills or make different conceptual demands on the candidates" (p. 4). Even when the topic is a constant, the wording of a prompt may affect examinees' performance. Brossell (1983) found significant differences in scores received on essays written in response to prompts that had a common topic but were designed with different rhetorical specifications. Furthermore, results of a comparative study conducted by Quellmalz, Capell, and Chou (1982) indicated that differing writing tasks did not measure the same thing. Referring to the TWE, Greenberg (1986) contends that the compare-contrast and chart-graph formats require "very different cognitive and linguistic skills" (p. 537). Given the potential disparity between formats, topics, and topic types, it is important for developers of the TWE to determine the extent to which present topic development and scoring procedures control for prompt difficulty.

Methods

Sample

The sample consisted of all candidates taking the TOEFL and TWE on October 28, 1989, at both domestic and foreign test centers. Examinees taking the TOEFL and the TWE typically are candidates seeking admission to colleges and universities in the United States and Canada.

At the time of the study, the TWE was administered four times a year--March, May, September, and October--as part of an operational TOEFL administration. The TOEFL was administered without the TWE the other eight months of the year. Therefore, examinees who elected to take the TOEFL during a combined TOEFL-TWE administration may have differed from the overall TOEFL examinee population. To date, an analysis of TOEFL scores from the operational program concerning any self-selection bias in TOEFL-TWE examinees compared to TOEFL examinees would suggest no differences in the two groups.

During the 1989 testing year (January through December), the mean TOEFL total converted score for the TOEFL examinees was 514.5 compared to 517.3 for the TOEFL-TWE examinees. The mean and standard deviation of the TOEFL total converted scores for the October 1989 TOEFL-TWE administration were 520.2 and 66.9 compared to a mean and standard deviation of 515.7 and 67.5 for the remaining three 1989 TOEFL-TWE administrations. The October 1989 TOEFL-TWE administration sample was slightly more able, but the slight difference should not threaten the generalizability of the results from this study.

Design of the Study

The purpose of the study was to examine the comparability of TWE prompts, focusing on two salient characteristics--topic and type. Topic refers to the subject of the prompt, the reference frame used by the examinee to produce an essay. The four topics selected for use in this study were typical of those used in past TWE administrations, focusing on learning styles, hometowns, cities, and children's leisure time. Type refers to the manner in which the essay question was posed, whether the examinee was implicitly or explicitly asked to compare and/or contrast points of view. The two characteristics, topic and type, were crossed to produce eight TWE prompts. A detailed description of the prompt development process is presented in the following section.

Therefore, the basic design was a two-factor, between-subjects design with the two factors being the prompt characteristics described above. The design is presented in Figure 1. The letters in the cells of the design are used to designate the eight topic-type combinations. A copy of each of these prompts is presented in Appendix A.

Figure 1
Experimental Design

TYPE	TOPIC			
	LEARNING STYLES	HOMETOWNS	BAR-CHART - CITIES	CHILDREN'S LEISURE TIME
IMPLICIT	A	B	C	D
EXPLICIT	E	F	G	H

Randomly equivalent samples of approximately 10,000 examinees responded to each of the prompts. The samples of examinees within each of the cells were formed by spiraling the eight prompts worldwide. This was accomplished by packaging the test forms in alternating order so that, upon distribution, random eighths of the examinees received a particular prompt. The effectiveness of the spiraling to produce samples with equivalent language ability was checked using the TOEFL total converted score as a measure of overall ability in English as a second language.

Prompt Development

To speed the process of developing and pretesting prompts that would fit the study, a special three-person topic development group met in May 1988 to develop parallel sets of prompts for use in the study. To ensure continuity with "regular" TWE prompts, the group began by attempting to create variations of prompts already approved by the TWE Core Reader Group for final form use.

The May group developed 30 prompts within four topic areas. Although many of these prompts were not reviewed by the TWE Core Reader Group before pretesting, the essays elicited by all the prompts were reviewed during the pretesting process by the TWE Core Reader Group to determine how well they worked. In the course of this development process, alternate versions of a few additional prompts were also pretested in an attempt to achieve a desirable balance of topics among the prompts to be used in the study.

The TWE Core Reader Group approved only those prompts that met the criteria for inclusion in a regular TOEFL TWE administration. In anticipation of the usually higher failure rate for chart-graph prompts, the May topic development group had developed more experimental chart-graph prompts than prose prompts, hoping that this would improve the chances of obtaining ideal sets for the study. In spite of this, the group was able to approve only one parallel pair of chart-graph prompts from among the pretest sets.

One problem these decisions posed for the study was that the chart-graph format was now confounded within topic. If performance on the chart-graph was different from the prose prompts, one could not unequivocally attribute the differences to the chart-graph format as opposed to the topic addressed.

TWE Scoring Process

Responses to the eight TWE prompts were scored using standard operational procedures for the TWE program. All papers were scored in a central location by trained TWE readers to ensure that standardized reading procedures were maintained.

Prior to the reading, eight rooms were organized, one for each of the eight prompts. Each room was assigned a room leader to oversee the room's activities and several table leaders. The room and table leaders were selected from the most able and experienced readers. In addition to the room-level management, a chief reader was assigned to oversee the activities in all eight rooms. The readers used during the October 1989 reading were assigned to rooms, and tables within rooms, to ensure a balance of experienced/ inexperienced and fast/slow readers. No attempt could be made to randomly assign readers to rooms (prompts). Therefore, any effects associated with reader variations are confounded with the prompts. However, given the training of the readers and the procedures in place, any impact due to reader differences was expected to be minimal.

In preparation for the TWE reading, the reading management team, consisting of the chief reader and the room and table leaders, selected sample papers or exemplar sets. These sets of papers were selected to highlight the six-point TWE scoring guide and were used to train readers prior to and throughout the main and clean-up readings[1] to maintain scoring consistency.

Each essay was independently scored by two readers. Small groups of readers were organized into tables and worked under the direct supervision of a table leader who monitored each reader's performance throughout the process. In addition, each batch of essays was scrambled between the first and second reading to ensure that readers were not unduly influenced by the sequence of essays. The TWE reading process is a criterion process, not a norm-referenced process. Essays were scored using the six-point scoring guide presented in Appendix B. A summary of the main points of that guide is presented in Figure 2.

When the ratings of the two readers differ by less than two points, the score reported for the TWE is the average of the two readers' ratings. Thus, scores range from 1.0 to 6.0 and are reported at .5 intervals. When the ratings of the two readers differ by two or more points, the paper is given to a member of the reading management team (a room or table leader) for adjudication. In the adjudication process, the examinee's score is the average of the room or table leader's rating and the rating closest to it. The rating that is furthest from the room or table leader's rating is discarded. In cases where the room or table leader's rating is equidistant from the two discrepant ratings, the room or table leader's rating is reported

[1]A clean-up reading is held approximately one week after the main reading. During the clean-up reading, essays that arrive late and those that were not processed in the main reading are scored.

as the score for the TWE. Typically, discrepancy rates for the TWE have been extremely low, ranging from .02 to .05.

Figure 2
TWE Scoring Guide

6	The essay demonstrates clear competence in writing on both the rhetorical and syntactic levels, though it may have occasional errors.
5	The essay demonstrates competence in writing on both the rhetorical and syntactic levels, though it will probably have occasional errors.
4	The essay demonstrates minimal competence in writing on both the rhetorical and syntactic levels.
3	The essay demonstrates some developing competence in writing, but it remains flawed on either the rhetorical or syntactic level or both.
2	The essay suggests incompetence in writing.
1	The essay demonstrates incompetence in writing.
No	No-response
Off	The essay has been written off-topic.

Statistical Analyses

The statistical analyses conducted for the present study can be divided into four main areas:

(1) analysis of the scoring procedures including discrepancy rates and reliability estimates

(2) summary of the reported TWE scores by prompt, including frequency distributions and descriptive statistics

(3) statistical comparison of the prompts using analysis of variance techniques with appropriate follow-up comparisons

(4) post hoc statistical comparison of prompts separately for males and females

Analysis of the Scoring Procedures. For each of the eight prompts, the discrepancy rate and correlation between the first and second readers' scores were calculated. A discrepancy is defined as a difference of two or more points in the scores assigned by the two readers. The discrepancy rate is the proportion of discrepant essays excluding off-topic and no-response essays.

The reliability of the reading process was estimated by correlating the assigned scores from the first and second readers. The correlation was then adjusted using the Spearman-Brown Prophecy formula to reflect the use of two readers. The correlations exclude the scores assigned by room and table leaders during the adjudication process for discrepant essays. Therefore, the correlation may tend to slightly underestimate the true reader reliability of the TWE reading. As with the discrepancy rate, off-topic and no-response papers were excluded from the reader reliability analyses.

Summary of the Reported TWE Scores. Separate frequency distributions and descriptive statistics were produced for each of the eight TWE prompts. Reported TWE scores, which included the adjudicated scores for discrepant essays, were used for these analyses. Correlations among the TWE scores and TOEFL converted section and total scores were also calculated.

Statistical Comparison of Prompts. The performance of the prompts was compared using a two-factor between-subject analysis of variance (ANOVA) design with the two factors being TOPIC (I, II, III, and IV) and TYPE (implicit and explicit). The questions of interest were:

(1) Are TWE prompts equivalent across topics?

(2) Are implicit and explicit TWE prompts equivalent?

(3) Is there an interaction between topic and topic type?

For this study, two or more TWE prompts were considered equivalent if they produced similar distributions of scores.

Prior to completing the ANOVAs, the comparability of the examinees responding to the eight prompts was assessed by comparing their TOEFL Total converted scores. This was done to check the spiraling process previously discussed. The TOEFL total converted scores were compared for the subgroups taking each of the eight prompts using a one-way ANOVA. The retention of the null hypothesis of no differences would provide evidence to justify the use of the two-way ANOVA. Based on the results of the ANOVA, appropriate post hoc comparisons or follow-up analyses were conducted using Tukey's Wholly Significant Difference (WSD) test (Myers, 1979).

Gender Level Statistical Comparison of Prompts. Methods discussed above for comparing the performance of prompts for the total sample were also completed separately for males and females.[2]

[2]A three-factor ANOVA, which included gender as the third factor, was also performed. However, because a significant interaction between gender and topic made it necessary to conduct separate follow-up comparisons for males and females, it was decided to report the separate two-factor ANOVAs. The results of both analyses were consistent.

Results

Approximately 84,000 examinees took the Test of Written English along with the TOEFL in October 1989. Of that number, 79,879 examinees had complete data and were included in the analyses, i.e., off-topic and no-response papers were eliminated. One of the assumptions in the design of the study was that spiraling would produce randomly equivalent groups of examinees taking each prompt. To test that assumption, the TOEFL scores for these eight groups were compared. Table 1 presents the means and standard deviations of the TOEFL total converted scores for the groups responding to each of the eight prompts. The number of examinees responding to each of these prompts was large, ranging from 9,589 to 10,593.

TOEFL total converted score means ranged between 519.7 and 521.9 and the standard deviations ranged between 66.0 and 67.3. A one-way analysis of variance was performed and found no significant differences among the means ($F=1.36$; $df=7,79871$; $p=0.22$). Thus, it was reasonable to conclude that the spiraling was effective in producing eight groups of examinees equivalent in their English language proficiency as measured by the TOEFL.

Analysis of the Scoring Procedures

Table 2 presents a summary of the TWE reader analysis. Each TWE paper is read by two different readers and the correlations between the first and second readers provides an indication of the consistency of the process. These correlations ranged between .74 and .79. An estimate of reader reliability can be obtained by stepping up these correlations using the Spearman-Brown Prophecy formula. Reader reliabilities ranged between .85 and .88. Third readings are required when the first and second readers' scores differ by two or more points. The discrepancy rates for these prompts were small, averaging about 2 percent of the total number of papers.

Summary of the Reported TWE Scores

Table 3 presents a frequency distribution and summary statistics for the TWE scores for each of the eight prompts. Prompts A - D contain an implicit comparison statement and prompts E - H contain an explicit comparison. Among the **implicit** comparison prompts, the means ranged between 3.76 and 3.84 and the standard deviations between .88 and .96. Among **explicit** comparison prompts, the means ranged between 3.73 and 4 01 and the standard deviations between .88 and .97.

Figure 3 presents a graphical representation of the data in Table 3. It plots the relative frequency (percentage) of each of the possible eleven score levels (1.0 to 6.0). Although the shapes of the eight distributions are similar, there are differences in the relative frequencies at each score level. In order to illustrate more clearly the differences in the eight score distributions, Table 4 presents the percentage of scores within selected score intervals for each of the eight prompts. The intervals chosen were "at or above 5.0," "at or below 2.0," "at or above 4.0," and "at or below 3.5." The latter two categories were chosen in order to compare the distributions within

the region of acceptable/unacceptable performance used by some institutions. When the scores are grouped into these intervals, differences among the prompts are more readily apparent. Among the implicit comparison prompts, the differences were small, generally no greater than 3.5 percent. Among the explicit comparison prompts, however, the differences were larger, ranging between 2 and 16 percent.

Table 5 presents the correlations between the TWE scores and the TOEFL converted section and total scores. The correlations for each of the eight prompts with both the total and section scores were very similar. Total converted scores correlate the highest with TWE scores; however, this can be explained by the larger range and reliability of the total scores. The Structure and Written Expression score, logically the one most similar to the writing task, had, with one exception, lower correlations than either Listening Comprehension or Vocabulary and Reading Comprehension. These correlations are consistent with those found by Way (1990). He noted that correlations between the TWE and TOEFL sections were related to the composition of the language groups in the samples analyzed. For Asian language groups (i.e., Japanese, Chinese, and Korean), the TWE correlates higher with Listening Comprehension. Since the majority of the TOEFL testing population are usually from Asian language groups, overall correlations of the TWE with TOEFL sections rarely indicate that the TWE is more highly related to the Structure and Written Expression section.

Statistical Comparison of Prompts

Tables 6 and 7 present the results of the 4-by-2 analysis of variance performed on the TWE scores for the total group. Table 6 presents a detailed ANOVA table and Table 7 provides the results of the follow-up analyses on the means[3]. Because both the main effects and the interaction between topic and topic type were significant, comparisons between the two topic types within each level of topic and comparisons among the four topics within each level of topic type were made.

The differences in means between prompts using an implicit comparison versus an explicit comparison statement were significant for three out of the four topics. These differences ranged between .04 and .20. In all three cases, the means for the explicit prompt were higher than for the implicit prompt. Within a topic type, i.e., implicit or explicit, there were fewer significant differences among the means. Within the implicit type, only topic IV (prompt D) was significantly different from the others. For the explicit type, both topics III and IV (prompts G and H) were significantly different.

Figure 4 provides a simple graphical representation of the results of Table 7. It plots the mean TWE scores as a function of topic and topic type. The differential effect of type on these four topics is clearly depicted, especially the difference between the implicit and explicit means for topic III, the "Cities" chart-graph.

---

[3]In Tables 7, 13, and 15 prompts A, B, C and D are represented as Implicit topics I through IV. Likewise, prompts E, F, G and H are represented as Explicit topics I through IV.

Post Hoc Comparisons for Males and Females

Although the spiraling of the prompts was not purposefully designed to produce randomly equivalent subgroups of examinees, it was thought that it might have been successful in producing equivalent groups of males and females. If so, one could look at the performance of the prompts as a function of gender. Tables 8 and 9 present the means and standard deviations of the TOEFL total converted scores for male and female examinees. For males, TOEFL total converted score means ranged between 523.6 and 525.5, whereas for females, the means ranged between 513.8 and 517.3. For both males and females, a one-way analysis of variance did not detect any significant differences among the means (Males: $F=0.46$; $df=7,47138$; $p=0.86$; Females: $F=1.53$; $df=7,32122$; $p=0.15$).

Table 10 presents the mean and standard deviation of the TWE scores for males and Table 11 presents the mean and standard deviation of the TWE scores for females. The means for the males range between 3.69 and 3.99, and the means for the females range between 3.78 and 4.03. Tables 12 through 15 present the results of the two 4-by-2 analyses of variance performed separately on male and female TWE scores. Tables 12 and 14 present the detailed ANOVA table and Tables 13 and 15 provide the results of the follow-up analyses on the means.

For both males and females, there were significant main effects and interactions between topic and topic type. For both males and females, the differences in the means between prompts using an implicit comparison vs. an explicit comparison were significant for three out of the four topics (although not the same three for both males and females). For males, the differences ranged in size between .06 and .19, and for all three topics the explicit comparison version produced higher scores than the version using an implicit comparison. For females, however, while the differences in means were within the same range, .06 to .20, the explicit comparison statement did not consistently produce higher scores than the implicit comparison.

Within a topic type, for both males and females, there were fewer significant differences among the means. Among the four topics using an implicit comparison, topic IV (prompt D) was significantly different from the others for males and topics II and III (prompts B and C) were significantly different from each other for females. The magnitude of these differences was .10 for the males and .06 for the females. Among the four topics using an explicit comparison statement, there were a few more significant differences. For males, topics III and IV (prompts G and H) and for females, topics II and IV (prompts F and H) were significantly different from the others. The magnitude of these differences ranged from .12 to .30 for the males and .06 to .21 for the females.

Figures 5 and 6 present graphical representations of the results of Tables 13 and 15. In these figures, the mean TWE scores are plotted as a function of topic and topic type. These two figures clearly depict the differential effect that type had on these topics for males and females. Clearly the explicit chart-graph produced the highest means for both males and females.

Discussion

The present study was conducted to investigate the effect of variations in features of TWE prompts on the resulting performance of examinees. At the time the study was proposed, its design was influenced by several issues surrounding the TWE. Among them were concerns regarding the comparability of the chart-graph and compare-contrast prose prompts, concerns that variations in the way similar prompts were worded might influence the way examinees respond, and concerns about whether TWE scores would vary simply due to differences in the subject matter of the prompt. The two features that were ultimately varied in the study were topic and topic type. Within topic, both the prose and chart-graph format were used[4].

The results of the study indicated that there were differences in the way the eight prompts performed. Most of the differences, though statistically significant, were small in magnitude. The largest significant difference between means for the total group was .28 and the smallest was .04. In terms of "effect size"[5] (Cohen, 1988), the range of observed differences, from largest to smallest, translates to .30 and .04, with more than 80 percent of the differences equal to effects of .20 or less. For the separate analyses of males and females, a somewhat similar range of magnitudes were obtained.

In terms of sheer magnitude, the largest differences were among the means for prompts using the explicit comparison. Generally, the smallest differences were obtained among the means for prompts using the implicit comparison. In all comparisons--total, male, and female--the largest difference was between the explicit version of topic III (Cities, chart-graph) and the explicit version of topic IV (Leisure Time, prose).

Across all comparisons--total, male, and female--the chart-graph with the explicit comparison statement (prompt G) produced the highest mean scores. However, when worded with an implicit comparison, the chart-graph (prompt C) produced mean scores that were reasonably similar to the other implicit prompts. As mentioned earlier, based on the design of this study, it cannot be conclusively stated that the chart-graph format, as opposed to the "Cities" topic, is responsible for producing these results, but previous experience with this format in the operational TWE program would indicate that it is likely. As noted on page 4, the TWE Core Reader Group and Reading Management had been concerned about differences in examinee writing that had been informally observed in essays elicited by chart-graph prompts. As a result, the TWE program had already temporarily discontinued the use of the chart-graph format pending further research.

---

[4]As mentioned previously, the inability to develop two parallel sets of chart-graph prompts resulted in the confounding of format within topic.

[5]Cohen (1988) defines "effect size" as the difference in means expressed in standard deviation units. As a rule of thumb, he considers effect sizes of .20, .50 and .80 to be indicative of small, medium and large effects.

Although the differences among the mean scores were small, i.e., none greater than .30, tne differences in the distribution of scores within particular score intervals were more dramatic, especially as they affected an examinee's ability to meet an institution's admission criterion. Certain topics, when combined with an explicit compare instruction, produced larger percentages of papers in the 4.0 and above category. Thus, while group performance, as measured by the mean score, was somewhat similar, there was a tendency for individuals to receive higher scores when presented with an explicit comparison. An exception to this was noted for topic IV (prompts D and H) where there was a reversal in the effect of the explicit comparison. Thus, these results can be explained as an interaction of both topic and topic type.

The fact that differences of .04 were found to be statistically significant attests to the high degree of power of the statistical test used. However, this was to be expected since power is a function of sample size and each of the prompts was administered to approximately 10,000 examinees. In evaluating the results from this study, therefore, it is necessary to focus on the importance of these differences for the TOEFL program, rather than on the fact that they were "significant."

At its inception, one of the goals of this study was to inform people involved in developing TWE prompts about how variations in wording and format might affect examinee performance. But in reviewing the outcomes of the study, the authors were faced with the following questions: What are acceptable limits of differences in scores obtained across different prompts in a major testing program? While a majority of the differences observed in this study was very small, i.e., between .04 and .10, what about the differences in means of .20 or .30? What are acceptable limits of differences in the percentage of examinees at different score points? There are no simple answers to these questions. While many of the differences in means observed in this study were so small as to be of no practical significance, differences observed across prompts in the numbers of examinees at each score level were not.

Since this was probably the first study of its kind to focus on the comparability of prompts in a major testing program, we are currently unable to compare these results with data from other programs. However, one overriding implication for this program is that when TWE considers using different formats and topic types, their impact should be researched before they are implemented in the operational program.

All of the analyses conducted in this study assumed that there were no reader effects, i.e., that the scoring guide was applied in the same manner by all readers, or if there were any effects due to the leniency or stringency of the different readers, they were balanced out across the different prompts. Because the readers could not be randomly assigned to prompts, and because no reader scored more than one of the eight prompts, there is a possibility that beyond the format, topic, and topic type, the readers contributed a source of variation in the scoring process. However, since this potential source of variation was not incorporated into the design at the outset, the data that were collected were insufficient to tease out these effects on a post hoc basis (Raymond and Houston, 1990). Therefore, the authors caution readers of

this report to evaluate these findings in light of this fact and suggest that reader variation and its effect on the scoring process be given a high priority for further research.

Suggestions for the Future

To guide further investigation, the quantitative results obtained here can be used as a framework within which qualitative aspects of TWE essays can be explored. Rhetorical and linguistic analyses can now be conducted on the essays that were elicited by these prompts in an effort to identify (1) patterns of English structure and rhetoric that may be characteristic of different score levels and (2) patterns that appear to be determined by the format, topic, and topic type of each prompt.

If any such patterns are observed at various score levels across prompts, they can be used to evaluate and improve the TWE scoring guide. If distinctive patterns are observed within prompts, they can be used to inform people who develop TWE prompts about possible effects of format, topic, and topic type on examinee writing. This could result in greater efficiency in developing prompts and could contribute to a higher success rate for prompts at pretesting.

If patterns of similarities and differences attributable to the formats, topics, or topic types of prompts can be documented, such documentation can give the program a basis for beginning to consider the relative importance of small differences in the performance of various prompts. Being aware of the statistical characteristics generated by the prompts currently in use would facilitate the evaluation of the performance of new formats and topic types as they are introduced into the TWE.

New kinds of prompts could be pretested in conjunction with established kinds of prompts so that qualitative aspects of the resulting essays could be noted during pretest evaluation. Later, when a new kind of prompt is administered at final form, it might be introduced as the only prompt in a small administration. Its score distributions might then be compared with those of the spiraled populations represented in the present study and it would be possible to check for relative patterns of variation across prompts.

Finally, the issue of reader variation could be investigated to ascertain the impact this source of variation has on score reliability and validity. While it is difficult and expensive to conduct this type of study during an administration of a large operational testing program, it may be possible to design the scoring sessions of a small TWE administration in such a way that only a subgroup of the essays and readers are involved in the research. This type of investigation would provide evidence of the effectiveness of the reader training and scoring management procedures and might lead to ways to control for possible reader effects.

APPENDIX A

TWE Prompts

A

Some people learn best when a classroom lesson is presented in an entertaining, enjoyable way. Other people learn best when a lesson is presented in a serious, formal way. Which of these two ways of learning do you prefer? Give reasons to support your answer.

B

After they complete their university studies, some students live in their hometowns. Others live in different towns or cities. Which do you think is better--living in your hometown or living in a different town or city? Give reasons for your answer.

C

(Chart reproduced on following page)

The chart shows the results of a study that compared three major cities. Four characteristics of each city were measured on a scale from 1 to 10, with 10 being the most favorable evaluation. In which of these cities would <u>you</u> prefer to live? Use information from the chart to support your choice.

D

Some people think that parents should plan their children's leisure time carefully. Other people believe that children should decide for themselves how to spend their free time. Which idea do you agree with? Give reasons for your choice.

E

Some people learn best when a classroom lesson is presented in an entertaining, enjoyable way. Other people learn best when a lesson is presented in a serious, formal way. Briefly compare these two ways of learning. Which of these two ways of learning do you prefer? Why?

F

After they complete their university studies, some students live in their hometowns. Others live in different towns or cities. Briefly compare the advantages of living in your hometown with the advantages of living in a different town or city after completing your studies. Which do you think is better?

G

(Chart reproduced on following page)

The chart shows the results of a study that compared three major cities. Four characteristics of each city were measured on a scale from 1 to 10, with 10 being the most favorable evaluation. Compare the advantages of living in each of these three cities. In which of these three cities would <u>you</u> prefer to live? Use information from the chart to support your choice.

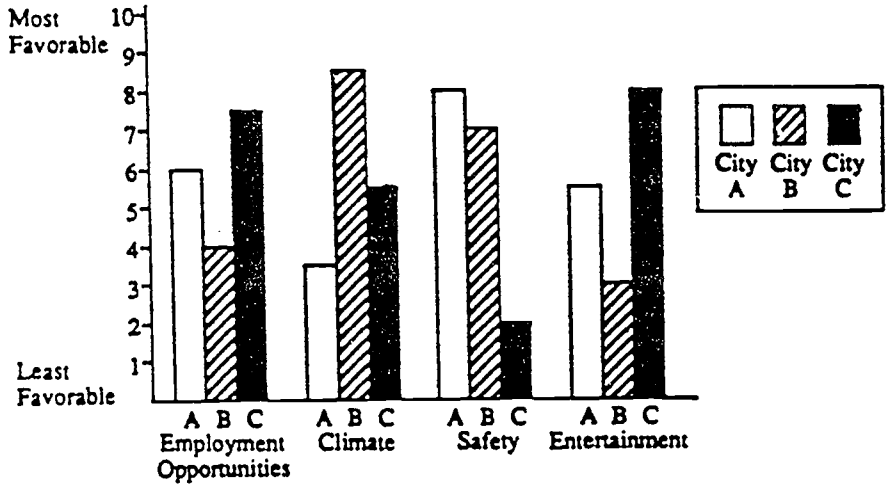
H

Some people think that parents should plan their children's leisure time carefully. Other people believe that children should decide for themselves how to spend their free time. Compare these two ideas. Which idea do you agree with? Give reasons for your choice.

TWE ESSAY QUESTION

(30 Minutes)

CHARACTERISTICS OF THREE CITIES



APPENDIX B

Test of Written English (TWE)
Scoring Guidelines

Readers will assign scores based on the following scoring guide. Though examinees are asked to write on a specific topic, parts of the topic may be treated by implication. Readers should focus on what the examinee does well.

Score

6 **Clearly demonstrates competence in writing on both the rhetorical and syntactic levels, though it may have occasional errors.**

A paper in this category
--is well organized and well developed
--effectively addresses the writing task
--uses appropriate details to support a thesis or illustrate ideas
--shows unity, coherence, and progression
--displays consistent facility in the use of language
--demonstrates syntactic variety and appropriate word choice

5 **Demonstrates competence in writing on both the rhetorical and syntactic levels, though it will have occasional errors.**

A paper in this category
--is generally well organized and well developed, though it may have fewer details than does a 6 paper
--may address some parts of the task more effectively than others
--shows unity, coherence, and progression
--demonstrates some syntactic variety and range of vocabulary
--displays facility in language, though it may have more errors than does a 6 paper

4 **Demonstrates minimal competence in writing on both the rhetorical and syntactic levels.**

A paper in this category
--is adequately organized
--addresses the writing topic adequately but may slight parts of the task
--uses some details to support a thesis or illustrate ideas
--demonstrates adequate but undistinguished or inconsistent facility with syntax and usage
--may contain some serious errors that occasionally obscure meaning

3 **Demonstrates some developing competence in writing, but it remains flawed on either the rhetorical or syntactic level, or both.**

A paper in this category may reveal one or more of the following weaknesses:
--inadequate organization or development
--failure to support or illustrate generalizations with appropriate or sufficient detail
--an accumulation of errors in sentence structure and/or usage
--a noticeably inappropriate choice of words or word forms

2 **Suggests incompetence in writing.**

A paper in this category is seriously flawed by one or more of the following weaknesses:
--failure to organize or develop
--little or no detail, or irrelevant specifics
--serious and frequent errors in usage or sentence structure
--serious problems with focus

1 **Demonstrates incompetence in writing.**

A paper in this category will contain serious and persistent writing errors, may be illogical or incoherent, or may reveal the writer's inability to comprehend the question. A paper that is severely underdeveloped also fall into this category.

Papers that reject the assignment or fail to address the question in any way must be given to the Table Leader. Papers that exhibit absolutely no response at all must be given to the Table Leader.

Figure 3

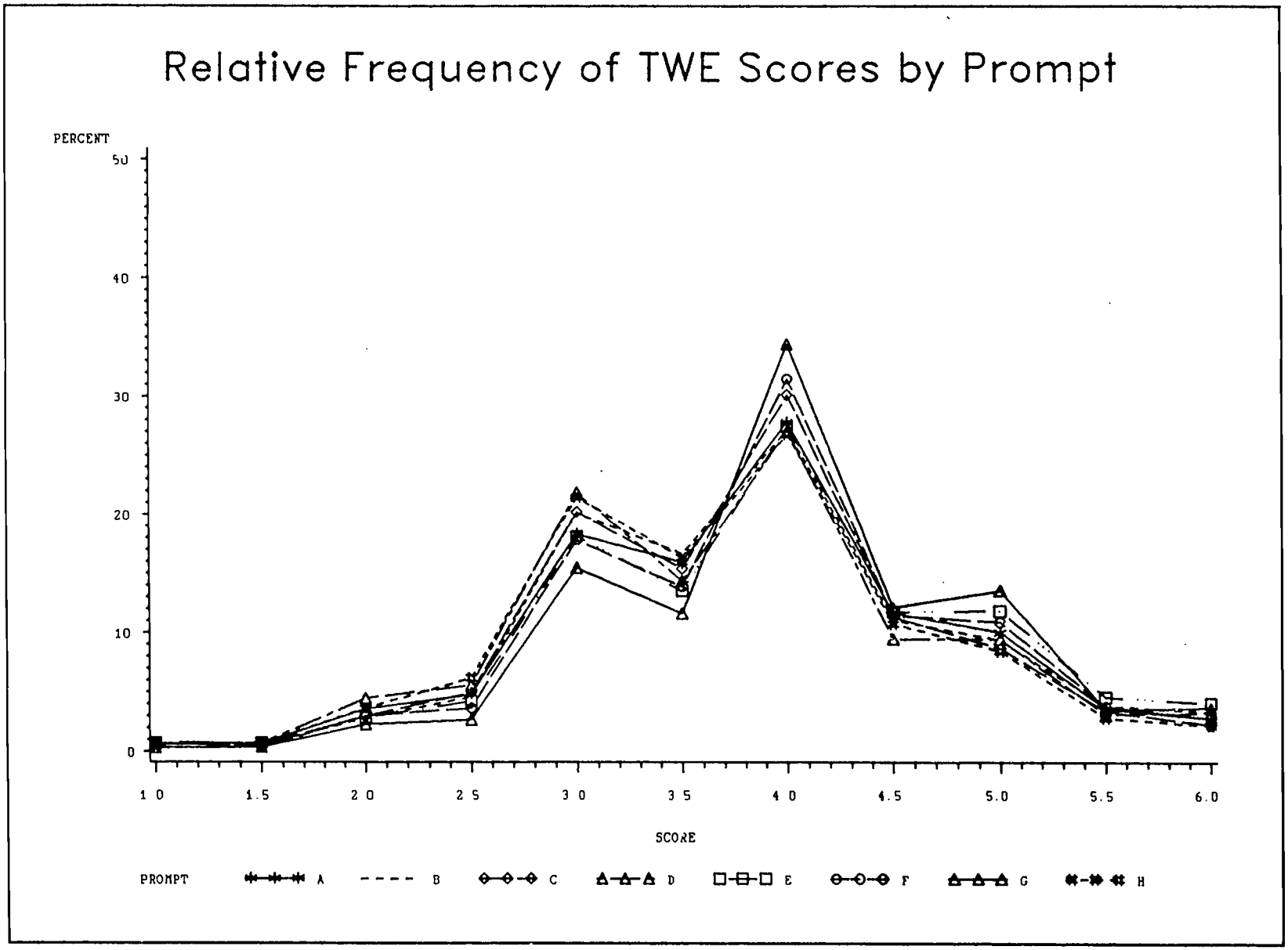

Relative Frequency of TWE Scores by Prompt
PERCENT
50
40
30
20
10
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
SCORE
PROMPT
A
B
C
D
E
F
G
H

Figure 4

Average TWE Score for the Total Group as a Function of Topic and Topic Type

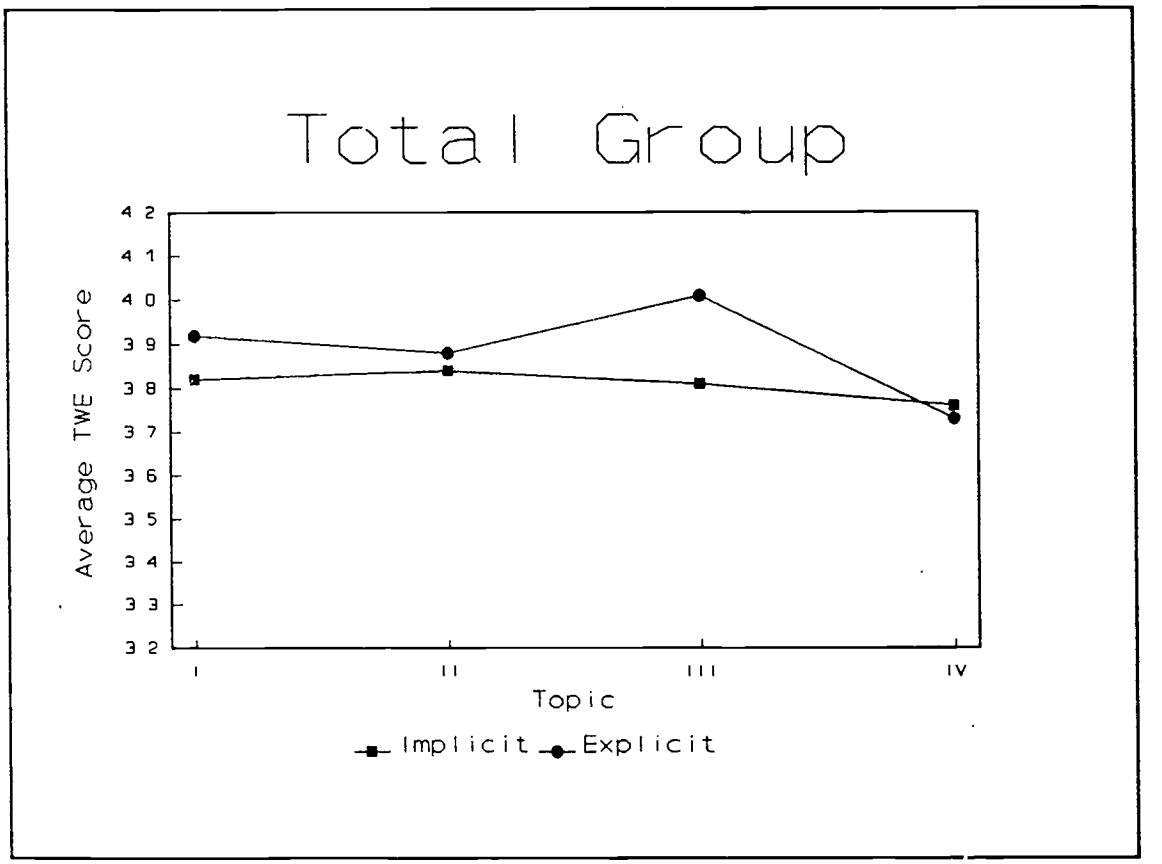


Figure 5

Average TWE Score for Males as a Function of Topic and Topic Type

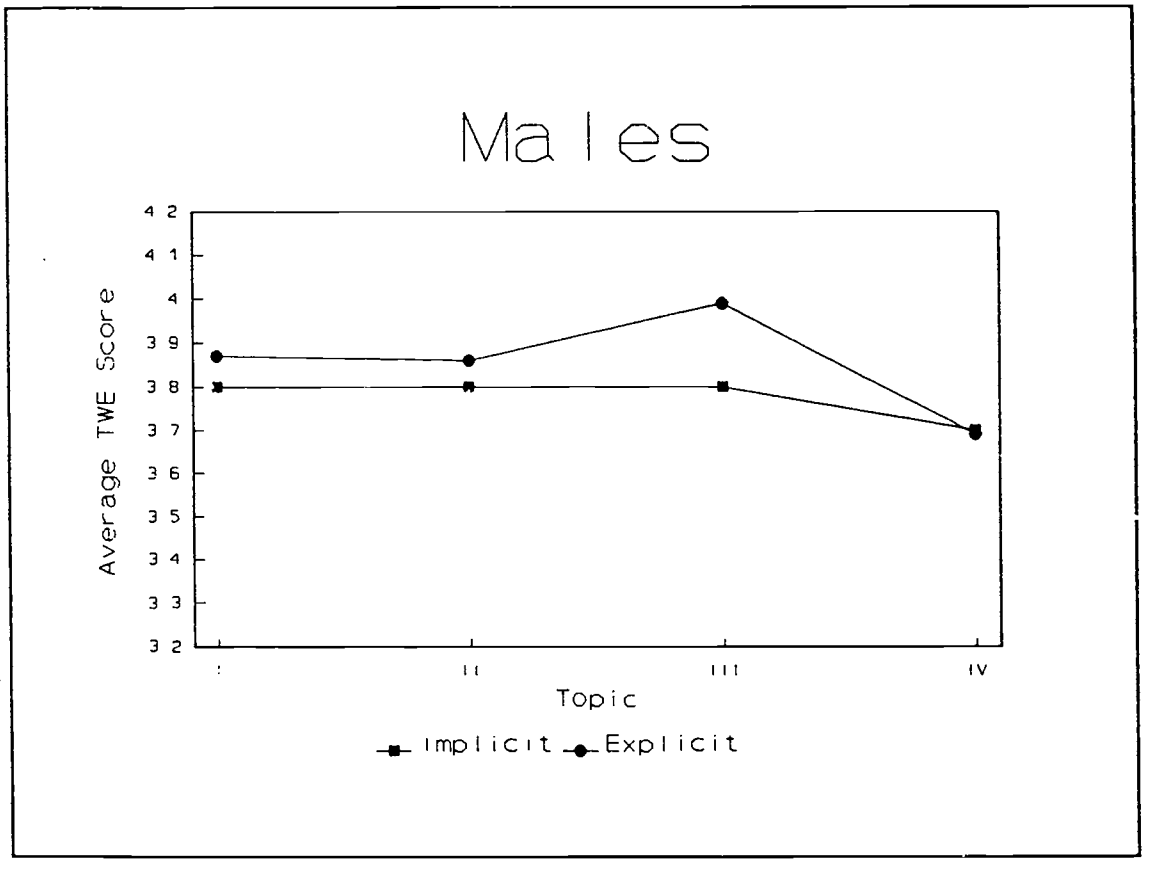


Figure 6

Average TWE Score for Females as a Function of Topic and Topic Type

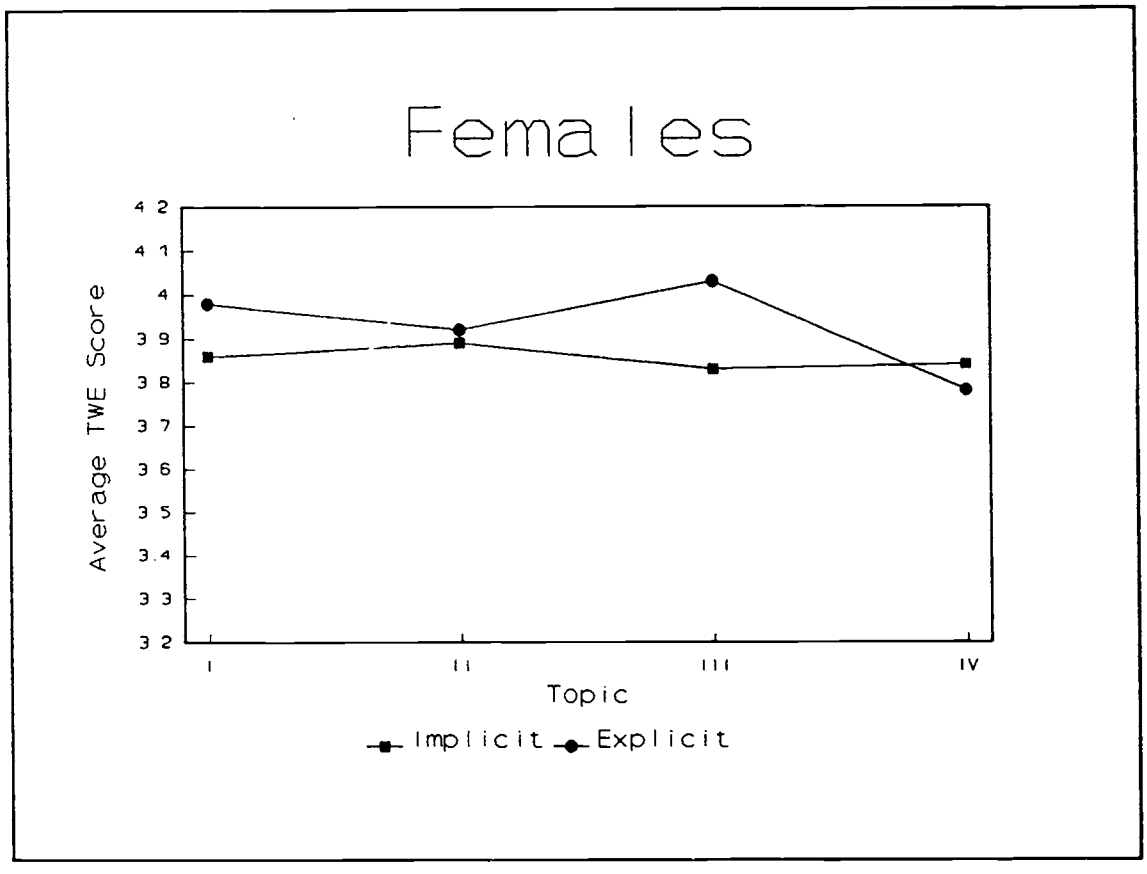


TABLE 1

Means and Standard Deviations of TOEFL Total Converted Scores for Examinees Responding to Each of the Eight Prompts

	Prompts							
	A	B	C	D	E	F	G	H
Sample Size	10593	9965	10000	9840	9956	9855	10081	9589
Mean	519.7	520.9	520.7	521.8	521.9	520.2	520.6	521.5
Std. Dev.	66.6	67.1	67.3	66.1	66.8	66.7	66.0	67.0

TABLE 2

Summary of TWE Reader Analysis

	Prompts							
	A	B	C	D	E	F	G	H
Correlation Between Readers	0.74	0.75	0.74	0.79	0.78	0.76	0.77	0.74
Reliability Estimate[a]	0.85	0.86	0.85	0.88	0.88	0.86	0.87	0.85
Discrepancy Rate	0.024	0.017	0.019	0.016	0.020	0.017	0.017	0.021

[a]Reader reliability is estimated by adjusting the correlation between readers with the Spearman-Brown Prophecy formula to reflect the use of two readers.

TABLE 3

Frequency Distribution and Summary Statistics of TWE Scores by Prompt

	Prompt															
	A		B		C		D		E		F		G		H	
Score	f	Percent	f	Percent	f	Percent	f	Percent	f	Percent	f	Percent	f	Percent	f	Percent
6.0	297	2.8	334	3.4	227	2.3	323	3.3	407	4.1	277	2.8	371	3.7	211	2.2
5.5	381	3.6	383	3.8	340	3.4	307	3.1	456	4.6	374	3.8	355	3.5	273	2.8
5.0	1063	10.0	934	9.4	869	8.7	932	9.5	1179	11.8	1074	10.9	1367	13.6	810	8.4
4.5	1234	11.6	1103	11.1	1131	11.3	932	9.5	1170	11.8	1127	11.4	1222	12.1	1031	10.8
4.0	2948	27.8	2712	27.2	3014	30.1	2665	27.1	2739	27.5	3102	31.5	3470	34.4	2568	26.8
3.5	1692	16.0	1661	16.7	1540	15.4	1415	14.4	1349	13.5	1368	13.9	1171	11.6	1565	16.3
3.0	1945	18.4	2008	20.2	2023	20.2	2150	21.8	1806	18.1	1758	17.8	1559	15.5	2059	21.5
2.5	507	4.8	457	4.6	492	4.9	546	5.5	421	4.2	358	3.6	268	2.7	592	6.2
2.0	380	3.6	280	2.8	299	3.0	441	4.5	284	2.9	296	3.0	230	2.3	349	3.6
1.5	75	0.7	54	0.5	36	0.4	71	0.7	73	0.7	56	0.6	37	0.4	69	0.7
1.0	71	0.7	39	0.4	29	0.3	58	0.6	72	0.7	65	0.7	31	0.3	62	0.6
Sample Size	10593		9965		10000		9840		9956		9855		10081		9589	
Mean	3.82		3.84		3.81		3.76		3.92		3.88		4.01		3.73	
Std. Dev.	0.93		0.92		0.88		0.96		0.97		0.91		0.88		0.91	
Skewness	-0.02		0.16		0.12		0.12		-0.03		-0.07		-0.05		0.07	
Kurtosis	0.16		0.09		0.13		0.02		0.03		0.29		0.25		0.17	

TABLE 4

Percentage of TWE Scores Within Selected Score Intervals

	Prompts							
	Implicit				Explicit			
	A	B	C	D	E	F	G	H
At or Above 5.0	16.4	16.6	14.4	15.9	20.5	17.5	20.8	13.5
At or Above 4.0	55.9	54.9	55.8	52.4	59.8	60.4	67.3	51.0
At or Below 3.5	44.1	45.1	44.2	47.6	40.2	39.6	32.7	49.0
At or Below 2.0	5.0	3.7	3.6	5.8	4.3	4.2	3.0	5.0

TABLE 5

Correlations Between TOEFL Converted Section and Total Scores and the TWE[a]

	Prompts							
	A	B	C	D	E	F	G	H
Listening Comprehension	0.64	0.64	0.62	0.63	0.65	0.64	0.62	0.62
Structure & Written Expression	0.62	0.61	0.61	0.60	0.62	0.62	0.61	0.63
Vocabulary & Reading Comprehension	0.64	0.63	0.63	0.61	0.65	0.63	0.63	0.64
Total Score	0.67	0.66	0.66	0.66	0.68	0.67	0.66	0.67

[a] Correlations of the TWE with each of the TOEFL scores have been corrected for the unreliability of the TOEFL scores by dividing the zero-order correlation by the square root of the TOEFL score reliability.

TABLE 6

Analysis of Variance Summary
Total Group Topic X Topic Type Results

SOURCE	DF	SS	F	PR > F
TOPIC	3	308.78	121.41	0.0001
TYPE	1	116.44	137.36	0.0001
TOPIC X TYPE	3	134.09	52.73	0.0001
ERROR	79871	67708.61		

TABLE 7

Follow-up Analysis of Significant Interaction
Total Group Topic X Topic Type Results

	TOPIC			
TOPIC TYPE	I	II	III	IV
Implicit	3.82[a]	3.84[a]	3.81[a]	[1]3.76
Explicit	3.92[b]	3.88[b]	4.01	[1]3.73

Letters to the right of the values indicate comparisons within a row (topic type) and numbers to the left of the values indicate comparisons within a column (topic). Values sharing the same letter or number are not significantly different at the .01 level.

TABLE 8

Means and Standard Deviations of TOEFL Total Converted Scores for Male Examinees Responding to Each of the Eight Prompts

	Prompts							
	A	B	C	D	E	F	G	H
Sample Size	6197	5840	5882	5882	5908	5785	6001	5651
Mean	523.6	524.8	524.9	524.8	525.1	524.6	524.3	525.5
Std. Dev.	66.4	66.9	66.9	66.4	66.7	66.7	66.0	66.3

TABLE 9

Means and Standard Deviations of TOEFL Total Converted Scores for Female Examinees Responding to Each of the Eight Prompts

	Prompts							
	A	B	C	D	E	F	G	H
Sample Size	4317	4042	4049	3894	3975	3999	4004	3850
Mean	514.1	515.1	514.3	517.0	517.3	513.8	515.0	515.6
Std. Dev.	66.4	67.2	67.5	65.4	66.7	66.3	65.9	67.7

TABLE 10

Means and Standard Deviations of TWE Scores (After Adjudication) for Male Examinees Responding to Each of the Eight Prompts

	Prompts							
	A	B	C	D	E	F	G	H
Sample Size	6197	5840	5882	5882	5908	5785	6001	5651
Mean	3.80	3.80	3.80	3.70	3.87	3.86	3.99	3.69
Std. Dev.	0.94	0.93	0.89	0.96	0.98	0.91	0.89	0.92

TABLE 11

Means and Standard Deviations of TWE Scores (After Adjudication) for Female Examinees Responding to Each of the Eight Prompts

	Prompts							
	A	B	C	D	E	F	G	H
Sample Size	4317	4042	4049	3894	3975	3999	4004	3850
Mean	3.86	3.89	3.83	3.84	3.98	3.92	4.03	3.78
Std. Dev.	0.92	0.90	0.86	0.95	0.95	0.91	0.88	0.89

TABLE 12

Analysis of Variance Summary
Male Topic X Topic Type Results

SOURCE	DF	SS	F	PR > F
TOPIC	3	245.56	94.88	0.0001
TYPE	1	74.93	86.85	0.0001
TOPIC X TYPE	3	63.42	24.50	0.0001
ERROR	47138	40667.00		

TABLE 13

Follow-up Analysis of Significant Interaction
Male Topic X Topic Type Results

	TOPIC			
TOPIC TYPE	I	II	III	IV
Implicit	3.80[a]	3.80[a]	3.80[a]	[1]3.70
Explicit	3.87[b]	3.86[b]	3.99	[1]3.69

Letters to the right of the values indicate comparisons within a row (topic type) and numbers to the left of the values indicate comparisons within a column (topic). Values sharing the same letter or number are not significantly different at the .01 level.

TABLE 14

Analysis of Variance Summary
Female Topic X Topic Type Results

SOURCE	DF	SS	F	PR > F
TOPIC	3	73.46	29.69	0.0001
TYPE	1	42.21	51.18	0.0001
TOPIC X TYPE	3	79.16	31.99	0.0001
ERROR	32122	26492.37		

TABLE 15

Follow-up Analysis of Significant Interaction
Female Topic X Topic Type Results

	TOPIC			
TOPIC TYPE	I	II	III	IV
Implicit	$3.86^{a,c}$	$^{1}3.89^{a}$	3.83^{c}	$3.84^{a,c}$
Explicit	3.98^{b}	$^{1}3.92$	4.03^{b}	3.78

Letters to the right of the values indicate comparisons within a row (topic type) and numbers to the left of the values indicate comparisons within a column (topic). Values sharing the same letter or number are not significantly different at the .01 level.

References

Bridgeman, B., & Carlson, S. (1983). Survey of academic writing tasks required of graduate and undergraduate foreign students (TOEFL Research Report No. 15). Princeton, NJ: Educational Testing Service.

Brossell, G. (1986). Current research and unanswered questions in writing assessment. In K. L. Greenberg, H. S. Wiener, & R. A. Donovan (Eds.), Writing assessment: Issues and strategies. New York: Longman.

Brossell, G. (1983). Rhetorical specification in essay examination topics. College English, 45, 165-174.

Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). Relationship of admission test scores to writing performance of native and nonnative speakers of English (TOEFL Research Report No. 19). Princeton, NJ: Educational Testing Service.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooper, P. L. (1984). The assessment of writing ability: A review of research (GRE Board Research Report No. 82-15R). Princeton, NJ: Educational Testing Service.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. E. Mosenthal, L. Tamor, & S. Walmsby (Eds.), Research on writing. New York: Longman.

Greenberg, K. L. (1986). The development and validation of the TOEFL writing test: A discussion of TOEFL Research Reports 15 and 19. TESOL Quarterly, 20, 531-544.

Hout, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. Review of Educational Research. 60, 237-263.

Legg, S. M. (1987, April). Understanding topic difficulty. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, DC.

Myers, J. L. (1979). Fundamentals of experimental design (3rd ed.). Boston: Allyn & Bacon, Inc.

Phillips, G. W. (1987, April). Essay topic equating: Statistical equating of direct writing assessment. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C.

Quellmalz, E. S., Capell, F., & Chou, C. P. (1982). Effects of discourse and response mode on the measurement of writing competence. Journal of Educational Measurement, 19, 241-258.

Stansfield, C. W., & Ross, J. (1988). A long-term research agenda for the Test of Written English. Language Testing, 5, 160-186.

Raymond, M. R., & Houston, W. M. (1990). Detecting and correcting for rater effects in performance assessment (ACT Research Report Series 90-14). Iowa City, IA: American College Testing.

Ruth, L., & Murphy, S. (1988). Designing writing tasks for the assessment of writing. Norwood, NJ: Ablex.

Way, W. D. (1990). TOEFL 2000 and section II: Relationships between structure, written expression, and the test of written English. Unpublished report, Educational Testing Service, Princeton, NJ.

57906-01202 • Y33M.5 • 275634 • Printed in U.S.A.