



TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 61
MARCH 1998

The Relationship Between Computer Familiarity and Performance on Computer-based TOEFL Test Tasks

Carol Taylor

Joan Jamieson

Daniel Eignor

Irwin Kirsch



The Relationship Between Computer Familiarity and Performance on Computer-based TOEFL Test Tasks

Carol Taylor, Joan Jamieson, Daniel Eignor, and Irwin Kirsch

Educational Testing Service
Princeton, New Jersey

RR 98-8



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

Copyright © 1998 by Educational Testing Service. All rights reserved.

No part of this report may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher. Violators will be prosecuted in accordance with both U.S. and international copyright laws.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logo, GRE, TOEFL, and the TOEFL logo are registered trademarks of Educational Testing Service. THE PRAXIS SERIES and the modernized ETS logo are trademarks of Educational Testing Service.

College Level Examination Program and CLEP are registered trademarks of the College Entrance Examination Board.

Pentium is a registered trademark of the Intel Corporation.

Authorware is a registered trademark of Macromedia, Inc.

Abstract

The increasing use of computer-based testing raises concerns about equity and bias. Specifically, many in the field of language testing are concerned that the introduction of a computer-based TOEFL test in 1998 will confound language proficiency with computer proficiency and thus bring construct-irrelevant variance to the measurement of examinees' English-language abilities.

In a Phase I study (Kirsch, Jamieson, Taylor, & Eignor, 1998), TOEFL examinees were surveyed regarding their computer familiarity and classified into one of three computer familiarity groups: low, moderate, and high. In this study, Phase II, more than 1,100 "low-computer-familiar" and "high-computer-familiar" examinees from 12 international sites were identified from the Phase I survey and administered a computer tutorial and a set of 60 computer-based TOEFL test items. The relationship between level of computer familiarity and performance on the computer-based items was then examined. The examinees in Phase II were largely representative of those in Phase I, who were representative of the general TOEFL test-taking population. Thus, results from this phase of the study are considered generalizable to the current TOEFL examinee population.

The effect of computer familiarity after adjustments for language ability was examined by performing a series of analyses of covariance (ANCOVAs), using TOEFL paper-and-pencil test score as the covariate. These analyses were followed by a series of ANCOVAs involving the computer familiarity variable and a number of other variables: gender, reason for taking the TOEFL test, times the TOEFL test had been taken, and location where the TOEFL test was taken. In a final set of analyses, the TOEFL paper-and-pencil test scores of the low- and high-computer-familiar examinees were weighted such that the groups had identical distributions on the covariate.

After controlling for language ability, the researchers found no meaningful relationship between level of computer familiarity and level of performance on computerized language tasks among TOEFL examinees who had completed the computer tutorial. This finding was consistent for all but one of the subgroups considered. A small but practically significant interaction between computer familiarity and reason for taking the test was found on the set of computerized reading items. Researchers concluded that there was no evidence of adverse effects on the computer-based TOEFL performance due to lack of prior computer experience.

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 (1996-97) members of the TOEFL Research Committee are:

Paul J. Angelis	Southern Illinois University at Carbondale
Micheline Chalhoub-Deville	University of Minnesota
JoAnn Crandall	University of Maryland Baltimore County
Fred Davidson	University of Illinois at Urbana-Champaign
Thom Hudson	University of Hawaii at Manoa
John A. Upshur (Chair)	Concordia University

To obtain more information about TOEFL products and services, use one of the following:

E-mail: toefl@ets.org

Web Site: <http://www.toefl.org>

Acknowledgments

An international study of this scale and complexity could not have been completed without the collaboration and assistance of many professionals from ETS, Sylvan Learning Systems, Inc., and the TOEFL overseas representatives. We wish to thank ETS staff in TOEFL Program Direction, Test Development, Research, and Operations Management whose support and financial commitments made this project possible. We also wish to acknowledge the TOEFL Research Committee for their encouragement and support throughout the study.

Some individuals, because of their special contributions, deserve mention. Charles Lewis, Donald Rock, Henry Braun, and Samuel Messick provided invaluable advice on the design of the study. Patricia Santiago and Terrie Mansmann provided guidance with operational issues. Test developers Felicia DeVincenzi, Philip Everson, Susan Nissan, Cynthia Potter, and Mary Schedl were responsible for developing the test items used in the study and reviewed the tutorials at various stages of development. The tutorial development team—Debra Pisacreta, Thomas Florek, Mary Lou Lennon, Louis Mang, Holly Knott, Janet Stumper, and Michael Ecker—worked on an accelerated schedule to develop and package the tutorial and test items. Thomas Florek advised on all equipment needs, oversaw the loading of the computer programs at Sylvan, made a site visit, and was on call to troubleshoot technical problems during data collection.

Sylvan Learning Systems, Inc. staff provided inestimable support with all logistics of the data collection. In retrospect we could not have managed the international computer administrations without Sylvan's partnership. Brett Lundeen negotiated contracts and arrangements with TOEFL overseas representatives. Kent Weatherley coordinated and oversaw all Sylvan activities. Sylvan's test administrators were extremely accommodating and stayed in close communication throughout the data collection. Special recognition and thanks to Isabelle Bazire, Mary Fortier, Bert Hendriksen, Mario Jimenez, Shaqeel Mubariz, Sylvia Neuteboom, Larry Perry, Susan Powell, Jay Lee, and Anand Ramachandran.

TOEFL representatives in Egypt, Japan, Korea, Mexico, Pakistan, Taiwan, and Thailand recruited local examinees and assisted with local arrangements and testing. Because of their diligent efforts, the response rates exceeded our expectations. While it is not possible to recognize all who contributed, we do want to acknowledge the local site coordinators: Vichitra Wongpiyanantakul, Injy Shaarawy, Meena Valli Mohammed, Laura Riveroll, Jai Ok Shim, Sharon Lin, and Satoka Setaka.

Michael Ecker devoted countless hours to setting up the data layouts and processing the data. Judy Pollack's data analyses were impeccable. Her patience, clarity in documenting procedures, scrutiny of the data, and accommodation of our schedules made the arduous task of report writing less difficult.

Charles Alderson, Lyle Bachman, Robert Boldt, JoAnne Crandall, William Grabe, Robert Kantor, Luis Laosa, Donald Rock, and Jacqueline Ross provided valuable feedback on an earlier version of this paper. Any remaining flaws are ours alone.

Finally, we wish to express gratitude to the TOEFL examinees who participated in this project. Our sincerest wish is that future test takers will be the direct beneficiaries of our efforts.

Table of Contents

	Page
Introduction	1
Computer Familiarity.....	1
Training.....	2
The Research Questions.....	4
Methodology	5
Instruments Used	5
Developing the Computer Familiarity Scale.....	5
Developing the TOEFL CBT Tutorial	5
Developing the TOEFL CBT Items	6
Procedures Followed.....	7
Selecting Equipment.....	7
Selecting Sites	7
Selecting Examinees.....	8
Matching Examinee and TOEFL Records	9
Training Test Administrators	10
Administering the Tutorial and CBT Items.....	10
Examinees Tested	11
Results	12
Comparability of the Sample	12
Differences in Familiarity Groups Prior to Adjustments for Ability	16
Effect of Computer Familiarity After Adjustments for Ability.....	19
Further Examination of Effects of Computer Familiarity	24
Summary and Conclusions	26
References	29

List of Tables

		Page
Table 1	Sites Selected by Testing Region.....	8
Table 2	Number of Examinees Tested	11
Table 3	Comparisons Between Phase I and Phase II Examinees.....	13
Table 4	Comparisons Between Phase I and Phase II Examinees on Computer Familiarity Scores by Computer Familiarity Group	15
Table 5	Comparisons Between Phase I and Phase II Examinees on TOEFL Paper-and-pencil Test Scores by Computer Familiarity Group.....	16
Table 6	Comparisons Between Computer-familiar and Computer-unfamiliar Examinees on TOEFL Paper-and-pencil Test Scores, CBT Scores, and Computer Familiarity Scores	17
Table 7	Correlations Among TOEFL Paper-and-pencil Test Scores, CBT Scores, and Computer Familiarity Scores	18
Table 8	Effect of Computer Familiarity on CBT Scores After Adjusting for Ability	19
Table 9	Effects of Computer Familiarity and Gender on CBT Scores After Adjusting for Ability	20
Table 10	Effects of Computer Familiarity and Reason for Taking the TOEFL Test on CBT Scores After Adjusting for Ability	21
Table 11	Actual and Adjusted Group Reading CBT Means for Cells Involved in Interaction of Familiarity Group Membership and Reason for Taking the TOEFL Test.....	22
Table 12	Effects of Computer Familiarity and Number of Times Tested on CBT Scores After Adjusting for Ability	23
Table 13	Effects of Computer Familiarity and Test Center Location on CBT After Adjusting for Ability.....	24
Table 14	Mean Weighted Difference on the TOEFL CBT Items Between Computer Familiar and Unfamiliar TOEFL Examinees for the Total Sample and by TOEFL Score Level	25

Introduction

Changing the mode of delivery of a standardized test such as the Test of English as a Foreign Language (TOEFL) from a paper-and-pencil format to a computerized format brings with it both “promises and threats” (Canale, 1986). Improvements planned for computerized TOEFL tests in the next decade, for example, include tailoring item administration to examinees’ ability levels, creating new item types which will allow constructed responses, enabling test takers to control the pace of the assessment (e.g., the speed with which the next listening item is presented), adding pictures and graphics to contextualize items, providing immediate feedback on machine-scored items, allowing flexible scheduling, and reporting scores faster. These improvements are intended to make the TOEFL test more meaningful to examinees, to English as a second/foreign language (ESL/EFL) teachers, to admissions officers, and to others who use the test scores.

Certain threats can also be foreseen in changing the mode of test delivery, however. Perhaps the greatest danger is that computerizing the TOEFL test could lessen the validity of the measure. If examinees are required to use a computer to take the test, their scores might reflect not only their level of English proficiency but also their level of computer proficiency. That is, two examinees with the same level of English ability might score differently on the computerized test: an examinee with a high degree of computer familiarity might score above another examinee with little computer familiarity. If this were to occur, then the test would no longer be measuring what it claimed it was. Score interpretations would be confounded with ability to use a computer. In other words, construct-irrelevant variance would be introduced into the measure (Messick, 1989).

The potentially confounding effect of computer familiarity on computerized test performance is a threat not unique to the TOEFL program. However, while many studies have investigated relationships between computer experience and variables such as age, gender, attitude, and anxiety (e.g., Loyd & Gressard, 1984; Marcoulides, 1988; Levin & Gordon, 1989; Kay, 1992; Miller & Varma, 1994), there is relatively little literature discussing the effects of computer familiarity on performance on computer-based language tasks. Further, there do not appear to have been any studies of this issue with an international student population.

In reviewing the literature, the authors found four studies that were particularly relevant. Two of these investigated whether computer familiarity affected test performance; the other two studies included computer training in their designs to reduce any negative effects that a lack of computer experience might have. Each of these investigations is discussed in turn.

Computer Familiarity

A study by Lee (1986) concluded that past computer experience significantly affected the scores of college students on a computerized test of arithmetic reasoning. In this study, the computerized arithmetic reasoning test score was the dependent variable, a paper-and-pencil arithmetic reasoning test score was the covariate, and computer experience was the independent variable. Computer experience was measured by a questionnaire, and scores on the questionnaire were used to divide students into three groups: “games,” low, and high. Lee found that the “games” group (defined as those students who had no computer experience or who had only played computer games) scored significantly lower, while there was no difference between

the low and high groups. Lee cautioned that computer testing may discriminate against those who have not worked with computers prior to testing, but also suggested that “minimal work with computers may be sufficient to prepare a person for computerized testing” (p. 732).

Two studies reported in an article by Mazzeo, Druesne, Raffeld, Checketts, and Muhlstein (1991) were primarily designed to investigate the comparability of scores from computerized and paper-and-pencil versions of two tests [College Level Examination Program® (CLEP®) English and Mathematics] but also included computer experience as a moderating variable. These studies used a counterbalanced design in which a random half of the group took the computerized version first, followed by the paper-and-pencil version; the other half took the paper-and-pencil version first. Examinees completed a computer familiarity questionnaire that was a modified version of Lee’s questionnaire (1986). However, whereas Lee (p. 729) found a skewed distribution, restricted variance, and low Cronbach’s alpha, and thus decided to use the questionnaire scores to form a categorical variable with three levels (games, low, high), Mazzeo et al. found a relatively normal distribution and summed the scores on the questions to form a continuous variable.

Computer familiarity was only analyzed for the mathematics test in Mazzeo et al.’s first study. No relationship was found between computer familiarity and scores on the paper-and-pencil section of the mathematics exam, but a small negative effect was found on the computer-administered section. In other words, examinees with greater computer familiarity had lower scores on the computer-based test items. Results of Mazzeo et al.’s second study indicated that for both sections of the mathematics test and for the English test, there was no evidence of a relationship between computer familiarity and performance on the computerized versions of the tests, after accounting for paper-and-pencil test scores.

In their discussion, Mazzeo et al. discounted the one significant finding of an effect of computer familiarity as anomalous because it was negative. Nonetheless, they wrote that because their sample size was small and their questionnaire was short and not field tested, the role of computer familiarity should not be dismissed.

To summarize these two studies on computer familiarity, Lee (1986) found that individuals with no computer experience (or only experience with game playing) received lower scores on a computerized test. A conflicting result was found by Mazzeo et al. (1991). In four out of five comparisons, computer experience was not related to performance on a computerized test. In the one case in which a significant relationship was found, it was the opposite of what was expected: higher computer familiarity was related to lower scores. Thus, although these studies addressed concerns similar to those of TOEFL test takers, one could not be certain from the researchers’ findings and interpretations that computer familiarity would have no effect on examinees’ performance on a computerized TOEFL test.

Training

In both of the studies discussed above, participants were given minimal computer training. In Lee’s (1986) study, students received instructions on how to enter their answers on the computer and were given three sample questions. In Mazzeo et al.’s (1991) report, the only training mentioned for the first study was in relation to the mathematics exam. Students were allowed to work through sample questions on the computer, but these items were not limited to the subject area tested. The computerized versions of the tests were revised before Mazzeo et

al.'s second study, and some of the changes involved training. Practice items were now limited to mathematics and English, and the set of practice items was expanded to include all item types on the test. Other changes concerned reviewing items and revising answers.

Two other studies investigated the role of computer familiarity from a different perspective. In these investigations, the researchers did not question whether computer familiarity would play a role; they assumed that it would. As a result, they explored whether or to what extent the effect of computer familiarity could be mitigated. In an early study, Johnson and White (1980) examined the effect of training on computerized test performance. Their participants were 20 senior citizens who were divided into two groups: one group was given training and experience with the computer, while the other received no training. The researchers reported that the group with training significantly outscored the other group, but this result must be cautiously interpreted due to the small number of participants and the fact that the participants do not represent the population of interest in the study.

Powers and O'Neill (1993) assessed the performance of college students on computerized reading and mathematics questions. Like Lee (1986), they used scores on paper-and-pencil versions of the items as covariates in the analyses. Independent variables included a measure of computer experience and scores on a computer attitude scale composed of measures of anxiety, confidence, and liking (viz., Loyd & Gressard, 1984). Experience was operationalized by response to the question: How often have you used a computer in the last five years (some time each week, some time each month, infrequently, never)? All students received on-line training on how to use a mouse, how to scroll, how to navigate through the test, and how to answer. Students were divided into two groups: one group could ask for any additional help needed (via an on-line help function), whereas the other group received only the on-line training.

Neither computer experience nor additional help accounted for a significant amount of variance in the computerized reading or mathematics scores after adjusting for paper-and-pencil test scores. Powers and O'Neill concluded that the familiarization procedures used in the study were sufficient to reduce to negligible the amount of variance introduced by the computer mode of administration. Although these results seem promising, several problems limit their applicability to the current study. Specifically, the Powers and O'Neill study defined familiarity in terms of only one question, included a fairly low number of examinees, and did not deal with the population of interest in this study.

In summary, this review of the literature on computer familiarity and experience revealed conflicting results. Lee (1986) found that prior computer experience significantly affected the performance of college students on a computerized mathematics test, while Mazzeo et al. (1991) found no evidence of a positive relationship between computer familiarity and performance on a computer-based test (CBT) after accounting for paper-and-pencil test scores. Powers and O'Neill (1993), who provided training components before administering a computerized mathematics and reading test, concluded that computer familiarization procedures reduced to an inconsequential level the amount of variance due to computer administration. These studies focused on the problem of interest in this study, but their findings were not considered generalizable to TOEFL's population of international examinees or to a language proficiency test.

The Research Questions

Educational Testing Service (ETS[®]) has announced that a computer-based TOEFL test will be made available in selected domestic and international test centers in 1998. In Phase I of this study, TOEFL examinees were surveyed about their computer familiarity in order to collect baseline data for the TOEFL program (Kirsch, Jamieson, Taylor, & Eignor, 1998). Sixteen percent of the sample of 89,620 TOEFL examinees, who were representative of the TOEFL population as a whole, were classified as having a low level of computer familiarity based on responses to a questionnaire. Although this percentage was much smaller than expected, when generalized to the TOEFL population, it amounts to more than 120,000 annual TOEFL test takers. Because low computer familiarity might affect test performance, the authors decided to follow the path taken by Johnson and White (1980) and Powers and O'Neill (1993) and design an on-line training package for TOEFL test takers. It was hoped that this training would minimize or eliminate any adverse effect due to prior level of computer familiarity.

To examine the impact of training in detail, it would have been desirable to create three groups of computer-familiar examinees and three separate groups of computer-unfamiliar examinees: one group who did not receive any on-line training, another group who received on-line training designed for native speakers of English (i.e., the basic ETS tutorial designed for the computer-based Graduate Record Examinations[®] and PRAXIS[®] tests), and a third group who received on-line training designed for non-native speakers of English. This design would have allowed an investigation of the effectiveness of different types of training. It was also vitally important to have an international sample of examinees. A design of this complexity would have required more examinees than were available, however, and the cost and time required would have been prohibitive. Consequently, while this study included an international sample, control groups were not included. Examinees with varying levels of computer familiarity worked through a single on-line training package or tutorial designed for non-native speakers of English.

The purpose of this study was to answer two related research questions. Of primary interest was the following question:

What is the relationship between level of computer familiarity and level of performance on a set of TOEFL CBT items, after administering an on-line training package or tutorial and after controlling for ability level using performance on the TOEFL paper-and-pencil test?

A secondary goal was to examine the effects of computer familiarity on performance by gender, reason for taking the TOEFL test, number of times the TOEFL test had been taken previously, and where the TOEFL test was taken. This research question was worded as follows:

What is the relationship between level of computer familiarity and performance on a set of TOEFL CBT items among selected subgroups of the population?

Methodology

To answer the research questions asked in the previous section, it was necessary to select a design that included performance on computer-based language tasks as the dependent variable, computer familiarity as the independent variable, and language ability as measured by TOEFL paper-and-pencil test scores as the covariate. Other independent variables included examinee characteristics such as gender, reason for taking the test, number of times tested, and where the test was taken. This section of the paper describes three aspects of the study: (1) instruments used, (2) procedures followed, and (3) examinees tested.

Instruments Used

Developing the Computer Familiarity Scale. Computer-familiar and computer-unfamiliar examinees were identified based on their responses to 11 items which were part of a 23-item questionnaire that focused on individuals' access to, attitude toward, and experience with computers as well as related technologies. The questionnaire was administered to all TOEFL examinees in April, 1996 and to examinees in China in May, 1996. The 11 items that were selected from the questionnaire all loaded heavily on the first factor of a two factor common factor solution with promax oblique rotation (see Harman, 1967). Based on a composite computer familiarity score created from the 11 items which ranged from 11 to 44, examinees were classified into one of three computer familiarity groups: low, moderate, and high. Two separate reports (Kirsch, Jamieson, Taylor, & Eignor, 1998; Eignor, Taylor, Kirsch, & Jamieson, 1998) provide detailed discussions of the procedures used to develop the questionnaire and the computer familiarity scale as well as the profile of TOEFL examinees with respect to their level of computer familiarity.

Developing the TOEFL CBT Tutorial. Prior to the current study, TOEFL program and test development staff created and tried out prototype item and response types planned for the TOEFL CBT using an existing ETS computer tutorial.¹ Program staff determined that a new tutorial would be needed to give TOEFL examinees sufficient computer training to ensure that their test scores would not be affected by lack of previous computer experience. In creating the new tutorial, a team of tutorial developers from ETS's research division consulted with TOEFL test development, program, and systems staff, and staff in these areas reviewed each section of the tutorial as it was developed. In the final stages of development, the tutorial was piloted with a small number of ETS staff and with international students from a local college-based intensive English program. The tutorial was designed to take 40 minutes, on average, to complete.

Although the new tutorial was based on the design of the standard ETS tutorial used in earlier research (Powers & O'Neill, 1993), it included several additional design features that were expected to help TOEFL examinees with varying levels of English-language skills and prior computer experience. These new design features included: (1) the use of simple language, (2) the use of graphics and animation to minimize reading requirements and simulate computer functions, (3) the provision for interaction through menus, (4) the provision for leveled practice

¹ ETS developed a set of basic tutorials for use with the computerized versions of the GRE and PRAXIS tests. These tutorials were composed of four sections: How to Use a Mouse, How to Select an Answer, How to Use the Testing Tools (e.g., the help and clock functions), and How to Scroll.

exercises where examinees who did not successfully complete an exercise would be routed to additional training and practice, and (5) the evaluation of examinee performance on the tutorial.

Another notable change in design concerned the part of the tutorial that informed examinees about how to respond to the new computer-based TOEFL items. Whereas the standard ETS tutorial presented a single “How to Answer” tutorial, the TOEFL tutorial presented three separate “How to Answer” tutorials. These corresponded to the three sections of the TOEFL test (listening comprehension, structure and written expression, and reading comprehension) and were presented prior to each test section. In other words, examinees were presented with the “How to Answer Listening” tutorial immediately before they were administered the listening comprehension items, etc. This modification was considered important because the computer-based TOEFL test would introduce a variety of new item stimuli and response types, and examinees would have an opportunity to practice items in conjunction with the corresponding test sections. A separate report provides a detailed description and evaluation of the tutorial (Jamieson, Taylor, Kirsch, & Eignor, 1998).

Developing the TOEFL CBT Items. TOEFL test developers prepared 20 computer-based test items for each of the three sections of the test (listening comprehension, structure and written expression, and reading comprehension) for a total of 60 items. The items were developed to have a broad range of difficulty and were to be scored right/wrong. The estimated internal consistency reliability for all 60 CBT items was computed to be .85 using coefficient alpha. This and other statistics reported in this section are based on the total sample of examinees used in the current study and are described later in this report.

The items for this study were representative of those planned for the actual computer-based TOEFL test, but new item types were overrepresented to some extent. While the structure and written expression items in the TOEFL computerized test will remain essentially the same as those in the paper-and-pencil test, the CBT listening and reading comprehension items will contain design features and response types that are unique to the computerized testing environment. Unlike the paper-and-pencil version, where all questions accompanying listening stimuli or a reading passage appeared on the same or adjoining pages, the computerized version displayed one question at a time on the screen. Another notable difference between the paper-and-pencil and computer-based versions was that in the CBT version examinees were required to answer each question in the order it was presented before proceeding to subsequent questions. In the paper-and-pencil version of the test, examinees had the option of perusing the set of questions associated with listening stimuli or a reading passage and then answering the questions in random order. Finally, the new response types will begin to move the test away from the exclusive use of the four-option multiple-choice, text-based response format. All of the new response types were represented in the 60 CBT items assembled for this study.

New design features in the 20-item listening comprehension section included the use of still photographs, diagrams, and pictures that provided context or information relevant to the audio stimuli. Another important change was that the volume of and progression through the test items were controlled by the examinee, rather than by a recording, as in the current large-group, paper-and-pencil testing situation. As a result of this change, examinees were able to determine for themselves the volume and pace at which they proceeded through the questions. The listening comprehension section also introduced academic discussions which contained longer stimuli and exchanges involving more than two speakers. Three new response types required examinees to: (1) click on a picture or letter where the letter may be placed on a diagram, chart, or picture,

(2) select two answers, and (3) match or order information presented in a lecture or academic discussion. The listening comprehension items ranged in difficulty from .30 to .96 with a mean difficulty of .66, using p-values (proportion getting the item correct) as the index of difficulty. The estimated reliability of these 20 items was .78 as measured by coefficient alpha.

Although no new item types were introduced in the 20-item structure and written expression section, the two existing item types were interspersed rather than administered in separate subsections as in the paper-and-pencil test version. The two item types required examinees to click on a correct word or phrase that best completed a sentence or that had to be changed in order for a sentence to be correct. Items for this section were selected from the existing pool of paper-and-pencil items. Their range of difficulty was .43 to .95 with a mean difficulty of .80, and the reliability of this section was estimated to be .77.

The 20-item reading comprehension section retained the passage length and general topic areas found in paper-and-pencil versions of the TOEFL test, but the way in which reading passages and comprehension questions were displayed was changed. In the computerized version, examinees were required to scroll through a text to read it. Although no passage extended beyond 350 words, the entire text was too long to be shown in one display. Location within the passage was noted at the top of the screen as "beginning," "more to follow," or "end." This 20-item section also contained two new item types that required examinees to work more directly with text. These included clicking on a word, phrase, or sentence within a passage, and inserting a sentence in a passage. The items in this section ranged in difficulty from .27 to .88 with a mean difficulty of .61; the estimated reliability was .80.

Procedures Followed

Selecting Equipment. The research team chose to use laptop computers for the study. This allowed testing teams to deal with the logistical challenges of setting up temporary testing sites and moving equipment internationally. It also presented a worst-case testing scenario where stimuli would be displayed on small laptop screens. This was considered important, since laptops will be used operationally in future CBT testing sites where low examinee volumes preclude the establishment of permanent CBT test centers. The equipment and software requirements were: 486 or Pentium® laptop, minimum of 8 megabytes of RAM, at least 100 megabyte hard drive, soundcard, 256 color monitor (480 by 640 resolution), headset, standard English keyboard, 3 1/2" high-density floppy drive, external mouse, mousepad, and Microsoft Windows 3.1. Authorware® (1995) was used to program both the tutorial and CBT items. A total of 50 laptop computers were used.

Selecting Sites. Site selection was based on a number of considerations, but the primary factors were overall geographic representation and proportional relation to annual TOEFL testing volumes. There were also several pragmatic considerations: (1) sufficient local testing volumes from which to recruit and test 100 participants per site (i.e., up to 500 potential participants within a 100-mile radius of a testing site), (2) the likelihood of finding both computer-familiar and computer-unfamiliar examinees at the site, (3) accessibility, and (4) the cost of local data collection. Study sites were pre-selected because of the logistical complexities of international computer-based data collection and because of time constraints (i.e., data collection had to be completed within three months).

Table 1 presents the proportion of annual TOEFL test volumes and sites selected by testing region. In general, the number of sites per geographic region was one site for each 10% of the annual TOEFL test volume. Exceptions reflected the need for geographic representation or an expectation that more than one site would be needed to obtain a sample of 100 participants. Thus, for example, although Latin America represented only 3% of the annual TOEFL examinee volume, one Latin American site, Mexico City, was included to fulfill geographic representation. Similarly, two European sites, France and Germany, were included because a review of the previous year's volumes suggested that two sites would be needed to obtain a sample of 100 participants. Also, permanent CBT test centers had already been established in both Frankfurt and Paris, making testing accommodations there easily accessible and data collection cost effective. In the end, 12 sites were selected for participation in the study.

Table 1
Sites Selected by Testing Region

Testing Region	Percent of Annual TOEFL Test Volume	Number of Sites Needed for Study	Sites Selected for Study
Asia	47.0	5	Bangkok, Thailand Karachi, Pakistan Seoul, Korea Taipei, Taiwan Tokyo, Japan
Near East/Africa	5.0	1	Cairo, Egypt
Europe	12.0	2	Frankfurt, Germany Paris, France
Pacific Islands	.5	0	
Latin America	3.0	1	Mexico City, Mexico
Canada	6.0	1	Toronto
United States	27.0	2	Oakland, California Washington, DC

Selecting Examinees. Study participants were selected based on their computer familiarity scores. For the current study, examinees in the high-computer-familiar group (i.e., those with total computer familiarity scores greater than 32.5 on the 11-44 scale) were identified as “computer familiar.” Examinees in the low-computer-familiar group (i.e., those with total computer familiarity scores of 11 to 22.5 on the 11-44 scale) were identified as “computer unfamiliar.” It was determined that a sample of 1,000 examinees (500 computer-familiar examinees and 500 computer-unfamiliar examinees) would provide sufficient statistical power for detecting both analysis of covariance (ANCOVA) main effects and interaction effects. To obtain a sample of at least 1,000 individuals worldwide, more than 5,000 examinees who met the computer familiarity criteria were sent letters of invitation. Invitations were generated randomly from lists of those who met the familiarity criteria.

An analysis of the Phase I computer familiarity questionnaire data revealed that the numbers of computer-familiar and computer-unfamiliar examinees varied widely across the 12 selected sites. Thus, each site was given a unique sampling plan so that when the samples were aggregated across sites at least 500 computer-familiar and 500 computer-unfamiliar examinees would be tested. This intentional undersampling of computer-familiar examinees and oversampling of computer-unfamiliar examinees was needed to ensure nearly equal size groups from which performance comparisons could be made. Because of the complexity of the study, there was concern about response rates, especially among computer-unfamiliar examinees, and about possible loss of data due to technical difficulties. Therefore, recruiters were advised to oversample by a specified number to ensure a minimum of 1,000 usable responses. Where there were insufficient numbers, all eligible examinees at a site were invited to participate.

Sylvan Learning Systems, Inc.² staff and TOEFL overseas test representatives assisted with the recruiting of examinees. At each site a local recruiting coordinator and team of assistants were selected. Three to eight weeks before the administration of the computerized tutorials and test items, each site coordinator received letters of invitation ready for mailing, written guidelines and protocols for recruiting examinees, scheduling and confirmation forms, and scheduling report forms. Each site also received candidate rosters of examinees classified as computer familiar and computer unfamiliar. As an incentive, sites were offered a monetary bonus if they recruited and tested their target numbers of examinees. They were offered an additional bonus if they completed the testing on schedule. Throughout the recruiting process, one of the authors of this paper monitored and advised on the recruiting efforts at each site.

Matching Examinee and TOEFL Records. While the sample was not stratified on other examinee background characteristics, it was seen as important to determine whether the sample included a good representation and distribution of background characteristics. Thus, examinees' TOEFL test registration numbers were used to link their CBT performance data to the TOEFL records containing their background information.

The following examinee background characteristics were identified for analysis in this study: (1) gender, (2) reason for taking the test, (3) number of times the test had been taken, and (4) location or site of test administration. Examinees were grouped into one of three self-reported reasons for taking the test: undergraduate admissions, graduate admissions, and other. For the variable representing the number of times the TOEFL test was taken, examinees were divided into two groups: those who reported that this was their first time to take a TOEFL test, and those who reported having taken two or more TOEFL tests prior to the current test administration. The location or site of test administration variable consisted of two groups: those who took the test at domestic sites (i.e., in the United States or Canada) and those who took it at foreign sites (i.e., all remaining locations).

Although there was keen interest in considering the effects of computer familiarity on performance by native language and native country, the unequal distribution of examinees across the sites precluded analyses using these subgroups. In Mexico City, for example, 50 examinees in the computer-familiar group and 6 in the computer-unfamiliar group were recruited, while in Taipei the numbers were 56 and 87, respectively. Thus, although looking solely at domestic and

² Sylvan Learning Systems, Inc. is the company that delivers computerized examinations worldwide for ETS.

foreign sites was in some ways inadequate, it was considered valuable to know whether examinees who tested in domestic and foreign sites could be considered comparable.

Training Test Administrators. While examinees were being recruited, test administrators from Sylvan Learning Systems, Inc. were trained by two of the authors of this paper. Test administrators were given a procedural manual which included detailed written directions for installing and testing the software and hardware configuration, checking that data were successfully copied, and returning data to ETS. Equipment and data collection checks were conducted on a daily basis. Administrators were also given detailed written directions for testing examinees; these included instructions on how to handle any interruptions in a candidate's testing session. Materials that accompanied the test administration procedures included an installation checklist for laptop configuration, consent and payment forms, certificates of appreciation, and irregularity report forms.

The authors oversaw the first installation of the tutorial and CBT item package on the laptops used in the study and observed testing in two sites, one domestic and the other foreign.

Administering the Tutorial and CBT Items. Certified test administrators administered the computerized tutorial and set of CBT items between June and August, 1996. Established Sylvan CBT centers were used in Paris, Frankfurt, Toronto, and Oakland. In collaboration with local TOEFL representatives and an ETS field service office, Sylvan teams set up temporary testing sites in the remaining eight locations (Bangkok, Cairo, Karachi, Mexico City, Seoul, Taipei, Tokyo, and Washington, DC). Both permanent and temporary sites provided individual testing stations and used the same laptops and hardware and software configurations. Standardized administration procedures were used in all sites, and Sylvan test administrators proctored each testing session.

At the beginning of every testing day, test administrators ran a short diagnostic program on each laptop to test the sound, color, and mouse movement. If there was an error that could not be easily corrected, administrators used a spare laptop. Test administrators, assisted by local TOEFL representatives who spoke the native language of the examinees, welcomed examinees, checked their identification, reviewed with them the purpose of the study, and took them to a computer station. Test administrators started the computers, then typed a password and each candidate's identification number twice. Examinees were then given up to 3 hours to complete the tutorial and 60 CBT items. In the event of an interruption of any kind, administrators followed specified procedures and completed an irregularity report that coded the type of interruption (e.g., computer equipment problem, electrical power failure, examinee misconduct) and the action taken. At the end of each candidate's session, the test administrators checked that the tutorial and CBT item data were copied to a Master Data Disk.

At the end of each testing week, test administrators copied all data to a Weekly Backup Disk and shipped the Master Data Disks to ETS with the examinees' signed consent and payment forms, ID forms, and any irregularity reports. Site coordinators used the appointment schedule forms to record any appointment cancellations, reschedulings, or no shows. These forms were later used with signed consent forms to verify the examinees that participated in the study. When testing was completed at a site, coordinators shipped the final set of Master Data Disks and all Backup Data Disks in separate envelopes to ETS. They returned all other study materials (i.e., appointment schedule forms, signed consent and payment forms, photo IDs, irregularity

reports) in a third envelope. As a final backup, data were stored on the hard drives of the laptops until data analysts processed the data at ETS.

Examinees Tested

Across the 12 sites, a total of 1,204 examinees participated in the study, 633 with a high level of computer familiarity and 571 with a low level of familiarity. Table 2 presents the numbers of examinees tested at each site. Of the examinees tested, 35 examinees (20 computer familiar and 15 computer unfamiliar) encountered technical difficulties that disrupted their sessions and the data collection. As a result, there were 1,169 usable data records, more than 500 each from the computer-familiar and computer-unfamiliar groups. Because a sufficiently large sample for the planned analyses had been obtained, because the technical interruptions seemed to occur randomly across the computer-familiar and computer-unfamiliar groups, and because the resolution of the 35 interrupted cases would have been extremely time-consuming and costly, the interrupted cases were excluded from further analyses.

Table 2
Number of Examinees Tested

Sites	Computer Familiar	Computer Unfamiliar	Total
Foreign			
Bangkok	50	92	142
Cairo	60	21	81
Frankfurt	47	6	53
Karachi	58	35	93
Mexico City	57	6	63
Paris	53	39	92
Seoul	48	80	128
Taipei	56	87	143
Tokyo	52	82	134
Domestic			
Oakland	44	40	84
Toronto	54	60	114
Washington, DC	54	23	77
Total Tested	633	571	1,204
Less Technical Interruptions	-20	-15	-35
Total Used	613	556	1,169

Results

This section of the paper focuses on three topics: (1) the degree of comparability of the TOEFL examinees who responded to the computer familiarity questionnaire in Phase I (see Kirsch et al., 1998) with those who received the computer tutorial and CBT items in Phase II; (2) differences between the computer-familiar and computer-unfamiliar groups in TOEFL paper-and-pencil test performance, CBT performance, and computer familiarity scores prior to adjustments for ability as measured by TOEFL paper-and-pencil test scores; and (3) the effects of computer familiarity after adjusting for language ability as measured by TOEFL paper-and-pencil test scores.

Comparability of the Sample

One important question is whether the examinees who participated in the current phase of this study included a good representation and distribution of TOEFL examinee characteristics compared to the approximately 60,000 examinees who were classified as high- and low-computer-familiar in the first phase. As reported in the previous section, examinees were selected for Phase II primarily to meet the computer familiarity criteria. Hence, computer-familiar examinees were undersampled and computer-unfamiliar examinees were oversampled by design in order to obtain the nearly equal numbers in the two groups that were needed for the planned analyses.

To examine the extent to which the sample included a representative distribution of other examinee characteristics, the Phase I and Phase II samples were compared with respect to gender, reason for taking the test (undergraduate admissions, graduate admissions, other), the number of times a TOEFL had been taken previously (once or more than once), where the test was taken (domestic or foreign site), English-language proficiency as reflected by TOEFL paper-and-pencil test scores, and computer familiarity scores based on the 11 items selected from the computer familiarity questionnaire.

Table 3
Comparisons Between Phase I and Phase II Examinees

Selected Background Characteristics	Phase I n = 59,275*				Phase II n = 1,169			
	Computer Familiar		Computer Unfamiliar		Computer Familiar		Computer Unfamiliar	
	n	%	n	%	n	%	n	%
Total	44,911	76%	14,364	24%	613	52%	556	48%
Gender								
Male	26,359	59%	5,624	39%	381	62%	230	41%
Female	17,864	40%	8,508	59%	227	37%	317	57%
Missing	688	1%	232	2%	5	1%	9	2%
Reason for Testing								
Undergraduate	15,112	34%	5,531	39%	196	32%	137	25%
Graduate	23,065	51%	5,886	41%	310	51%	285	51%
Other	6,734	15%	2,947	20%	107	17%	134	24%
Missing	---	---	---	---	---	---	---	---
Times Test Was Taken								
Once	23,462	52%	7,448	52%	326	53%	268	48%
More than once	21,449	48%	6,916	48%	287	47%	288	52%
Missing	---	---	---	---	---	---	---	---
Where Test Was Taken								
Domestic Site	13,621	30%	4,430	31%	142	23%	116	21%
Foreign Site	31,290	70%	9,934	69%	471	77%	440	79%
Missing	---	---	---	---	---	---	---	---

*Note: The n of 59,275 does not include the 30,471 examinees classified as moderately computer familiar.

With respect to many of the background characteristics studied, the sample of examinees who participated in Phase II was very similar to the groups who responded to the computer familiarity questionnaire in Phase I. For example, the percentages of male and female examinees in the computer-familiar and computer-unfamiliar groups were almost identical in Phases I and II. The comparisons for the number of times tested are also favorable: computer-familiar examinees in the Phase I and Phase II groups are very similar in terms of number of times the test was taken.

Among examinees who were unfamiliar with computers, however, there is a clear difference between the Phase I and II groups with respect to reason for taking the test. In Phase I, 39% reported taking the TOEFL test for undergraduate admissions compared to only 25% in Phase II. Conversely, 41% reported taking the TOEFL test for graduate admissions in Phase I compared to 51% in Phase II. One plausible explanation may be that an increased number of computer-unfamiliar examinees were tested in Korea, Taiwan, and Thailand in Phase II in order to obtain an adequate sample of computer-unfamiliar examinees across the 12 sites. In Phase I (Kirsch et al., 1998), 60% of the combined group of Korean, Thai, and Chinese examinees indicated that

they took the TOEFL test for graduate admissions, 31% for undergraduate admissions, and 9% for other reasons. Thus, by increasing the numbers of computer-unfamiliar examinees from these three sites, the overall proportion of Phase II examinees who indicated taking the TOEFL test for graduate admissions also increased.

There was also some difference between Phase I and Phase II examinees with respect to where they had taken the TOEFL paper-and-pencil test. In Phase II, both computer-familiar and computer-unfamiliar examinees were oversampled in the foreign sites. This was due in part to the difficulty of recruiting examinees domestically during the summer months and the desire to have strong international representation.

It is also important to consider the extent to which the Phase II examinees are representative of Phase I examinees with respect to computer familiarity. As shown in Table 4, t-tests revealed a statistical difference in computer familiarity scores between examinees in Phase I and those in Phase II who were classified as computer familiar. However, there were no practical differences between the groups (i.e., differences of 20% or more of the pooled standard deviation; see Cohen, 1988). In assessing significance, the authors placed greater emphasis on the practical difference measures than on the actual tests of statistical significance because fairly large sample sizes were used to perform many of the statistical tests in this study. With large sample sizes, differences can be statistically significant when the actual differences are a fairly small percentage of a standard deviation. This is certainly the case for the computer-familiar Phase I and Phase II groups in Table 4.

Table 4
Comparisons Between Phase I and Phase II Examinees on Computer Familiarity Scores by
Computer Familiarity Group

Group	Phase I: Computer Familiarity Score			Phase II: Computer Familiarity Score			Significance	
	n	Mean	SD	n	Mean	SD	Statistical	Practical
Computer familiar	45,141	37.60	3.15	613	37.88	3.20	t = -2.17*	9% of SD NO
Computer unfamiliar	14,481	17.41	3.62	556	17.67	3.49	t = -1.67	7% of SD NO

Note. A small number of cases were analyzed but not matched to other examinee records. These cases are included in this table, so there are slightly more Phase I examinees here than in other tables.

*p < .05.

In addition to looking at computer familiarity scores across both groups of computer-familiar and computer-unfamiliar examinees, it is interesting to compare the groups' English-language proficiency as measured by TOEFL paper-and-pencil test scores. Table 5 shows the means, standard deviations, statistical and practical significance tests for the TOEFL paper-and-pencil section and total scores for computer-familiar and computer-unfamiliar examinees in Phases I and II. Once again, t-tests were conducted between the means of the computer-unfamiliar and computer-familiar groups of examinees in the two phases.

While there were no statistical or practical differences in TOEFL section or total scores for examinees who were familiar with computers, statistically significant differences in reading comprehension and total scores were found for examinees with a low level of computer familiarity. The observed difference in means for reading, 1.04 points, approached practical significance, but it did not reach the 1.22 points (i.e., 20% of the pooled standard deviation) needed to be considered meaningful. A difference of at least 10.89 points in total mean scores would be considered meaningful, and here the observed difference in means was 5.83 points. Thus, although there were statistically significant differences in reading comprehension and total mean scores between Phase I and Phase II examinees classified as computer unfamiliar, the differences were not large enough to be considered meaningful in a practical sense.

Table 5
Comparisons Between Phase I and Phase II Examinees on TOEFL Paper-and-Pencil Test Scores by Computer Familiarity Group

Group/ TOEFL Section	Phase I: TOEFL Test Score			Phase II: TOEFL Test Score			Significance	
	n	Mean	SD	n	Mean	SD	Statistical	Practical
Computer familiar	44,911			613				
Listening		52.74	6.49		52.90	6.15	t = -.62	3% of SD NO
Structure		53.37	7.41		53.86	6.94	t = -1.60	7% of SD NO
Reading		54.74	6.40		54.94	6.03	t = -.77	3% of SD NO
Total		536.17	60.08		538.97	56.72	t = -1.15	5% of SD NO
Computer unfamiliar	14,364			556				
Listening		50.01	6.36		50.06	6.03	t = -.16	1% of SD NO
Structure		51.33	7.91		51.98	6.47	t = -1.92	10% of SD NO
Reading		51.81	7.21		52.85	6.09	t = -3.35*	17% of SD NO
Total		510.51	64.28		516.34	54.47	t = -2.11*	11% of SD NO

Note. In analyses for this and subsequent tables, $\alpha = .05$. A Bonferroni adjustment was not made here because the adjustment would have resulted in a family-wise alpha level of .006 and a decidedly more conservative test of significance. Given the context of this study, the more liberal criterion of $\alpha = .05$ was seen as the preferred standard. This is true for the other tables as well.

* $p < .05$

The data presented in Tables 3 through 5, then, suggest that no extreme differences with respect to important examinee background characteristics were introduced as part of Phase II sampling. The small differences that were introduced probably resulted from the deliberate selection of nearly equal numbers of computer-familiar and computer-unfamiliar examinees and the intentional oversampling of computer-unfamiliar examinees in Korea, Taiwan, and Thailand.

Differences in Familiarity Groups Prior to Adjustments for Ability

It is important to examine the relationship between level of performance on the TOEFL paper-and-pencil test, the CBT items, and the computer familiarity score prior to controlling for language ability. Table 6 shows that, with no adjustment for language ability, statistical and practical differences exist between the computer-familiar and computer-unfamiliar groups on all three measures (TOEFL test, CBT items, and computer familiarity scores). The difference in performance on the TOEFL paper-and-pencil test between the two groups of Phase II examinees is consistent with the data seen in Phase I (Kirsch et al., 1998), where examinees with high levels of computer familiarity also had significantly higher TOEFL test scores than those with low levels of familiarity.

There is also a large significant effect between the two groups of Phase II examinees and their level of computer familiarity. By design, there is a difference of 6 standard deviations on the familiarity scale between the computer-familiar and computer-unfamiliar groups. It is important to reiterate that the familiarity score was formed with the expectation that just this sort

of large difference would result. In other words, the two extremes of the computer familiarity classification were used to select participants for the present study.

It is interesting to note that the magnitudes of practical differences (i.e., the differences expressed as percentages of the pooled standard deviations) between the two computer familiarity groups on the TOEFL paper-and-pencil tests and the CBT items were quite similar. That is, for both groups, the difference in means for the TOEFL paper-and-pencil total score was 41% of a pooled standard deviation, and for the CBT total score, 40%. Comparable differences in means can be seen for the corresponding listening, structure, and reading section scores, where percentages of standard deviations ranged from 28 to 47% for the paper-and-pencil test sections, and from 31 to 42% for the CBT sections. To the extent that computer familiarity interacts with CBT performance, larger differences as expressed in standard deviation units might be expected between the computer-unfamiliar and computer-familiar groups.

Table 6
Comparisons Between Computer-familiar and Computer-unfamiliar Examinees on TOEFL Paper-and-pencil Test Scores, CBT Scores, and Computer Familiarity Scores

Instrument/ Section	Computer Familiar		Computer Unfamiliar		Significance	
	Mean	SD	Mean	SD	Statistical	Practical
Paper-and-pencil TOEFL						
Listening	52.90	6.2	50.06	6.0	t = 7.98*	47% of SD YES
Structure	53.86	6.9	51.98	6.5	t = 4.76*	28% of SD YES
Reading	54.94	6.0	52.85	6.1	t = 5.87*	34% of SD YES
Total	538.97	56.7	516.34	54.5	t = 6.94*	41% of SD YES
CBT Items						
Listening	13.55	4.0	11.93	3.7	t = 7.12*	42% of SD YES
Structure	16.18	3.1	15.20	3.5	t = 5.12*	30% of SD YES
Reading	12.45	4.2	11.16	3.8	t = 5.28*	31% of SD YES
Total	42.18	9.9	38.29	9.4	t = 6.77*	40% of SD YES
Familiarity	37.89	3.2	17.68	3.5	t = 103.4*	606% of SD YES

*p < .05.

The comparisons presented in Table 6 would be more meaningful if they were based on the performance of equally able groups of computer-unfamiliar and computer-familiar examinees. In this case, a suitable measure of ability would be performance on the TOEFL paper-and-pencil test. The preferred method of performing such an analysis would be to form matched groups of computer-familiar and computer-unfamiliar examinees, where matching is performed using TOEFL paper-and-pencil test scores. Such a design requires large sample sizes, however—much larger than feasible given the complexities of the present study. Hence, a decision was made to perform a series of analyses of covariance (ANCOVAs), with TOEFL paper-and-pencil test total score serving as the covariate. Such analyses essentially involve a comparison of “adjusted”

means for the two groups, where the adjustment takes into consideration performance on the covariate, in this case, TOEFL paper-and-pencil test total score.

Prior to discussing these ANCOVAs, it is useful to present the correlations between the covariate (TOEFL paper-and-pencil test total score) and the various CBT scores used as dependent variables in the ANCOVAs. These data, along with other within- and across-measure correlations, are presented in Table 7. As this table shows, the correlations of most meaning for the ANCOVAs are: TOEFL paper-and-pencil test total score with listening CBT score (.74), TOEFL paper-and-pencil test total score with structure CBT score (.74), TOEFL paper-and-pencil test total score with reading CBT score (.72), and finally, TOEFL paper-and-pencil test total score with total CBT score (.84). In all cases, the covariate shares a relatively high degree of relationship with the dependent measure.

Also of interest in Table 7 are the patterns of within-measure (paper-and-pencil test or CBT) and across-measure correlations. The pattern of CBT correlations is very similar to the pattern of TOEFL paper-and-pencil test correlations, providing some validity evidence for the newly constructed CBT measures. The across-measure correlations are also quite high, particularly given that some of the CBT scores used in the correlations are based on only 20 items. The CBT items seem to be behaving as expected; that is, scores on the paper-and-pencil and CBT sections of the same kind correlate more highly with each other than with scores on any of the other sections.

Table 7
Correlations Among TOEFL Paper-and-Pencil Test Scores, CBT Scores, and Computer Familiarity

Instrument/ Section	TOEFL Paper-and-Pencil Test				CBT Items			
	Listen	Struct	Read	Total	Listen	Struct	Read	Total
Paper-and-pencil Test								
Listening	1.00							
Structure	.65	1.00						
Reading	.64	.76	1.00					
Total	.86	.91	.90	1.00	.74	.74	.72	.84
CBT Items								
Listening					1.00			
Structure	.56	.73			.59	1.00		
Reading	.58	.64	.69		.67	.64	1.00	
Total	.71	.75	.78	.84	.88	.83	.89	1.00
Familiarity	.22	.13	.17	.19	.19	.15	.16	.19

Effect of Computer Familiarity After Adjustments for Ability

The primary issue of interest in this study was the relationship between level of computer familiarity and level of performance on the CBT items after administering the tutorials and controlling for language ability level using performance on the TOEFL paper-and-pencil test. Results from the ANCOVAs contained in Table 8 address this issue. In each of the first three comparisons shown in this table, the score on a 20-item set of listening, structure, or reading CBT items is the dependent variable, group membership (computer familiar or computer unfamiliar) is the independent variable, and TOEFL paper-and-pencil test total score is the covariate. The last comparison differs only in that the 60-item total CBT score is used as the dependent variable.

For each of the comparisons shown in Table 8, tests of statistical and practical significance are presented. A much different picture of performance emerges after adjusting for ability (i.e., TOEFL paper-and-pencil test score) than emerged from the data in Table 6. Whereas all of the differences in Table 6 were both statistically and practically significant, after adjusting for ability only two of the differences are statistically significant. Further, these two differences, on the 20-item sets of listening and structure CBT items, are clearly not practically significant. In other words, after taking ability into account, there is no meaningful relationship between the performance of the computer-unfamiliar and computer-familiar groups on the various sets of CBT items.

Table 8
Effect of Computer Familiarity on CBT Scores After Adjusting for Ability

Effect (Predictor)	Significance	
	Statistical	Practical
Listening CBT Familiarity	F = 4.41*	(SD = 4.0) .35 of a point or 9% of SD NO
Structure CBT Familiarity	F = 5.76*	(SD = 3.3) .32 of a point or 10% of SD NO
Reading CBT Familiarity	F = 3.80	(SD = 4.2) .28 of a point or 7% of SD NO
Total CBT Familiarity	F = 3.39	(SD = 10) .53 of a point or 5% of SD NO

*p < .05.

To determine the effects of computer familiarity on performance among selected subgroups, a series of ANCOVAs was performed using more independent variables than simply computer familiarity group membership. Table 9 presents the results of tests of the effects of both computer familiarity and gender on the various CBT scores; in all tests, TOEFL paper-and-pencil total test scores served as the covariate. Note that, in addition to tests for statistical and practical significance for the main effects (familiarity group membership and gender), this table presents tests for the interaction between these two variables. The results indicate that while there are

statistically significant differences in performance between males and females on the 20-item sets of listening and structure CBT items and on the 60-item total set after adjusting for ability, these results do not approach the level of practical significance used to assess differences in the study. Furthermore, it is noteworthy that the statistically significant differences are for gender and not familiarity or the interaction of familiarity with gender. It may therefore be concluded that there is no meaningful relationship between computer familiarity and gender on any of the 20-item CBT measures or on the 60-item CBT total.

Table 9
Effects of Computer Familiarity and Gender on CBT Scores After Adjusting for Ability

Effect (Predictor)	Statistical	Significance	
		Practical	
Listening CBT		(SD = 4.0)	
Familiarity	F = 1.90	.24 of a point or 6% of SD	NO
Gender	F = 8.70*	.50 of a point or 13% of SD	NO
Gender X Familiarity	F = .07	.04 of a point or 1% of SD	NO
Structure CBT		(SD = 3.3)	
Familiarity	F = 3.57	.26 of a point or 8% of SD	NO
Gender	F = 4.33*	.28 of a point or 8% of SD	NO
Gender X Familiarity	F = .05	.04 of a point or 1% or SD	NO
Reading CBT		(SD = 4.2)	
Familiarity	F = 2.07	.30 of a point or 7% of SD	NO
Gender	F = .94	.18 of a point or 4% of SD	NO
Gender X Familiarity	F = .24	.26 of a point or 6% or SD	NO
Total CBT		(SD = 10.0)	
Familiarity	F = 1.30	.38 of a point or 4% of SD	NO
Gender	F = 4.32*	.68 of a point or 7% of SD	NO
Gender X Familiarity	F = .24	.16 of a point or 2% or SD	NO

*p < .05.

The data presented in Table 10 are comparable to those presented in Table 9 except that, in this case, reason for taking the TOEFL paper-and-pencil test (for undergraduate or graduate admissions) is used as an independent variable along with familiarity group membership.

Table 10
Effects of Computer Familiarity and Reason for Taking the TOEFL Test on CBT Scores
After Adjusting for Ability

Effect (Predictor)	Significance		
	Statistical	Practical	
Listening CBT (SD = 4.0)			
Familiarity	F = 1.21	.22 of a point or 6% of SD	NO
Reason for Testing	F = 1.28	.22 of a point or 6% of SD	NO
Reason X Familiarity	F = .69	.16 of a point or 4% of SD	NO
Structure CBT (SD = 3.3)			
Familiarity	F = 7.90*	.44 of a point or 13% of SD	NO
Reason for Testing	F = 3.28	.28 of a point or 8% of SD	NO
Reason X Familiarity	F = .55	.12 of a point or 4% of SD	NO
Reading CBT (SD = 4.2)			
Familiarity	F = .26	.10 of a point or 2% of SD	NO
Reason for Testing	F = 4.08*	.42 of a point or 10% of SD	NO
Reason X Familiarity	F = 15.29*	.82 of a point or 20% of SD	YES
Total CBT (SD = 10.0)			
Familiarity	F = .52	.28 of a point or 3% of SD	NO
Reason for Testing	F = .03	.06 of a point or 1% of SD	NO
Reason X Familiarity	F = 9.06	1.14 of a point or 11% of SD	NO

*p < .05.

Table 10 reveals a number of statistically significant differences, two involving main effects and one involving an interaction. Only one difference reaches the level of practical significance considered in this study. This difference involves performance on the 20-item reading CBT and the interaction between familiarity group membership and reason for taking the TOEFL paper-and-pencil test (for undergraduate or for graduate admissions).

Table 11 contains both the unadjusted (or actual) and adjusted means for the four cells involving the interaction between familiarity and reason for taking the TOEFL paper-and-pencil test. The actual and adjusted means are based on the 20-item reading CBT. These data show that, after adjusting for TOEFL paper-and-pencil test scores, the scores of computer-unfamiliar undergraduates are significantly higher than those of any of the other three groups. In addition, analyses indicated that the adjusted means for the other groups are not statistically different from one another (these data are not contained in Table 11). To date, it has been difficult to generate an explanation for this practically significant interaction.

Table 11
Actual and Adjusted Group Reading CBT Means for Cells Involved in Interaction of
Computer Familiarity Group Membership and Reason for Taking the TOEFL Test

Group	Actual Means		Adjusted Means	
	Undergraduate	Graduate	Undergraduate	Graduate
Computer Familiar	11.94	12.75	11.67	12.08
Computer Unfamiliar	11.14	10.75	12.29	11.11

The data presented in Table 12 are comparable to those in Table 10 except that, in this case, number of times tested (on the TOEFL paper-and-pencil test) was used as an independent variable along with computer familiarity group membership. Recall that this variable was used to form two groups: those who reported that this was the first time they had taken a TOEFL test, and those who reported having taken two or more TOEFL tests prior to the current test administration.

Table 12 contains a number of statistically significant differences, many having to do with differences in CBT performance between the two groups formed with respect to the number of times tested variable. Yet only one difference approaches the level of practical significance considered in this study. After adjusting for ability using TOEFL paper-and-pencil test scores, there is a practically significant difference in performance on the 20-item reading CBT set between examinees reporting that this was the first time they had taken a TOEFL paper-and-pencil test and examinees reporting they had taken it more than once. Since this variable does not interact with familiarity, however, it will not be considered further. In sum, it may be concluded that there is no relationship between examinees' computer familiarity and the number of times they had taken the TOEFL test.

Table 12
Effects of Computer Familiarity and Number of Times Tested on CBT Scores After
Adjusting for Ability

Effect (Predictor)	Significance		
	Statistical	Practical	
Listening CBT (SD = 4.0)			
Familiarity	F = 4.37*	.34 of a point or 9% of SD	NO
Times Taken	F = 9.67*	.50 of a point or 13% of SD	NO
Times Taken X Familiarity	F = 10.30*	.52 of a point or 13% of SD	NO
Structure CBT (SD = 3.3)			
Familiarity	F = 5.71*	.32 of a point or 10% of SD	NO
Times Taken	F = .10	.04 of a point or 1% of SD	NO
Times Taken X Familiarity	F = .64	.10 of a point or 3% of SD	NO
Reading CBT (SD = 4.2)			
Familiarity	F = 2.13	.26 of a point or 6% of SD	NO
Times Taken	F = 22.28*	.82 of a point or 20% of SD	YES
Times Taken X Familiarity	F = .03	.02 of a point or 0% of SD	NO
Total CBT (SD = 10.0)			
Familiarity	F = 2.66	.52 of a point or 5% of SD	NO
Times Taken	F = 13.99*	1.18 of a point or 12% of SD	NO
Times Taken X Familiarity	F = 1.10	.34 of a point or 3% of SD	NO

*p < .05.

The data presented in Table 13 are comparable to those shown in Table 12 except that, in this case, location of the paper-and-pencil testing site (domestic or foreign) is used as an independent variable along with computer familiarity group membership.

A number of statistically significant differences are found in Table 13, but none of the results approach the level of practical significance used to assess differences in this study. Thus, it may be concluded that there is no meaningful relationship between computer familiarity and test site location.

Table 13
Effects of Computer Familiarity and Test Center Location on CBT Scores After Adjusting for Ability

Effect (Predictor)	Significance		
	Statistical	Practical	
Listening CBT			
		(SD = 4.0)	
Familiarity	F = .37	.12 of a point or	3% of SD NO
Location	F = 4.37*	.42 of a point or	11% of SD NO
Location X Familiarity	F = 4.41*	.42 of a point or	11% of SD NO
Structure CBT			
		(SD = 3.3)	
Familiarity	F = .28	.08 of a point or	2% of SD NO
Location	F = 12.82*	.56 of a point or	17% of SD NO
Location X Familiarity	F = 6.20*	.40 of a point or	12% of SD NO
Reading CBT			
		(SD = 4.2)	
Familiarity	F = .81	.20 of a point or	5% of SD NO
Location	F = .01	.02 of a point or	1% of SD NO
Location X Familiarity	F = .76	.18 of a point or	4% of SD NO
Total CBT			
		(SD = 10.0)	
Familiarity	F = .07	.10 of a point or	1% of SD NO
Location	F = .11	.12 of a point or	1% of SD NO
Location X Familiarity	F = 4.62*	.82 of a point or	8% of SD NO

*p < .05.

Further Examination of Effects of Computer Familiarity

The analysis of covariance procedure asks what the mean difference would be in the dependent variable if the means were the same on the covariate. As the analyses presented here indicate, there does not appear to be any meaningful relationship between computer familiarity and total number correct on the set of computer-based TOEFL test tasks. One potential problem with this set of analyses is that they may underadjust for the true difference between computer-familiar and computer-unfamiliar examinees for several reasons. One is that the analyses only took one covariate into account (i.e., TOEFL paper-and-pencil total test score). It is possible, for example, that high computer familiarity is associated with social and economic status. Yet the TOEFL program does not collect any information that would allow the researchers to estimate this effect or take it into account in any of the analyses. Another factor which may contribute to an underadjustment is the measurement error associated with the covariate. A third reason has to do with the relatively large mean difference (approximately 40% of a standard deviation) which exists between the scores of the computer familiar and unfamiliar examinees on the TOEFL paper-and-pencil test.

Another way to investigate the potential impact of computer familiarity on TOEFL CBT performance is to ask what would happen if the two groups had identical distributions on the

covariate. This question was addressed by weighting the two samples such that there were equal numbers of examinees in each group within 10-point score intervals on the TOEFL paper-and-pencil test. The total test score, rather than any of the three subtest scores, was used in this analysis because most institutions use the total test score for making their admissions decisions. This procedure parallels the differential item functioning standardization procedure as described in Dorans and Holland (1993).

Table 14 presents the summary results from an analysis of variance comparing the weighted mean differences on the total set of computer-based TOEFL test tasks between the computer familiar and unfamiliar examinees. This table shows that there is, on average, a 1.3 score point advantage for the familiar group, which is statistically but not practically significant. These analyses were also run by aggregating examinees into important score intervals – those below 500, those between 500 and 549, and those at or above 550. These results are also presented in Table 14. As shown here, the mean difference ranges between 1.4 points for examinees with TOEFL scores below 500, to 1.1 points for those with scores between 500 and 549. These differences are all less than 20 percent of a pooled standard deviation.

Table 14
Mean Weighted Difference on the TOEFL CBT Items Between Computer Familiar and Unfamiliar TOEFL Examinees for the Total Sample and by TOEFL Score Level

Effect (Predictor)	Weighted Mean (SD)		Mean Difference	Significance	
	Computer Familiar	Computer Unfamiliar		Statistical	Practical
Total Group	41.1 (9.0)	39.9 (9.8)	1.3 points	t = 2.24*	(SD=10.0) 13% of SD NO
Paper-and-pencil TOEFL Test					(SD=10.0)
Up to 499	37.1 (8.4)	35.7 (7.7)	1.4 points	t = 1.72	14% of SD NO
500 to 549	35.8 (8.0)	34.7 (8.4)	1.1 points	t = 1.32	13% of SD NO
550 plus	49.8 (5.7)	48.5 (6.3)	1.3 points	t = 2.21*	13% of SD NO

*p = <.05.

In sum, the data presented here and earlier in this paper can be seen as providing a lower and upper bound estimate of the size of the effect which might be expected from delivering a computer-based TOEFL test after administering a tutorial and only taking differences in the TOEFL paper-and-pencil total test scores into account. Further, while the data presented here provide an improved estimate of the adjusted difference between these two groups, it is still not a complete adjustment since other relevant covariates related to computer familiarity and ability could not be taken into consideration as they were not measured in this study.

Summary and Conclusions

In the previous phase of this two-phase study, a small but significant relationship was found between computer familiarity and TOEFL paper-and-pencil test scores. The major purpose of this phase of the study was to examine the relationship between level of computer familiarity and level of performance on a set of TOEFL CBT items, after administering a tutorial and controlling for ability using performance on the TOEFL paper-and-pencil test. A secondary purpose was to examine this relationship among selected subgroups (i.e., gender, reason for taking the test, number of times tested, and test site location). Examining the effects of computer familiarity on CBT performance was seen as important to address the concern that the planned introduction of the computer-based TOEFL test may confound the measurement of TOEFL examinees' English proficiency with computer familiarity.

Before addressing the research question, it was important to determine whether the sample included a good representation and distribution of TOEFL examinee characteristics. While it was hoped that the participants in this study would be representative of those in Phase I, examinees in the Phase II sample were selected based on their computer familiarity scores. Comparisons revealed that the two groups (i.e., Phase I and Phase II examinees) were extremely similar in terms of gender, number of times tested, level of computer familiarity, TOEFL test scores, and, for computer-familiar examinees, reason for taking the test. Therefore, the findings of this study are generally applicable to the current TOEFL test-taking population as a whole.

The data also revealed that, with no adjustment for language ability, statistical and practical differences in performance exist on the three measures used in the study (i.e., TOEFL paper-and-pencil test scores, CBT scores, and computer familiarity questionnaire scores). Examinees who were familiar with computers had significantly higher TOEFL test scores and CBT scores than those who were not. It may be that TOEFL examinees with high levels of computer familiarity in general have more opportunities for language and computer instruction and use.

The correlations between TOEFL test scores (the covariate) and CBT scores (the dependent variables) were also important. For both section and total scores, TOEFL paper-and-pencil test scores shared a relatively strong relationship with CBT scores. Moreover, the pattern of CBT within-measure correlations was quite similar to the pattern of TOEFL paper-and-pencil test within-measure correlations, providing some validity evidence for the CBT items. It appears that the CBT items used in the study provide a reasonable instrument from which to study the issues of computer familiarity raised in this paper and to generalize to the planned TOEFL CBT. If the CBT items had not performed as intended, any findings of effects would be questionable.

Having addressed the comparability of the samples and instruments, the first research question was addressed (i.e., What is the relationship between level of computer familiarity and level of performance on a set of TOEFL CBT items, after administering a tutorial and controlling for ability?). After administering the tutorial and adjusting for ability as measured by TOEFL paper-and-pencil test scores, there were no practical differences between computer-unfamiliar and computer-familiar examinees on CBT listening, structure, reading, or total scores. That is, all observed differences in scores between the two familiarity groups were 20% or less than the total group standard deviation. (In actuality, they were all 10% or less.) Thus, in response to the first research question, there does not appear to be a meaningful relationship between examinees'

level of computer familiarity and level of performance on the computerized language tasks after administering a CBT tutorial and controlling for ability level.

The second research goal was to examine this relationship among selected subgroups (i.e., gender, reason for taking the test, number of times tested, and test site), and the findings were similar across subgroups. After adjusting for ability and examining 16 possible main effects and 16 possible interactions, only one difference of practical significance was found. No meaningful differences were found between male and female examinees, between examinees tested at domestic and foreign sites, or between examinees taking the TOEFL test once or more than once. Only the reading CBT showed an interaction between familiarity and reason for taking the test, and here the differences just reached practical significance (i.e., 20% of a standard deviation).

To further examine the possible impact of computer familiarity on TOEFL CBT performance, the TOEFL paper-and-pencil test scores of the two groups of computer familiar and unfamiliar examinees were weighted such that the groups had identical distributions on the covariate. These analyses were repeated for groups of examinees within three score ranges – those below 500, those between 500 and 549, and those at or above 550. Statistically significant differences were found which resulted in advantages of between 1.1 and 1.4 score points for the computer-familiar examinees. However, none of these differences were practically significant; that is, the differences ranged from 11 to 14% of a standard deviation. Even though these differences are small, it seems reasonable to expect that the effect of computer familiarity will diminish over time as examinees gain experience with the tutorial and practice CBT test tasks using a computer and as the technology becomes more accessible around the world.

While the main questions of interest could be addressed in this study, there are several other questions that could not be answered. First is whether there are differences in effects based on native language and native country. The sample size would not support the analyses needed to answer this question directly, so the issue was explored more broadly using test site location (domestic or foreign). Also, given practical constraints that precluded the use of control groups, it is not known to what extent the specially designed tutorial was necessary to eliminate or minimize performance differences due to prior level of computer familiarity. As noted earlier, a separate report describes the tutorial in greater detail (Jamieson et al., 1998). In addition, while the study sample appears to be generalizable to the current test-taking population and while there is evidence within the TOEFL program that the testing population has remained fairly stable over the past several years, it is not known how characteristics of the TOEFL examinees who will be taking the operational computer-based TOEFL test in the future may change. At this point, forecasting possible changes in the TOEFL test-taking population, either as a result of the introduction of computer-based testing or other international social or political events, would be speculative. Finally, it is not known what other factors such as social and economic status may impact computer familiarity and performance.

In conclusion, what is known is that, after administering the CBT tutorial and controlling for language ability as measured by TOEFL paper-and-pencil test scores, there were no meaningful differences in performance between candidates with low and high levels of computer familiarity either for the TOEFL examinee population overall or for any of the subgroups considered in this study. The study found no evidence that lack of prior computer familiarity might have adverse effects on TOEFL CBT scores. Furthermore, it seems likely that examinee access to and experience with computers will continue to increase and thus, any small observed differences will likely diminish further over time.

This international study was conducted to address concerns about computer familiarity expressed by TOEFL program staff, members of the language teaching and testing community, and TOEFL committee members. Computer familiarity is only one of many important issues involved in moving the TOEFL test to computer, however. Continued research on equity issues will be needed as more experience with CBT procedures is gained and as test items that require greater computer manipulation are developed for future generations of computer-based tests. For example, at present, little is known about the effects of various multimedia or how what is being measured actually changes when presented in a CBT environment rather than a paper-and-pencil environment. Thus, the current series of studies (i.e., Kirsch et al. [1998], Eignor et al. [1998], Taylor et al. [1998], and Jamieson et al. [1998]) is viewed as the first in a series of research efforts needed as the TOEFL program undergoes the transition to computer-based testing.

References

- Authorware 3.0 [computer software]. (1995). San Francisco, CA: Macromedia, Inc.
- Canale, M. (1986). The promise and threat of computerized adaptive assessment of reading comprehension. In C. Stansfield (Ed.), *Technology and language testing* (pp. 29-44). Washington, DC: TESOL.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In H. Wainer & P. W. Holland (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Eignor, D., Taylor, C., Kirsch, I., & Jamieson, J. (1998). *Development of a scale for assessing the level of computer familiarity of TOEFL examinees*. (TOEFL Research Report No. 60). Princeton, NJ: Educational Testing Service.
- Harman, H. H. (1967). *Modern factor analysis*. Chicago: The University of Chicago Press.
- Jamieson, J., Taylor, C., Kirsch, I., & Eignor, D. (1998). *Design and evaluation of a computer-based TOEFL tutorial*. (TOEFL Research Report No. 62). Princeton, NJ: Educational Testing Service.
- Johnson, D., & White, C. (1980). Effects of training on computerized test performance in the elderly. *Journal of Applied Psychology*, *65*, 357-358.
- Kay, R. (1992). An analysis of methods used to examine gender differences in computer related behavior. *Journal of Educational Computing Research*, *8*, 277-290.
- Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees*. (TOEFL Research Report No. 59). Princeton, NJ: Educational Testing Service.
- Lee, J. (1986). The effects of past computer experience on computer aptitude test performance. *Educational and Psychological Measurement*, *46*, 727-733.
- Levin, T., & Gordon, C. (1989). Effect of gender and computer experience on attitudes toward computers. *Journal of Educational Computing Research*, *5*, 69-88.
- Loyd, B., & Gressard, C. (1984). The effects of sex, age, and computer experience on computer attitudes. *AEDS Journal*, *40*, 67-77.
- Marcoulides, G. (1988). The relationship between computer anxiety and computer achievement. *Journal of Educational Computing Research*, *4*, 151-158.

- Mazzeo, J., Druesne, B., Raffeld, P., Checketts, K., & Muhlstein, A. (1991). *Comparability of computer and paper-and-pencil scores for two CLEP general examinations*. (College Board Report No. 91-5). Princeton, NJ: Educational Testing Service.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)* (pp. 13-103). New York: Macmillan.
- Miller, F., & Varma, N. (1994). The effects of psychosocial factors on Indian children's attitudes toward computers. *Journal of Educational Computing Research, 10*, 223-238.
- Powers, D., & O'Neill, K. (1993). Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. *Educational Assessment, 1*, 153-173.



50% RECYCLED PAPER
10% Post Consumer Waste

Cover Printed on Recycled Paper

57906-15478 • U48M1.5 • 275757 • Printed in U.S.A.