DOCUMENT RESUME

ED 395 941 TM 025 008

AUTHOR	Secolsky, Charles
TITLE	Accounting for Random R' ponding at the End of the Test in Assessing Speededness on the Test of English as a Foreign Language. TOEFL Research Reports, Report 30.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-89-11
PUB DATE	Jan 89
NOTE	36p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Adults; *English (Second Language); *Language Tests; Pretests Posttests; Regression (Statistics); *Responses; *Scoring; *Test Items; Time Management
IDENTIFIERS	Randomization; *Speededness (Tests); *Test of English as a Foreign Language

ABSTRACT

The usual assessment of speededness for rights-only scored tests does not account for the possibility that examinees respond in a random or patterned fashion to the items at the end of the test as the time limit approaches. This study represented an attempt to determine if Sections 2 and 3 of the Test of English as a Foreign Language (TOEFL) are truly speeded according to established criteria. Two exploratory techniques employing regression analyses were used in an attempt to account for the possibility of random or patterned responses at the end of each section. One technique provided an estimate of the degree to which all examinees truly reached the 75% point on the sections, and the second provided an estimate of the degree to which all examinees truly completed the last set of items. Support for the results was obtained from an examination of the number of items not reached and the number to which examinees responded in a patterned fashion. Findings are limited to the extent that one can attribute items not reached and patterned responding to the effect of speededness. Four administrations of the TOEFL (two pretest and two other) for 9,160 examinees in all were studied. Results suggest that Section 3 might be slightly speeded for pretest administrations, but more study is needed to confirm this finding. (Contains 9 tables and 10 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 395 941

TOEFL®

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 30
JANUARY 1989

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

□ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Accounting for Random Responding at the End of the Test in Assessing Speededness on the Test of English as a Foreign Language

Charles Secolsky

ETS
EDUCATIONAL TESTING SERVICE

025008

ERIC

BEST COPY AVAILABLE

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

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

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

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

Current (1988-89) members of the TOEFL Research Committee include the following:

Patricia L. Carrell (Chair)	Southern Illinois University
Lily Wong Fillmore	University of California at Berkeley
Fred Genesee	McGill University
Russell G. Hamilton	Vanderbilt University
Frederick L. Jenks	Florida State University
Harold S. Madsen	Brigham Young University

Accounting for Random Responding at the End of the Test in Assessing Speededness on the Test of English as a Foreign Language

by

Charles Secolsky

Educational Testing Service
Princeton, New Jersey 08541

RR-89-11

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

Copyright © 1989 by Educational Testing Service. All rights reserved.

Unauthorized reproduction in whole or in part is prohibited.

TOEFL, the TOEFL logo, ETS, and the ETS logo are trademarks of Educational Testing Service, registered in the USA, and in many other countries.

ERIC
Full Text Provided by ERIC

Abstract

The usual assessment of speededness for rights-only scored tests such as the Test of English as a Foreign Language (TOEFL) does not account for the possibility that examinees respond in a random or patterned fashion to the items at the end of the test as the time limit approaches. Therefore, for TOEFL, speededness measures that are based only on the number of items not reached may underestimate the degree to which the test is actually speeded.

The present study represented an attempt to determine if Sections 2 and 3 of TOEFL are truly speeded according to established criteria. Two exploratory techniques employing regression analyses were used in an attempt to account for the possibility that examinees responded randomly or in a patterned fashion to unconsidered items at the end of each section. One technique provided an estimate of the degree to which all examinees truly reached the 75 percent point on the sections. The second technique provided an estimate of the degree to which examinees truly completed the last set of items. Support for the results was obtained from an examination of the number of items not reached and the number of items at the end of the sections to which examinees responded in a patterned fashion. The findings are limited to the extent that one can attribute items not reached and patterned responding to the effect of speededness.

Four administrations of TOEFL were studied: two non-pretest administrations and two longer pretest administrations. The results suggest that Section 3 for pretest administrations may be slightly speeded. It is recommended that more observational or survey methods be used to confirm this finding. If the finding is confirmed, it would be recommended that the TOEFL program investigate ways to increase the amount of time allotted per item.

Table of Contents

Page

Introduction 1

Method 3

Determining Whether Virtually All Examinees Reached the First 75 Percent of the Items 4

Determining Whether 80 Percent of Examinees Reached the Last Set of Items 6

Results 11

To What Extent Did Virtually All Examinees Reach the 75 Percent Point of the Sections? 11

To What Extent Did 80 Percent of Examinees Reach the Last Set of Items? 16

Discussion 20

Recommendation 23

References 25

List of Tables

Table Page

1 Numbers and Percentages of Items in the First, Middle, and Last Item Sets for Pretest and Non-Pretest Administrations 5

2 Deltas for Last Six Items of Section and Six Difficult Items from First and Middle Parts of Section for Sections 2 and 3 of TOEFL for Four Administrations 8

3 Numbers and Proportions of Examinees with Standardized Residuals Below z = -1.645 and Above z = +1.645 for Middle, Last, and Both Middle and Last Item Sets for Section 2 of TOEFL for Four Administrations 12

4 Numbers and Proportions of Examinees with Standardized Residuals Below z = -1.645 and Above z = +1.645 for Middle, Last, and Both Middle and Last Item Sets for Section 3 of TOEFL for Four Administrations 13

5 Percentage Completing Section, Percentage Completing 75 Percent of Section, Number of Items Reached by 80 Percent of Examinees, Gulliksen's Index of Speededness, and Stafford's Speededness Quotient for Section 2 and Section 3 14

6 Numbers and Proportions of Examinees Indicating the Same Response to the Last 25 Percent of Items for Sections 2 and 3 of TOEFL for Four Administrations 15

7 Numbers and Proportions of Examinees With Standardized Residuals and Pseudo Standardized Residuals Below z = -1.645 and Above z = +1.645 for Section 2 of TOEFL for Four Administrations 16

8 Numbers and Proportions of Examinees With Standardized Residuals and Pseudo Standardized Residuals Below z = -1.645 and Above z = +1.645 for Section 3 of TOEFL for Four Administrations 17

9 Numbers and Proportions of Examinees Indicating the Same Responses to the Last Four, Five, and Six Items for Sections 2 and 3 of TOEFL for Four Administrations 18

Acknowledgements

The author acknowledges the contributions made by several individuals to this project. Lawrence Stricker and Gary Schaeffer provided insightful comments at the proposal stage. Robert Altman, Hunter Breland, Gerald De Mauro, Marilyn Hicks, Philip Oltman, and Janice Scheuneman provided valuable suggestions for revising the report. Patricia Savage created the datasets and ran data analysis programs, and Nancy Parr provided editorial assistance. The author also thanks the TOEFL Research Committee for funding this project. The contents of this report are solely the responsibility of the author.

Introduction

One of the problems that confronts developers of rights-only scored tests is the determination of the appropriate amount of time to allot for the test or test section. When enough time is provided for most examinees to complete a test, the test is considered a "power" test. However, when a test is intended as a power test but too few examinees complete most of it, the test must also be considered speeded to some degree. A test is considered to be essentially unspeeded by Educational Testing Service (ETS) criteria if virtually all examinees reach 75 percent of the items and at least 80 percent of examinees reach the last item (Hecht & Swineford, 1981; Swineford, 1974). These speededness criteria are more appropriate for formula-scored tests, in which examinees are penalized for incorrect responses. For rights-only scored tests, such as the Test of English as a Foreign Language (TOEFL), examinees are encouraged to respond to items even though these items have not truly been qyqympted. Therefore, many responses at the end of the test may be random or patterned responses. For this reason, the methods and criteria currently used by ETS to assess whether a test is speeded likely underestimate the degree to which a rights-only scored test or test section is actually speeded.

The purpose of the present investigation was to assess speededness on Sections 2 and 3 of TOEFL using two exploratory methods that attempt to account for the possibility that examinees respond randomly or in a patterned fashion to the items at the end of these sections due to insufficient time. (Since Section 1, Listening Comprehension, is paced by a recording and all examinees are presented all the items, Section 1 cannot be considered speeded according to the same criteria.) The intention was not to develop a new index of speededness for TOEFL but to generate data that could be used to evaluate whether the time limits for Section 2 and Section 3 are appropriate.

Approaches to the assessment of speededness have involved both single and multiple test administrations. These approaches have been reviewed by Donlon (1980) and Rindler (1979). The approach of Cronbach and Warrington (1951)·is a multiple administration approach that utilizes the correlation between speed and power conditions for parallel forms of the same test. Cronbach and Warrington define the index of speededness, Tau, as

$$\text{Tau} = 1 - \frac{r_{A_t B_p} \cdot r_{A_p B_t}}{r_{A_t B_t} \cdot r_{A_p B_p}}$$

where the r's are correlations between scores for parallel forms of the same test (A and B), and t and p refer to time limit (speed) and power conditions, respectively. The index, a measure of the difference between time limit and power measures represented by parallel tests, approaches zero as the correlation between time limit and power true scores approaches 1.0.

As a multiple administration approach, Cronbach and Warrington's index is sensitive to the effects of time limits. However, because it is a multiple administration approach, it is administratively impractical for many testing programs. The Reilly-Donlon and biserial methods discussed by Donlon (1980) attempt to estimate the time limit and power condition correlation in a single test administration. Both these exploratory methods, which are not discussed here, are dependent on the assumption of normality in the rate at which examinees complete a test.

The single administration approaches of Stafford (1971) and Gulliksen (1950) are functions of the number of not-reached items. Stafford's speededness quotient is simply $NR_i/(W_i + O_i + NR_i)$, where NR_i = the number of not-reached items (excluding omits), W_i = the number of incorrect responses, and O_i = the number of omitted items (excluding not-reached items). Gulliksen's ratios involve the standard deviation of the number of items not answered correctly (s_x) (which includes not-reached, omitted and incorrect responses), the standard deviation of the number of items answered incorrectly (s_w), and the standard deviation of the number of items not reached (s_{nr}). As the value of the ratio s_w/s_x becomes very small, s_{nr} becomes large and the test is a speed test. On the other hand, as s_{nr}/s_x becomes very small, s_w becomes large and the test is a power test. These two ratios, s_w/s_x and s_{nr}/s_x define the extent to which a test is measuring speed and power, respectively.

Perhaps because of the difficulties inherent in interpreting Gulliksen's ratios when the values of the ratios are high, ETS adopted a simple set of guidelines for determining whether a test is a power test (Rindler, 1979). The criteria for a power test are that virtually all examinees reach the first 75 percent of the items and at least 80 percent of the examinees complete the test.

The problem with the above measures of speededness for rights-only scored tests is that they are not sensitive enough to the possibility that some portion of the examinee group did not have enough time to truly attempt the items near the end of the test. In reality, some nonnegligible portion of the examinee group may have responded with random or patterned responses to the items at the end of the test or test section as the time limit approached. For this reason, Rindler (1979) has criticized the single administration approaches such as those of Gulliksen and Stafford and the ETS criteria.

Bejar (1985) has developed indices for detecting speededness on TOEFL that are sensitive to the random responses at the end of the test which may be due to a lack of time. According to Bejar, a test is speeded if performance on the most difficult items is not solely a function of ability. One index he proposes compares the observed performance on the most difficult items of the test to performance predicted by the item response theory model for these items. Essentially, for many of the difficult items, if the observed proportion of examinees responding correctly exceeds the proportion predicted by the model, a section is considered by Bejar to be speeded.

The theory underlying Bejar's procedure was that on the difficult items, lower ability examinees would perform better than predicted due to random or patterned responding. However, by basing his index of item fit on all fifteen examinee ability intervals of the IRT theta scale and on cases where predicted performance was greater than observed perfc~mance, Bejar may have incorporated sources of error into the index that might not be attributable to speededness. Secolsky (1985) adapted Bejar's item level index and computed it only on the lowest seven out of fifteen examinee ability intervals and only on cases where observed performance was greater than predicted performance.

In the present study, two different applications of regression analysis were employed in an attempt to determine if each of the two ETS criteria for a speeded test has been met when taking random or patterned responding into account. In addition, the indices suggested by Gulliksen (1950) and Stafford (1971) were computed and the ETS criteria evaluated without taking into account random or patterned responding. These latter measures are presented in order to portray the extent to which Sections 2 and 3 of TOEFL are to be considered speeded according to the usual assessment of speededness. The proportion of examinees responding to the last set of items with the same response was also determined in an effort to evaluate the ETS speededness criteria.

In the remainder of the report, the discussion is o ganized around the questions posed by the two ETS criteria for an unspeeded test: (1) Have virtually all of the examinees truly reached the first 75 percent of the items? (2) Have 80 percent of the examinee group truly reached all of the items?

Method

The techniques developed for the study were applied to four administrations of TOEFL. Two administrations were pretest administrations (Administration I and Administration II), and two administrations were non-pretest administrations (Administration III and Administration IV). In pretest administrations, pretest items are interspersed in the set of operational items for Sections 2 and 3. Section 1 (Listening Comprehension) is paced by a recording, and therefore the rate of responding is held constant. For pretest administrations, examinees are given 35 minutes to

BEST COPY AVAILABLE

answer 60 items for Section 2 (Structure and Written Expression) and 65 minutes to answer 90 items for Section 3 (Reading Comprehension and Vocabulary). For the non-pretest administrations, examinees are given 25 minutes to answer 40 items for Section 2 and 45 minutes to answer 60 items for Section 3.

For the pretest administrations, the examinees selected for the study were those who were administered test forms with the items in the same order. (Items appear in different orders in other, scrambled forms.) For Administration I, the results are based on the responses of 1,624 examinees; for Administration II, results are based on the responses of 1,042 examinees. Likewise, for the two non-pretest administrations, the examinees selected for the study had taken the same test format. For Administration III, results are based on the responses of 2,766 examinees, while for Administration IV, results are based on the responses of 3,728 examinees. Since the test formats are spiraled (rotated) in operational administrations, the four groups of examinees can be considered spaced samples of the total group of test takers for the respective administrations.

Foreign examinees are not administered pretest items. Therefore, pretest administration analyses are based only on examinees tested at domestic centers. For the non-pretest administrations, however, analyses are based on both domestic and foreign examinees.

Determining Whether Virtually All Examinees Reached the First 75 Percent of the Items

To answer the first question (Did virtually all examinees complete the first 75 percent of the items?), regression analyses were employed. First, simple linear regression was performed using the scores on the last 25 percent of the items in each section as the criterion. A second regression analysis was performed using as the criterion scores on approximately the 15 percent of items immediately preceding the last 25 percent of the items. In both cases, the predictor variable was the score on the set of items representing approximately the first 60 percent of the items in each section. The raw score for the predictor is very unlikely to be affected by speededness since the first 60 percent of the items are not located near the end of the section. By performing separate linear regressions, predicted scores were obtained for both the last 25 percent of items and the immediately preceding 15 percent of items.

For Section 2, the raw score for the first 60 percent of items (X) (the predictor) was based on responses to two item types (Structure and Written Expression). Support for including the item scores from two different item types in the predictor stems from the fact that scores for these two types have been typically highly correlated (about r = .90) (see, for example, Hicks, Secolsky & Skelton, 1987). Therefore, including items from a different item type in the predictor does not seem to present a serious problem. As with Section 2, the first set of items of Section 3 was based on two different item types (Vocabulary and Reading Comprehension). The corrected correlation typically found between Vocabulary and Reading Comprehension is about .90 (see, for example, Hicks, Secolsky & Skelton, 1987). The numbers and percentages of items in the first, middle, and last sets of items are presented in Table 1 for pretest and non-pretest administrations.

Table 1
Numbers and Percentages of Items in the First, Middle, and Last Item Sets for Pretest and Non-Pretest Administrations

Pretest Administrations (Admin. I and Admin. II)

	Section 2		Section 3	
	No. of Items	%	No. of Items	%
First	34	56.7	56	62.2
Middle	11	18.3	11	12.2
Last	15	25.0	23	25.6

Non-Pretest Administrations (Admin. III and Admin. IV)

	Section 2		Section 3	
	No. of Items	%	No. of Items	%
First	23	57.5	37	61.7
Middle	7	17.5	8	13.3
Last	10	25.0	15	25.0

After predicted scores were obtained on the last set of items ($\hat{Y}_{last}$) and the immediately preceding set of items ($\hat{Y}_{mid}$), residuals were computed for both regressions and then standardized. The standardized residual is $z = (Y - \hat{Y})/s_{y \cdot x}$, where $s_{y \cdot x}$ is the standard error of estimate. Assuming that the errors in prediction are normally distributed, probabilities can be computed that an examinee's observed score falls within or outside some specified range. To conclude that virtually all examinees completed the first 75 percent of the items (the first speededness criterion), the proportion of examinees with standardized residuals (z_{mid} and z_{last}) both below some criterion (such as $z = -1.645$) must be small. If Y_{mid} and Y_{last} are both improbably low relative to their predicted counterparts, where there is less than approximately a 5 percent chance of obtaining each score by chance alone, it is likely that these examinees either responded in a random or patterned fashion or did not reach the last 25 percent of the items. Random or patterned responding to the last 25 percent of the items is likely due to speededness. However, it is also possible that examinees responded in such a way because the items were too difficult. An index of speededness for this criterion was computed by dividing the proportion of examinees with z's below -1.645 on both item sets by the proportion of examinees with z's below -1.645 on either the last or middle item set, whichever corresponding proportion was smaller. The index ranges from approximately 0 to 1.0.

Since it might be difficult for low-ability examinees to score significantly below predicted scores, the proportion of examinees with observed scores on both middle and last item sets significantly above predicted scores was also determined. These examinees had standardized residuals above $z = +1.645$ for both the middle and last item sets. Along the line of Bejar's (1985) work, these proportions represent lower-ability examinees performing better than predicted on the difficult items at the end of the section. As with the proportion of examinees with z's below -1.645, an index of speededness was computed as the proportion of examinees with z's above +1.645 for both the middle and last item sets by dividing the proportion by either the proportion for the middle item set or the last item set, whichever was smaller. Both the proportions (those below $z = -1.645$ and those above $z = +1.645$) were compared to the proportion of examinees that reached the 75 percent point on the sections and the numbers and proportions of examinees that responded with the same response (e.g. A,A,A....) to all of the last 25 percent of the items.

Determining Whether 80 Percent of Examinees Reached the Last Set of Items

The second part of the study involved answering the question of whether 80 percent of the examinee group truly completed the last set of items in each section. The technique used to answer this question also employed regression analysis. The technique capitalized on the fact that both Sections 2 and 3 of TOEFL both contain two parts with items in each part loosely sequenced so as to increase in order of difficulty. While it was possible to determine the relative extent to which examinees completed the last sets of items in the sections, it was not totally possible with the procedure to determine precisely the percentage of examinees that truly reached the last set of items. To support the validity of the results obtained from using this procedure, data were collected on the proportion of examinees that did not reach the last item as well as the proportion of examinees that responded to the last set of items with the same response.

With this procedure, for both Sections 2 and 3, for both pretest and non-pretest administrations, three raw scores were obtained. One score (Y_1)

was based on responses to the last six items in the sections (Written Expression items for Section 2 and Reading Comprehension items for Section 3). These items are typically the most difficult items in the second parts of the sections. For these items, mean deltas[1] were computed after adjusting for dropout (i.e., examinees who did not reach the items). A second score (Y_2) was computed for a set of six equally difficult items from the first and middle parts of the sections. A third score (X) was based on responses to the remaining items, which consisted of both item types and excluded items that had been contained in the last 25 percent of the items in the sections. Twenty-five percent was used as a cut-off to ensure that the scores for the predictor items were likely to be unaffected by speededness, especially if the sections were found not to be speeded according to the speededness criterion that virtually all examinees reached the 75 percent point for the section. Table 2 contains the deltas for the items on which Y_1 and Y_2 scores were based for each administration.

Y_2 scores were then regressed onto X. From the regression equation, predicted scores ($\hat{Y}_2$) were computed. Also computed was the standard error of estimate, $s_{y \cdot x}$, which was computed as $s_{y2 \cdot x} = (1 - r^2_{y2 \cdot x})^{1/2}$. Instead of then computing standardized residuals for observed Y_2 scores, standardized residuals were computed by substituting Y_1 scores for Y_2 scores. If standardized residuals were computed for Y_2 scores, 5 percent of observed scores would be expected to be significantly below their predicted counterparts using a z score of -1.645. However, if so-called pseudo standardized residuals were computed using the observed Y_1 scores, the proportion of examinees with pseudo standardized residuals below $z' = -1.645$ would be greater than .05 if random or patterned responding was occurring due to speededness. The pseudo standardized residuals are computed using the observed Y_1 score, the predicted $\hat{Y}_2$ score and the standard error of estimate for Y_2. Or,

$$z' = \frac{Y_1 - \hat{Y}_2}{s_{y2 \cdot x}}$$

[1] Delta is an index of item difficulty used at Educational Testing Service. It is a function of the proportion of examinees correctly responding to an item. In practice, the index ranges from 3.5 (easy) to 22.5 (difficult).

Table 2
Deltas for Last Six Items of Section and Six
Difficult Items from First and Middle Parts of Section for
Sections 2 and 3 of TOEFL for Four Administrations

Section 2

Administration I

	Deltas for First Set of Six Items Struct. & W. E.		Deltas for Last Six Items Written Exp.
	11.4		12.4
	13.3		12.7
	12.8		14.6
	13.2		14.9
	12.2		9.3
	12.3		12.1
Mean	12.53	Mean	12.67

Administration II

	11.8		13.5
	11.5		11.0
	14.5		12.5
	12.6		15.3
	13.2		13.4
	14.0		12.2
Mean	12.93	Mean	12.98

Administration III

	9.0		11.0
	10.7		8.9
	11.4		12.0
	12.1		11.2
	12.7		12.1
	8.1		8.7
Mean	10.66	Mean	10.65

Administration IV

	13.2		13.7
	16.0		13.6
	10.8		13.8
	11.4		12.7
	11.4		12.1
	13.3		11.4
Mean	12.68	Mean	12.88

Table 2 (continued)

Section 3

	Deltas for First Set of Six Items Vocab. & R.C.		Deltas for Last Six Items Reading Comp.
	Administration I		
	3.4		13.6
	12.3		12.1
	14.2		12.6
	15.3		14.2
	11.4		13.0
	11.4		11.2
Mean	13.00	Mean	12.95
	Administration II		
	11.2		12.7
	13.6		11.9
	15.3		12.2
	14.3		13.7
	13.2		14.9
	11.8		14.0
Mean	13.23	Mean	13.23
	Administration III		
	12.0		11.2
	12.0		13.8
	12.9		11.6
	11.5		10.4
	12.6		11.8
	11.7		13.8
Mean	12.12	Mean	12.10
	Administration IV		
	12.4		11.7
	13.7		11.5
	13.9		14.1
	13.6		15.2
	13.8		14.6
	13.3		14.5
Mean	13.45	Mean	13.60

The standardized residuals are pseudo residuals in the sense that the observed scores are based on the last six items in the section and the predicted scores and standard error of estimate are based on the six equally difficult items from the first part of the section. Since observed and predicted scores are based on different sets of items, the proportion of examinees with pseudo standardized residuals below, say, $z' = -1.645$ can exceed .05 even if the errors in prediction are normally distributed. Because item difficulty is to some extent being controlled, the proportion of examinees with pseudo standardized residuals below $z' = -1.645$ in excess of .05 may be an indication of the extent to which the section is speeded, which includes the possibility that examinees responded with random or patterned responses to the items at the end of the section. Random or patterned responding at the end of the test would be an indication that either examinees did not have enough time to complete the items or that examinees found the items so difficult they could not eliminate any distractors.

The assessment of speededness for the second speededness criterion does not appear to enable the detection of speededness for examinees with very low scores on the predictor, X. Therefore, the obtained proportion of examinees with Y_1 scores significantly below $\hat{Y}_2$ scores should be viewed as an underestimate. An adjustment is therefore necessary for the inability of the second procedure to detect speededness for low-scoring examinees.

The adjustment consists of lowering the proportion of examinees with pseudo standardized residuals below $z' = -1.645$ required to claim a section is speeded according to the second speededness criterion. Instead of claiming a section is speeded if more than 25 percent of examinees have Y_1 scores significantly below $\hat{Y}_2$ scores, a lower percentage is required. Therefore, to make the claim that fewer than 80 percent of examinees truly completed the last six items, more than approximately 10-15 percent of examinees would have to have obtained values of z' below -1.645. The 25 percent criterion was derived from adding 20 percent to the 5 percent of the distribution that would be expected if there were no differences between Y_1 scores and Y_2 scores. The 10-15 percent criterion was derived from taking into account the fact that the proportion of examinees with pseudo standardized residuals below $z' = -1.645$ might be an underestimate of the pool of potentially affected examinees. However, the 10-15 percent criterion must be viewed as an extremely rough estimate of the proportion needed to claim a section is speeded, since the second procedure is actually best suited to provide the relative extent to which the sections are speeded and cannot pinpoint the exact percentage of examinees truly completing each section.

As part of the analysis, an examination was also made of the proportion of examinees with pseudo standardized residuals above z' = +1.645. This made it possible to detect speededness for low-scoring examinees along the line of Bejar's (1985) work. Significantly higher performance on the last set of six items than predicted for the first set of six items would logically be attributable to random or patterned responding, whereby examinees did not have the time to select the most attractive distractor as their choice of the best answer.

The method for assessing speededness according to the second speededness criterion was to regress scores based on a set of items from mostly one item type onto a criterion and to compute standardized residuals using a second item type. This did not seem to pose a problem since the scores on the two parts of each section are highly correlated (about r = .90 or greater). The values obtained from the regression procedure used to evaluate the second speededness criterion were compared to Gulliksen's (1950) index of speededness, Stafford's (1971) speededness quotient, the proportion of examinees responding systematically with the same response to the last items in the section, and the ETS criterion of the proportion of examinees reaching the last item in the section without accounting for random or patterned responding.

Results

To What Extent Did Virtually All Examinees Reach the 75 Percent Point on the Sections?

Tables 3 and 4 present the numbers and proportions of examinees with standardized residuals below z = -1.645 and above z = +1.645 for the middle, last, and both middle and last item sets for Sections 2 and 3 of TOEFL. As can be seen from these data, the proportions of examinees with standardized residuals below z = -1.645 on both middle and last item sets range from .009 to .015 for Section 2 and from .010 to .016 for Section 3. This indicates that relatively small proportions of the same examinees scored lower than predicted on both the middle and last item sets. The index obtained by dividing the proportion of examinees scoring significantly below predicted on both item sets by the proportion of examinees scoring significantly below predicted on either item set, whichever was smaller, can range from approximately 0 to 1.0. The values for the index were relatively low, from .159 to .234 for Section 2 and from .139 to .250 for Section 3.

BEST COPY AVAILABLE

Table 3
Numbers and Proportions of Examinees with Standardized Residuals
Below z = -1.645 and Above z = +1.645 for Middle, Last, and
Both Middle and Last Item Sets for Section 2 of TOEFL
for Four Administrations

	n below	Prop. below	Index	n above	Prop. above	Index
	Pretest Administrations					
Admin. I (n = 1624)						
Middle Item Set	106	.065		63	.039	
Last Item Set	93	.057		53	.033	
Both Item Sets	20	.012	.216	12	.007	.224
Admin. II (n = 1042)						
Middle Item Set	63	.060		30	.029	
Last Item Set	58	.056		36	.035	
Both Item Sets	13	.012	.223	3	.003	.099
	Non-Pretest Administrations					
Admin. III (n = 2766)						
Middle Item Set	186	.067		117	.042	
Last Item Set	180	.065		71	.026	
Both Item Sets	42	.015	.234	17	.006	.236
Admin. IV (n = 3728)						
Middle Item Set	228	.061		124	.033	
Last Item Set	221	.059		107	.029	
Both Item Sets	35	.009	.159	11	.003	.102

As for differences between pretest and non-pretest administrations for the first speededness criterion that virtually all examinees reach the 75 percent point on each of the sections, no clear pattern emerged. From these data, it does not appear that the sections were speeded according to the first speededness criterion.

Table 4
Numbers and Proportions of Examinees with Standardized Residuals Below z = -1.645 and Above z = +1.645 for Middle, Last, and Both Middle and Last Item Sets for Section 3 of TOEFL for Four Administrations

	Pretest Administrations					
	n below	Prop. below	Index	n above	Prop. above	Index
Admin. I (n = 1624)						
Middle Item Set	104	.064		58	.036	
Last Item Set	119	.073		35	.022	
Both Item Sets	26	.016	.250	2	.001	.057
Admin. II (n = 1042)						
Middle Item Set	72	.069		31	.030	
Last Item Set	112	.107		99	.095	
Both Item Sets	10	.010	.139	6	.006	.192
	Non-Pretest Administrations					
Admin. III (n = 2766)						
Middle Item Set	163	.059		109	.039	
Last Item Set	176	.064		86	.031	
Both Item Sets	37	.013	.227	16	.006	.187
Admin. IV (n = 3728)						
Middle Item Set	229	.061		141	.038	
Last Item Set	230	.062		128	.034	
Both Item Sets	48	.013	.211	13	.003	.103

The numbers and proportions of examinees with standardized residuals above z = +1.645 were also computed for each section for each administration. These results also appear to be in the "safe" range for these sections. A relatively small proportion of apparently lower-ability examinees scored significantly higher than predicted on both the middle and last item sets. The fact that smaller proportions of examinees had standardized residuals above z = +1.645 than below z = -1.645 suggests that the distribution of the errors in prediction may have been positively skewed. Other data that can be brought to bear on whether the sections were speeded according to the first speededness criterion are the proportions of examinees reaching the 75 percent point on each section. These data are contained in Table 5.

Table 5
Percentage Completing Section, Percentage Completing 75 Percent of Section, Number of Items Reached by 80 Percent of Examinees, Gulliksen's Index of Speededness, and Stafford's Speededness Quotient for Section 2 and Section 3

Section 2

Speededness Measure	Admin. I	Admin. II	Admin. III	Admin. IV
Percentage Completing Section	98.4	98.2	99.1	98.4
Percentage Completing 75 Percent of Section	99.7	99.8	99.8	99.8
Number of Items Reached by 80 Percent of Examinees	60	60	40	40
Gulliksen's Index of Speededness	.03	.02	.02	.02
Stafford's Speededness Quotient	.01	.01	.01	.01

Section 3

Speededness Measure	Admin. I	Admin. II	Admin. III	Admin. IV
Percentage Completing Section	92.2	93.2	96.4	94.5
Percentage Completing 75 Percent of Section	99.6	99.8	99.6	99.7
Number of Items Reached by 80 Percent of Examinees	90	90	60	60
Gulliksen's Index of Speededness	.08	.03	.02	.04
Stafford's Speededness Quotient	.03	.02	.02	.02

Table 5 shows the percentage completing 75 percent of the section for Sections 2 and 3. The percentages are very high and, as a set, quite homogeneous: 99.7 percent - 99.8 percent for Section 2 and 99.6 percent - 99.8 percent for Section 3. However, these data do not include examinees who may have responded randomly or with patterned responses to the items at the end of the section as the time limit approached, and therefore cannot be considered a pure indication of speededness, especially for a rights-only scored test. More realistically, some portion of the examinee group may not have reached the 75 percent point on the test. This is supported by the the number and proportion of examinees that responded with the same response to the last 25 percent of items in each of the sections. Table 6 presents these data.

Table 6
Numbers and Proportions of Examinees Indicating the Same Response to the Last 25 Percent of Items for Sections 2 and 3 of TOEFL for Four Administrations

	Pretest Administrations					
	Section 2			Section 3		
Administration	Number of Successive Identical Responses	n	Prop.	Number of Successive Identical Responses	n	Prop.
Administration I	15	11	.007	23	7	.004
Administration II	15	4	.004	23	5	.005
	Non-Pretest Administrations					
Administration III	10	3	.001	15	4	.001
Administration IV	10	11	.003	15	13	.003

As can be seen from Table 6, a very small proportion of examinees responded with the same response to the last 25 percent of the items in each section. These data provide an indication that some very small proportion of examinees may have guessed randomly at the last 25 percent of items. The data would more likely reflect the presence or absence of speededness if one added to it: (1) the proportion of examinees that responded in a random or patterned fashion without responding with the same response, (2) the proportion that did not reach the 75 percent point (i.e., failed to respond to the last 25 percent of items), and (3) the proportion of examinees that in some combination responded randomly, in a patterned fashion, or failed to respond to the last 25 percent of items. However, such occurrences do not appear widespread enough to claim the sections were speeded according to the first speededness criterion. On the average, 99.7 percent of examinees reached the last 25 percent of items (from Table 5), and, at most, only .7 percent of examinees responded with the same response to the last 25 percent of items. In total, it can be roughly estimated that slightly more than 1 percent of the examinee group did not truly reach the last 25 percent of items in the sections. It appears that this percentage is associated with the pretest administrations rather than the non-pretest administrations. For the non-pretest administrations, the proportion of examinees that did not truly reach the 75 percent point appears to be less than 1 percent.

To What Extent Did 80 Percent of Examinees Reach the Last Set of Items?

While it was not possible to determine the exact percentage of examinees that truly completed each of the sections, it was possible to generate data to indicate the relative extent to which examinees completed the sections. Tables 7 and 8 present the numbers and proportions of examinees with pseudo standardized residuals below z' = -1.645 and above z' = +1.645 for Sections 2 and 3, respectively, of TOEFL for four administrations.

Table 7
Numbers and Proportions of Examinees With Standardized Residuals and Pseudo Standardized Residuals Below z = -1.645 and Above z = +1.645 for Section 2 of TOEFL for Four Administrations

Pretest Administration	n below	Prop. below	n above	Prop. above
Admin. I (n = 1624)				
Standardized Residual	93	.057	47	.029
Pseudo Stand. Residual	102	.062	39	.024
Admin. II (n = 1042)				
Standardized Residual	65	.062	27	.026
Pseudo Stand. Residual	93	.089	37	.036
Non-Pretest Administration				
Admin. III (n = 2766)				
Standardized Residual	188	.067	99	.036
Pseudo Stand. Residual	274	.099	105	.038
Admin. IV (n = 3728)				
Standardized Residual	197	.053	263	.071
Pseudo Stand. Residual	361	.097	332	.089

Table 8
Numbers and Proportions of Examinees With Standardized Residuals and Pseudo Standardized Residuals Below z = -1.645 and Above z = +1.645 for Section 3 of TOEFL for Four Administrations

Pretest Administration	n below	Prop. below	n above	Prop. above
Admin. I (n = 1624)				
Standardized Residual	82	.050	73	.045
Pseudo Stand. Residual	264	.163	121	.075
Admin. II (n = 1042)				
Standardized Residual	57	.055	47	.045
Pseudo Stand. Residual	113	.108	62	.060
Non-pretest Administration				
Admin. III (n = 2766)				
Standardized Residual	170	.061	72	.026
Pseudo Stand. Residual	204	.074	80	.029
Admin. IV (n = 3728)				
Standardized Residual	182	.049	158	.042
Pseudo Stand. Residual	289	.076	185	.050

The highest proportions of examinees with pseudo standardized residuals below z' = -1.645 were for Section 3 for Administration I (.163), a pretest administration and Section 3 for Administration II (.108), also a pretest administration. For the other sections, the proportion of examinees with pseudo standardized residuals below z' = -1.645 was below .10. For sections for those administrations for which fewer than 10 percent of the examinees obtained pseudo standardized residuals below z' = -1.645, observed scores for the six difficult items at the end of the sections were not much lower than would be expected using observed scores for the six equally difficult items from the first part of the sections.

Before discussing whether the TOEFL sections were speeded according to the second speededness criterion -- that 80 percent of examinees complete the section -- it seems appropriate to discuss those sections with the highest proportions of examinees with pseudo standardized residuals below z' = -1.645. To assist in evaluating the relative extent to which the sections may be truly

BEST COPY AVAILABLE

speeded, the reader is referred to Table 9, which presents the numbers and proportions of examinees indicating the same response for the last four, five, and six items for Sections 2 and 3.

Table 9
Numbers and Proportions of Examinees Indicating the Same Responses to the Last Four, Five, and Six Items for Sections 2 and 3 of TOEFL for Four Administrations

Section 2

Administration	Last Four Items		Last Five Items		Last Six Items	
	n	Prop.	n	Prop.	n	Prop.
Admin. I (n = 1624)	64	.039	57	.035	43	.026
Admin. II (n = 1042)	62	.060	47	.045	22	.031
Admin. III (n = 2766)	128	.046	110	.040	30	.011
Admin. IV (n = 3728)	73	.020	48	.013	33	.009

Section 3

	n	Prop.	n	Prop	n	Prop.
Admin. I (n = 1624)	206	.127	166	.102	145	.089
Admin. II (n = 1042)	115	.110	97	.093	73	.070
Admin. III (n = 2766)	117	.042	87	.031	73	.026
Admin. IV (n = 3728)	260	.070	197	.053	144	.039

For Section 3 for both pretest administrations (Administration I and Administration II), examinees scored somewhat lower on the last set of six items than was predicted for them on the first set of six items. The proportion of examinees with pseudo standardized residuals below z' = -1.645 for Administration I was highest (.163) (see Table 8). From Table 2, it can be observed that the mean delta for the first and last sets of six items were closely matched, with the mean delta for the first item set only slightly exceeding the mean delta for the last item set. From Table 9, one can see that Administration I also had the highest proportion of examinees with successive identical responses to the last six items (.089). This correspondence also held for Section 3 of Administration II. For Administration II, the proportion of examinees with pseudo standardized residuals below $z' = -1.645$ was second highest (.108) (see Table 8), while the proportion of examinees with successive identical responses to the last six items was second highest (.070).

The pattern does not hold for Section 2. In fact, a negative relationship can be roughly observed for Section 2 between the proportion of examinees with pseudo standardized residuals below $z' = -1.645$ and the number of examinees with successive identical responses to the last four items. For Section 2 for Administration II, the proportion of examinees with pseudo standardized residuals below $z' = -1.645$ was .089 (see Table 7), while there were 62 examinees (.060) who responded with identical responses to the last four items. For Administration III, the proportion of examinees with pseudo standardized residuals below $z' = -1.645$ was .099, while the number of examinees with identical responses to the last four items was 128 (.046). The proportion of examinees with pseudo standardized residuals below $z' = -1.645$ for Administration IV (.097) may be due to the fact that, for this section, the matching of deltas for the item sets was poorest (see Table 2). With the items at the end of the section on the average considerably more difficult, the observed scores for this item set would likely be lower, thereby resulting in a greater proportion of examinees with pseudo standardized residuals below $z' = -1.645$. It is likely that the proportion would be lower if the item sets were matched more closely in terms of mean delta.

Based on the results, if the number of successive identical responses at the end of a section can be considered a rough indicator of the relative extent to which random or patterned responding is occurring due to speededness, then proportions of examinees with pseudo standardized residuals below a certain z' score may hold promise, when applicable, for identifying potentially speeded sections. However, one must also consider the possibility that examinees responded in a patterned fashion at the end of a section because the items at the end are the most difficult. One problem with the second procedure, of course, is the difficulty in finding an adequate match for the last item set in terms of mean item difficulty. A second problem lies in determining how great the proportion must be before a section can be considered speeded according to the second speededness criterion (i.e., that 80 percent of examinees complete the section).

Other data that have traditionally had a bearing on the question of speededness are contained in Table 5. These measures are the percentage completing the section, number of items reached by 80 percent of the examinees, Culliksen's index of speededness, and Stafford's speededness quotient. As can be seen from the table, Section 3 appears slightly more speeded than Section 2 according to the second speededness criterion. Also, in agreement with the procedures developed for this study, for Section 3, the pretest administrations appear to be more speeded than the non-pretest administrations.

As for addressing the question of whether the sections are speeded in absolute terms according to the second speededness criterion, it is not yet possible to make a definitive determination. However, it can be said that Section 3 for pretest administrations appears more speeded than the other sections. If one were to use the cut-off of 10-15 percent to claim a section was speeded, Section 3 for Administration I and Section 3 for Administration II may have been slightly speeded. The proportion of examinees with pseudo standardized residuals below $z'= -1.645$ for Administration I was .163 (see Table 8). The proportion of examinees with pseudo standardized residuals below $z' = -1.645$ for Administration II was .108. However, without guidelines connecting the proportions of examinees with pseudo standardized residuals below $z' = -1.645$ with proportions of examinees completing a section, it is difficult to know if the sections were speeded in absolute terms. Data that tend to corroborate the tentative conclusion that Section 3 pretest administrations were slightly speeded are based on the proportion of examinees completing these sections and the proportion of examinees responding with identical responses to the last four items. Of the 1,624 examinees included in the group that took Administration I, 206 examinees or 12.7 percent responded with the same response to the last four items of Section 3. For Administration II, this figure was 11 percent. If one added to these figures the proportion of examinees that did not complete each of the sections (100 percent - 92.2 percent = 7.8 percent for Administration I and 100 percent - 93.2 percent = 6.8 percent for Administration II), the percentages would rise to 20.5 percent for Administration I and 17.8 percent for Administration II. Since these figures still do not include the proportion of examinees that may have responded in a random or patterned fashion to the last set of items in the section without having responded with the same response, the proportions truly reaching the last set of items for both Administration I and Administration II may not have exceeded 80 percent.

From the results, it appears that 80 percent or more of the examinee group truly completed Section 2 for all four administrations. In addition, Section 3 non-pretest administrations did not appear to be speeded according to this criterion. For these sections, the pseudo standardized residuals were less than .10. The percentages of examinees not completing each section, the traditional measures of speededness, and the proportion of examinees responding with the same response tended to confirm this finding.

Discussion

At present, most speededness measures are based on the notion that speededness is present only when examinees fail to respond to items at the end of the test or test section. For rights-only scored tests such as TOEFL, however, it is possible that some portion of the examinee group responds in a random or patterned fashion to unconsidered items at the end of the test as the time limit approaches. For this reason, present methods for assessing speededness are likely to underestimate the degree to which a test or section is actually speeded. The present study utilized two exploratory yet relatively simple approaches to the problem of assessing the extent to which the TOEFL is truly speeded according to established criteria.

To address the question of whether virtually all examinees reached the 75 percent point on the test, Sections 2 and 3 of TOEFL for four administrations were divided into three item sets. The procedure applied was based on the premise that if an examinee scored lower than predicted on both the last and middle item sets, it was possible that the examinee may have either "not reached" the last 25 percent of items or guessed randomly at these items. While there was evidence that examinees either did not reach the last 25 percent of items or responded in a patterned fashion to these items, the problem was not widespread enough to warrant a claim that the sections were speeded. For the two non-pretest administrations, there is virtually no evidence of speededness according to the first speededness criterion. To support the claim that the sections were not speeded according to the first speededness criterion, an examination was also made of the proportion of examinees reaching the 75 percent point on the test as well as the proportion of examinees that responded with the same response to the last 25 percent of the items. These data both indicate that a very small proportion of examinees did not truly reach the 75 percent point. If one were to interpret the criterion strictly, perhaps it could be said that the pretest administrations are very slightly speeded. However, because there are potentially other reasons such as item difficulty and motivational factors, that may be related to the fact that examinees did not truly reach the 75 percent point, a claim that the pretest administration sections were not speeded according to the first speededness criterion is more plausible.

The second procedure addressed the question of whether at least 80 percent of examinees truly completed each of the sections. While the first procedure may be applicable to other testing programs, the second procedure is limited in use to tests that either have two or more parts or do not have items ordered in terms of increasing difficulty. The second procedure entails matching, with respect to item difficulty, the set of items at the end of the section with items near the beginning or middle of the section. Essentially, the procedure is based on the proportion of examinees that score lower on the last set of six items than predicted for the first set of six items. If the item sets were perfectly matched and scores for the item sets were perfectly correlated, it would be expected that the two item sets produce identical results in terms of the proportions of examinees with standardized residuals below a certain value of z'. However, if the item sets are matched only in terms of mean difficulty, it is possible that the distribution of scores for the two item sets can differ in terms of variability, skewness, and kurtosis. If examinees were affected by speededness at the end of a section, their observed scores on the last item set would likely be lower than predicted for the first item set. Therefore, there is a greater likelihood that examinees would obtain standardized residuals below $z' = -1.645$. Assuming mean deltas are equal for the two item sets, other examinees unaffected by speededness would likely score higher on the last item set. The result would be a more platykurtic distribution of residuals for the last item set than for the first item set. The resulting proportion of examinees with pseudo standardized residuals below z' = -1.645 may, therefore, take on values greater than .05.

From the results of the study, it appears possible that, for Administrations I and II, fewer than 80 percent of examinees truly completed Section III. These sections had pseudo standardized residuals above .10. However, there are potentially a number of factors that would tend to reduce the proportion not truly reaching the last set of items. One possibility is that some portion of the examinee group responded with the same response to the last four items because the four successive identical responses were perceived to be the best choices as answers to the last four questions. A second factor has to do with whether the examinees responded with the same response because the last questions were very difficult; the examinees had enough time but guessed blindly at each item with the same response.

If one can assume that random or patterned responding is due to speededness, one can safely assert that Section 3 pretest administrations were slightly speeded according to the second speededness criterion. However, one cannot be sure that patterns or omissions at the end of the section constitute a speededness effect, especially given the graduated difficulty of the items. Nevertheless, speededness cannot be ruled out as a possible cause. The items in the second part of Section 3 are reading comprehension items, which take more time per item since examinees must read a passage, then read the items associated with the passage, and then, in many cases, refer back to the passage. Possibly, because there are 45 reading comprehension items in Section 3 pretest administrations, less than 80 percent of the examinee group appears to have truly completed the section in at least one of the two pretest administrations studied.

Although the results for both procedures appear reasonable, some problems exist for these procedures. For both procedures, there is the problem of test-taking style. Both procedures assume that the items are answered in sequence. However, it is likely that some portion of the examinee group goes through the test quickly, omitting items about which they are unsure and returning to these items at the end of the time limit. While it may be more likely that items at the end of the test or test section are not answered or responded to randomly, there is nonetheless some portion of the examinee group that responds randomly to items in the middle of the test or test section.

A problem with the first procedure is that the measure appears to reach a limit in the proportion of examinees that can be identified with z's below -1.645 on both middle and last item sets. When the proportion of examinees affected by speededness is large, the procedure becomes less effective. However, the procedure appears ideal when testing whether some relatively small proportion of the examinee group (25 percent or less) is affected by speededness on at least the last 25 percent of the items. The procedure is also more effective when the middle and last item sets are large enough so that a sizable portion of the examinee group can obtain observed scores significantly below predicted scores. A second limitation of the second procedure lies in its inability to detect speededness when responses to both the first and last sets of six items are affected by speededness.

Contrary to what one might think, the procedures themselves do not appear to be adversely affected by the potential confounding of item difficulty and speededness. With the first procedure, the difficulty of items is taken into account by the simple linear regression of the scores for the last item set onto the scores for the first item set. For the second procedure, by matching the first and last sets of six items in mean item difficulty, the procedure detects that portion of the difficulty of the last item set that may be attributable to speededness.

The methods used in conducting this study were exploratory and, in many ways, specific to TOEFL. The first procedure was intended to detect whether a relatively small proportion of the examinee group did not reach the 75 percent point on each section. In addition, it was necessary for the sections to contain enough items so examinees could score significantly below predicted scores on both middle and last item sets. The second procedure was specific to TOEFL in that it was necessary to match in terms of mean difficulty the items at the end of the sections with items near the beginning and middle of the sections. To bolster the validity of decisions made with these procedures, it is recommended that an examination be made of the percentage not completing each section as well as the proportion of examinees that respond to the last set of items with the same response. A single generalized index of speededness has yet to be developed which can account for random or patterned responding.

Recommendation

The results of the study suggest that Section 3 for pretest administrations may be slightly speeded according to the second speededness criterion. It is possible that 80 percent or more of the examinee group did not truly complete the section within the given time limit. However, it must be noted that the criteria against which TOEFL is being assessed for speededness are in some sense arbitrary. Why is 80 percent completing the test the standard for a nonspeeded test? Swineford (1974) contends that the 80 percent who finish the test on time are likely to include all the able examinees, while the other 20 percent of examinees would be unlikely to increase their scores if the time limit were extended. If one were to accept the standard, it would be recommended that another investigation (e.g., survey, observational study) be conducted that more directly determines the extent of the problem. If the findings of this study are confirmed, it would be recommended that the TOEFL program investigate ways to increase the amount of time allotted per item. The time limit for Section 3 pretest administrations could, for example, be extended slightly from 65 minutes to between 68 and 70 minutes. An increase of five minutes would result in an increase in the number of seconds per item for Section 3 pretest administrations from 43.3 to 46.7. This compares to 45 seconds, which is presently the number of seconds allotted each item for Section 3 non-pretest administrations, which consist of 60 items (30 vocabulary and 30 reading comprehension).

References

Bejar, I. I. (1985). *Test speededness under number-right scoring: an analysis of the Test of English as a Foreign Language* (RR-85-11), Princeton, NJ: *Educational Testing Service*.

Cronbach, L. J. & Warrington, W. G. (1951). Time limit tests: estimating the reliability and degree of speeding. *Psychometrika*, *14*, 167 - 188.

Donlon, T. F. (1980). *An exploratory study of the implications of test speededness*. GRE Board Professional Report, GREB No. 76-9P, Princeton, NJ: Educational Testing Service.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons.

Hecht, L. W. & Swineford, F. (1981). *Item analysis at Educational Testing Service*. Princeton, NJ: Educational Testing Service.

Hicks, M. M, Secolsky, C. & Skelton, C. F. (1987). *Test Analysis, Test of English as a Foreign Language* (Form 3HTF8; August 1986 Administration). Princeton, NJ: Educational Testing Service.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement*, *4*, 261-270.

Secolsky, C. (1985). *Evaluation of speededness indices for use with TOEFL*. Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Stafford, R. E. (1971). The speededness quotient: a new descriptive statistic for tests. *Journal of Educational Measurement*, *8*, 275-278.

Swineford, F. (1974). *The test analysis manual*. (SR-74-06) Princeton, NJ: Educational Testing Service.

TOEFL Research Reports currently available...

Report 1. *The Performance of Native Speakers of English on the Test of English as a Foreign Language.* John L D Clark November 1977

Report 2. *An Evaluation of Alternative Item Formats for Testing English as a Foreign Language.* Lewis W Pike June 1979

Report 3. *The Performance of Non-Native Speakers of English on TOEFL and Verbal Aptitude Tests* Paul J. Angelis. Spencer S Swinton. and William R Cowell October 1979

Report 4. *An Exploration of Speaking Proficiency Measures in the TOEFL Context.* John L. D Clark and Spencer S Swinton October 1979

Report 5. *The Relationship between Scores on the Graduate Management Admission Test and the Test of English as a Foreign Language* Donald E Powers December 1980.

Report 6. *Factor Analysis of the Test of English as a Foreign Language for Several Language Groups.* Spencer S Swinton and Donald E Powers. December 1980.

Report 7. *The Test of Spoken English as a Measure of Communicative Ability in English-Medium Instructional Settings* John L D Clark and Spencer S Swinton. December 1980.

Report 8. *Effects of Item Disclosure on TOEFL Performance* Gordon A Hale. Paul J Angelis. and Lawrence A. Thibodeau December 1980.

Report 9. *Item Performance Across Native Language Groups on the Test of English as a Foreign Language* Donald L Alderman and Paul W Holland August 1981

Report 10. *Language Proficiency as a Moderator Variable in Testing Academic Aptitude.* Donald L Alderman. November 1981

Report 11. *A Comparative Analysis of TOEFL Examinee Characteristics. 1977-1979.* Kenneth M. Wilson. July 1982.

Report 12. *GMAT and GRE Aptitude Test Performance in Relation to Primary Language and Scores on TOEFL.* Kenneth M. Wilson July 1982

Report 13. *The Test of Spoken English as a Measure of Communicative Ability in the Health Professions: Validation and Standard Setting* Donald E Powers and Charles W Stansfield. January 1983

Report 14. *A Manual for Assessing Language Growth in Instructional Settings.* Spencer S Swinton. February 1983.

Report 15. *Survey of Academic Writing Tasks Required of Graduate and Undergraduate Foreign Students* Brent Bridgeman and Sybil Carlson September 1983

Report 16. *Summaries of Sudies involving the Test of English as a Foreign Language. 1963-1982.* Gordon A Hale. Charles W. Stansfield. and Richard P Duran. February 1984

Report 17. *TOEFL from a Communicative Viewpoint on Language Proficiency: A Working Paper.* Richard P. Duran. Michael Canale. Joyce Penfield. Charles W Stansfield. and Judith E Liskin-Gasparro February 1985.

Report 18. *A Preliminary Study of Raters for the Test of Spoken English* Isaac I. Bejar. February 1985.

Report 19. *Relationship of Admission Test Scores to Writing Performance of Native and Nonnative Speakers of English* Sybil B Carlson. Brent Bridgeman. Roberta Camp. and Janet Waanders. August 1985.

Report 20. *A Survey of Academic Demands Related to Listening Skills.* Donald E Powers. December 1985

Report 21. *Toward Communicative Competence Testing Proceedings of the Second TOEFL Invitational Conference* Charles W Stansfield May 1986

Report 22. *Patterns of Test Taking and Score Change for Examinees Who Repeat the Test of English as a Foreign Language.* Kenneth M. Wilson. January 1987

Report 23. *Development of Cloze-Elide Tests of English as a Second Language.* Winton Manning. April 1987.

Report 24. *A Study of the Effects of Item Option Rearrangement on the Listening Comprehension Section of the Test of English as a Foreign Language.* Marna Golub-Smith August 1987.

Report 25. *The Interaction of Student Major-Field Group and Text Content in TOEFL Reading Comprehension.* Gordon A Hale. January 1988.

Report 26. *Multiple-Choice Cloze Items and the Test of English as a Foreign Language.* Gordon A Hale. Charles W. Stansfield. Donald A Rock. Marilyn M Hicks. Frances A. Butler. and John W Oller. Jr. March 1988

Report 27. *Native Language. English Proficiency. and the Structure of the Test of English as a Foreign Language.* Philip K Oltman. Lawrence J. Stricker. and Thomas Barrows. July 1988.

Report 28. *Latent Structure Analysis of the Test of English as a Foreign Language* Robert F Boldt. November 1988

Report 29. *Context Bias in the Test of English as a Foreign Language* William H Angoff January 1989

Report 30. *Accounting for Random Responding at the End of the Test in Assessing Speededness on the Test of English as a Foreign Language* Charles Secolsky January 1989

ERIC Full Text Provided by ERIC

BEST COPY AVAILABLE

00102 • S128P3 • 275631 • Printed in U.S.A