Technical Report

TR-13

Concurrent Calibration of Dichotomously and Polytomously Scored TOEFL Items Using IRT Models

K. Linda Tang
Daniel R. Eignor
Concurrent Calibration of Dichotomously and Polytomously Scored TOEFL Items Using IRT Models

K. Linda Tang
Daniel R. Eignor

Educational Testing Service
Princeton, New Jersey

RR-97-6
Abstract

In order to meet the needs of the Test of English as a Foreign Language (TOEFL®) constituencies, the TOEFL program is sponsoring a development project known as TOEFL 2000. Drawing from current linguistic theory and models of communicative competence, it is anticipated that the new test or test battery developed by the TOEFL 2000 project will likely be designed to test all four language skills -- reading, writing, listening, and speaking -- in an integrated fashion. However, one compromise level or position on integration of skills is one in which reading and writing would be tested together, and listening and speaking also tested together. It is also assumed that the test will largely be performance-based, meaning a substantial portion of the items on the test will likely be constructed-response items, and an examinee’s score on such items will be in one of multiple ordered categories.

Two groups of item response theory (IRT) models have been developed to calibrate items with multiple ordered categories (i.e., polytomously scored items): (a) the partial credit model (Masters, 1982) and the generalized partial credit model (Muraki, 1992); and (b) the graded response model (Samejima, 1969,1972). These models have been used jointly with the dichotomous three parameter logistic (3PL) IRT model to concurrently calibrate dichotomously and polytomously scored items for the National Assessment of Educational Progress (NAEP). However, the performance of these polytomous IRT models and the concurrent calibration of dichotomous and polytomous scored items have not been investigated with data from the TOEFL examinee population.

The purpose of this study was to obtain a good understanding of the performance of a combination of dichotomous and polytomous IRT models with TOEFL data. TOEFL Vocabulary and Reading Comprehension and Test of Written English (TWE®) items, and TOEFL Listening Comprehension and Test of Spoken English (TSE®) items were concurrently calibrated using a combination of the generalized partial credit model and the 3PL IRT model. The two sets of combined items were also concurrently calibrated using a combination of the graded response model and the 3PL IRT model.

The results of this study indicate that data from a reading/writing combination made up of the TOEFL Vocabulary and Reading Comprehension section and the TWE were reasonably well fit by a combination of the 3PL and generalized partial credit models or 3PL and graded response models. In a similar fashion, data for a listening/speaking combination made up of the TOEFL Listening Comprehension section and selected tasks from the TSE were also reasonably well fit by the 3PL/generalized partial credit and 3PL/graded response model combinations.

A variety of comparisons across the generalized partial credit and graded response models seem to indicate some preference for using the generalized partial credit model when PARSCALE is used as the calibration program. The results of this study provide useful information about test construction and item calibration procedures that might later be used for the TOEFL 2000 project.
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
Micheline Chalhoub-Deville
JoAnn Crandall
Fred Davidson
Thom Hudson
John A. Upshur (Chair)

Southern Illinois University at Carbondale
University of Minnesota
University of Maryland Baltimore County
University of Illinois at Urbana-Champaign
University of Hawaii at Manoa
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

Special acknowledgment is due the following for their substantial contribution to the success of this project:

Eiji Muraki for his advice regarding the conduct of the study.

Steve Laue and Judy DeChamplain for conducting much of the data analyses.

Isaac Bejar, Eiji Muraki, Mary Schedl, and Spencer Swinton for their constructive comments on a draft of this report.

Eugenia Tye for her secretarial assistance.

The TOEFL Research Committee members for their encouragement and support.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td><strong>Method</strong></td>
<td></td>
</tr>
<tr>
<td>Item and Examinee Data</td>
<td>3</td>
</tr>
<tr>
<td>The IRT Models and Their Parameter Interpretation</td>
<td>6</td>
</tr>
<tr>
<td>Calibration Procedures</td>
<td>12</td>
</tr>
<tr>
<td><strong>Results</strong></td>
<td></td>
</tr>
<tr>
<td>Dimensionality</td>
<td>16</td>
</tr>
<tr>
<td>Effectiveness of Fit</td>
<td>16</td>
</tr>
<tr>
<td>Comparison of Models</td>
<td>24</td>
</tr>
<tr>
<td><strong>Discussion and Further Exploration</strong></td>
<td>30</td>
</tr>
<tr>
<td><strong>References</strong></td>
<td>31-32</td>
</tr>
<tr>
<td><strong>Appendices</strong></td>
<td>33, 38, 43, 66</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Items in TOEFL Vocabulary and Reading Comprehension and TWE Concurrent Calibration</td>
<td>4</td>
</tr>
<tr>
<td>Table 2</td>
<td>Items in Listening Comprehension and TSE Concurrent Calibration</td>
<td>5</td>
</tr>
<tr>
<td>Table 3</td>
<td>Testlet Score Frequencies for Reading Comprehension Sets in TOEFL/TWE Forms A-C</td>
<td>7</td>
</tr>
<tr>
<td>Table 4</td>
<td>TWE Score Frequencies TOEFL/TWE Forms A-C</td>
<td>8</td>
</tr>
<tr>
<td>Table 5</td>
<td>Testlet Score Frequencies for Mini-talk Sets in TOEFL/TSE Forms X and Y</td>
<td>9</td>
</tr>
<tr>
<td>Table 6</td>
<td>TSE Task Score Frequencies TOEFL/TSE Forms X and Y</td>
<td>10</td>
</tr>
<tr>
<td>Table 7</td>
<td>Percentage of Test Variance for the Five Principal Components with Largest Eigenvalues</td>
<td>17</td>
</tr>
<tr>
<td>Table 8</td>
<td>Item Fit Statistics for the Polytomous Items TOEFL/TWE Form A</td>
<td>20</td>
</tr>
<tr>
<td>Table 9</td>
<td>Item Fit Statistics for the Polytomous Items TOEFL/TSE Form X</td>
<td>21</td>
</tr>
<tr>
<td>Table 10</td>
<td>Average Root Mean Squared Difference (RMSD) between the Estimated True Scores and the Observed Scores and the Correlations between the Estimated True Scores and the Observed Scores TOEFL/TWE</td>
<td>25</td>
</tr>
<tr>
<td>Table 11</td>
<td>Average Root Mean Squared Difference (RMSD) between the Estimated True Scores and the Observed Scores and the Correlations between the Estimated True Scores and the Observed Scores TOEFL/TSE</td>
<td>26</td>
</tr>
</tbody>
</table>
List of Figures

| Figure 1     | Plots of eigenvalues for TOEFL/TWE Forms A-C | 18 |
| Figure 2     | Plots of eigenvalues for TOEFL/TSE Forms X-Y | 19 |
| Figure 3     | Bivariate Plots of Observed vs. Estimated True Scores TOEFL/TWE | 27-28 |
| Figure 4     | Bivariate Plots of Observed vs. Estimated True Scores TOEFL/TSE | 29 |

List of Appendices

| Appendix A   | Description of Pre-July 1995 TSE Sections 4-7 from Bulletin of Information for TSE | 33 |
| Appendix B   | Selected 3PL Item-Ability Regression Plots TOEFL/TWE Form A and TOEFL/TSE Form X | 38 |
| Appendix C   | Overall Generalized Partial Credit Model and Graded Response Model Plots and Fit Plots for Each Response Category TOEFL/TWE Form A | 43 |
| Appendix D   | Overall Generalized Partial Credit Model and Graded Response Model Plots and Fit Plots for Each Response Category TOEFL/TSE Form X | 66 |
Introduction

In order to meet the needs of the TOEFL constituencies, the TOEFL program is sponsoring a development project known as TOEFL 2000, which has as its goal the construction of a new test or test battery that takes into consideration current understandings of communicative competence and language assessment (see Bachman, 1990). Although the project is still in its early stages, its goal has been clearly articulated in the 1993 Project Plan for TOEFL 2000: "The program is attempting to do what has not been done before: build a performance-based test of second language ability with a theoretically defensible framework that meets recognized psychometric standards (i.e., the AERA/APA/NCME Standards for Educational and Psychological Testing, 1985)."

Drawing from current linguistic theory and models of communicative competence, it is anticipated that the new test or test battery will likely be designed to test all four language skills -- reading, writing, listening, and speaking -- in an integrated fashion. With such a scenario, dependencies among the skills are likely to be built into the performance tasks on the test, and the usual psychometric models, such as IRT models, which are based on (conditional) independence among responses to items or tasks, are not likely to be applicable. New psychometric models, now in their infancy, will need to be further developed. (See Mislevy, 1995, for one such approach.) At present, it is unclear whether such models can support development of test forms for an ongoing testing program, so different levels of integration of skills are being considered.

For example, instead of integrating tasks and scores on all four skills, reading and writing could be tested together, with scores on both skill areas combined into one overall score. Likewise, listening and speaking could be tested together, with these scores combined into a separate overall score. Depending on the level of independence among the tasks testing these combinations, it may be possible with this scenario to make use of existing psychometric models based on IRT. However, certain of these IRT models, which are described in the following section, have not yet been used with TOEFL data.

Regardless of the level of skills integration demonstrated in the final test, it is assumed that the test will largely be performance-based, meaning a substantial portion of the items on the test will likely be constructed-response items, and an examinee's score on such items will be in one of multiple ordered categories. If integration of skills within the test is at the level that the use of IRT can be considered, the dichotomous 3PL IRT model, currently used to calibrate TOEFL multiple-choice items, each of which has only two response categories (right or wrong), will not be appropriate for these constructed-response items, and a new IRT model will be needed. However, to boost overall test reliability, it is possible that the new test could include some multiple-choice items along with the constructed-response items. In this case, the dichotomous 3PL model would need to be used in conjunction with one of the other models.

Two groups of IRT models have been developed to calibrate items with multiple ordered categories (i.e., polytomously scored items): (a) the partial credit model (Masters, 1982) and the generalized partial credit model (Muraki, 1992); and (b) the graded response model (Samejima, 1969, 1972). These polytomous models are generalized from the dichotomous IRT models and reduce to the dichotomous IRT models when the number of response categories is two. In
other words, a particular dichotomous IRT model can be thought of as a special case of the corresponding polytomous IRT model in which the number of response categories is two.

These polytomous IRT models have been used to successfully calibrate polytomously scored items using both simulated and real data, and have demonstrated good model-data fit (Masters, 1982; Muraki, 1992; Samejima, 1972). These models have been successfully applied to large-scale testing programs, such as the National Assessment of Educational Progress (NAEP) (see the NAEP 1992 Technical Report, 1993) and the Maryland School Performance Assessment Program (MSPAP) (see Huynh & Ferrara, 1994). In addition, these models have been used jointly with the dichotomous 3PL IRT model to concurrently calibrate dichotomously and polytomously scored items for NAEP. Methods for linking item parameters from different calibrations to a common IRT scale have also been developed for both the graded response model (Baker, 1992) and the generalized partial credit model (Muraki & Chang, 1994). Therefore, it appears that most of the major issues in applying the polytomous IRT models to large-scale testing program data have been successfully addressed. If the degree of integration of the new TOEFL 2000 test is at a level that using IRT can be considered, then one of the polytomous models described above would appear to be the most promising candidate for use with the performance-based components on the new test. However, as mentioned previously, the performance of these models on data from the TOEFL examinee population has not yet been investigated.

Using a new test that would integrate reading and writing into a single score and listening and speaking into another score, then combine items in the TOEFL Vocabulary and Reading Comprehension section (Section 3) with a TWE item, and combine items in the TOEFL Listening Comprehension section (Section 1) with items in the TSE test, should provide a good approximation to such a test. Three different combinations or "forms" of TOEFL Section 3 and the TWE test and two different combinations or forms of TOEFL Section 1 and the TSE test were created for the study. In all cases, pre-July 1995 or unrevised TOEFL and TSE forms were used because data from the revised TOEFL test were not available for use when the project began. However, the results based on this available TOEFL data should clearly provide useful initial pilot information in developing the new test.

The purpose of this study was to obtain a good understanding of the performance of a combination of dichotomous and polytomous IRT models on data obtained from the TOEFL examinee population.

The research questions addressed in the study were:

1. Can dichotomous and polytomous IRT models be successfully applied to concurrently calibrate combinations of TOEFL Vocabulary and Reading Comprehension and TWE items? Can these models also be successfully applied to concurrently calibrate combinations of TOEFL Listening Comprehension and TSE items?

2. Which of the two polytomous IRT models, namely, the generalized partial credit model or the graded response model, will produce a better fit to the TOEFL/TWE and TOEFL/TSE data and thus be a better candidate for possible use with TOEFL 2000 test data?

To address these issues, each of the three forms of the TOEFL Vocabulary and Reading Comprehension and TWE items and each of the two forms of the TOEFL Listening Comprehension and TSE items were first concurrently calibrated using
a combination of the generalized partial credit model and the 3PL IRT model. Following this, all sets of combined items were concurrently calibrated using a combination of the graded response and 3PL IRT models. Appropriate comparisons across models were then made.

With respect to the second research question above, having to do with which of the two polytomous models might perform better, a search of the literature located an article of relevance by Maydeu-Olivares, Drasgow, and Mead (1994). Maydeu-Olivares et al. compared fit of the graded response model with the Thissen and Steinberg (1986) extension of the partial credit model, which is a slightly more general model than the generalized partial credit model used in this study. Maydeu-Olivares et al. found that both models fit equally well with the data sets they studied, indicating that more or less comparable fit might be expected as an outcome in this study. In the study of Maydeu-Olivares et al., however, only simulated data were used.

Method

**Item and Examinee Data**

**Item Data:** TOEFL Vocabulary and Reading Comprehension items and the TWE item. In the TOEFL Vocabulary and Reading Comprehension section for the forms used in this study, there were 29 discrete Vocabulary items and 29 Reading Comprehension items that were associated with five short passages. There were from four to six Reading Comprehension items related to each passage. The discrete Vocabulary items were dichotomously scored. The Reading Comprehension items were grouped according to passage; the items for each passage were considered to be a testlet-type item, and were polytomously scored (see Thissen, Steinberg, & Mooney, 1989). For example, suppose a passage had five items, then the passage would have six score categories. The examinee who answered none of these five items correctly would reach score category 1, the examinee who answered one of the five items correctly would reach score category 2, and so on. The examinee who answered all of the five items correctly would reach score category 6 for this testlet-type item.

The scores for a TWE item range from 1.0 to 6.0 in increments of 0.5. (Scores from two raters on a 1 to 6 scale were averaged.) Each score category was treated as a response category, that is, the TWE item had 11 response categories. By directly transferring the 11 discrete score points of the TWE item to 11 response categories, no response information was lost.

The item types, number of items, and the number of response categories for the three TOEFL Vocabulary and Reading Comprehension and TWE combined test forms in this study (referred to as TOEFL/TWE Forms A-C) is summarized in Table 1.

---

1Form A-C was administered in October 1994, February 1995, and May 1995, respectively.
Table 1

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Number of Items</th>
<th>Number of Response Categories for Each Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary\textsuperscript{a}</td>
<td>29</td>
<td>2</td>
</tr>
<tr>
<td>Reading Comprehension\textsuperscript{b}</td>
<td>5</td>
<td>5 - 7</td>
</tr>
<tr>
<td>TWE</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Total Number of Items</td>
<td>35</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{a}Discrete items  
\textsuperscript{b}Testlet-type items

Item Data: TOEFL Listening Comprehension items and TSE items. The TOEFL Listening Comprehension section for the forms used in this study consisted of 20 discrete statement items, 15 discrete dialogue items, and 4 mini-talks, with 3 to 4 discrete items associated with each mini-talk. The discrete statement and dialogue items were dichotomously scored. The mini-talk items were grouped according to the specific mini-talk; all items associated with each mini-talk were considered to constitute a testlet-type item, and were polytomously scored.

The TSE consisted of a number of specific speaking tasks, all of which were rated on four-point scales with a variety of different scores created for each task. A total of nine tasks or items were included in the concurrent calibrations performed. These nine tasks constituted Sections 4-7 of the pre-July 1995 TSE forms used in this study. The descriptions of these sections from the Bulletin of Information for TSE are contained in Appendix A of this paper. These nine tasks and the TOEFL items used in the two combined test forms in this study (referred to as TOEFL/TSE Forms X and Y)\textsuperscript{1} are described in Table 2.

\textsuperscript{1}Forms X and Y were administered in November 1994 and February 1995, respectively.
Table 2

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Number of Items</th>
<th>Number of Response Categories for Each Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statements&lt;sup&gt;a&lt;/sup&gt;</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Dialogues&lt;sup&gt;a&lt;/sup&gt;</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>Mini-talks&lt;sup&gt;ac&lt;/sup&gt;</td>
<td>4</td>
<td>3-4</td>
</tr>
<tr>
<td>Story Telling Based on Pictures&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Answer Questions about a Picture&lt;sup&gt;b&lt;/sup&gt;</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Describe Things and Give Opinion&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Explain a Schedule or Notice&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total Number of Items</strong></td>
<td><strong>48</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Note:**

<sup>a</sup> Items from TOEFL Listening Comprehension Section.

<sup>b</sup> Items from TSE.

<sup>ac</sup> For the mini-talks items, there are three or four discrete items associated with a common passage or extended conversation, each of which is scored right or wrong in TOEFL. These items were combined into a testlet-type item, and the scoring method was similar to that used for the Reading Comprehension items described in the previous section.

For pre-July 1995 forms of the TSE test, four types of competency scores were created, namely, pronunciation, grammar, fluency, and overall comprehensibility. The overall comprehensibility score was the only score assigned to every item in the test. In the present study, only the overall comprehensibility score was used because it is a global rating of examinee proficiency in dealing with a variety of complex speech tasks and, as just mentioned, is scored for all items. There were two additional item types in the pre-July 1995 TSE test not listed in Table 2, the passage reading and sentence completion questions. Passage reading would appear to be a measure of reading and pronunciation alone rather than overall speaking ability. Sentence completions were ten discrete items that involved reading and speaking instead of listening and speaking. Therefore, these two item types were not included in the present study.

**Examinee data.** With the current TOEFL test, the IRT scales used in equating are defined on a within-section basis using data from the domestic examinee population because the domestic population is more homogeneous and stable than the foreign examinee population. It is very likely that any equatings carried out for TOEFL 2000 tests will also make use of domestic samples. Therefore, response data from 1,500 domestic examinees were randomly sampled from each of
the three most recent pre-July 1995 TOEFL administrations when the TWE test was also administered. The three sets of response data provided three replications.

Table 3 contains the score frequencies for each of the TOEFL reading comprehension sets from each of the three TOEFL/TWE forms when each set is scored as a testlet. Table 4 contains the score frequencies for each of the TWE essays given at each of the three combined administrations. (Recall that each essay is scored twice on a 0 to 6 scale and the individual scores averaged.)

Similarly, TOEFL examinee response data from two recent pre-July 1995 TOEFL administrations when TSE was also administered were used in the study. The reason for conducting two instead of three replications for the TOEFL/TSE concurrent calibration is that only two recent TOEFL/TSE administrations had close to 500 examinees, and this sample size is close to the minimum calibration sample size of 500 examinees for operational use recently recommended by Muraki and Bock (1993) for the PARSCALE program. In addition, because the TOEFL/TSE domestic examinee sample sizes were so small, it was decided to include all available response data (both foreign and domestic) when exploring the possibility of dichotomous and polytomous item concurrent calibration for the TOEFL/TSE combinations.

Table 5 contains the score frequencies for each of the TOEFL mini-talk item sets from each of the two TOEFL/TSE forms when each set is scored as a testlet. Table 6 contains the score frequencies for each of the nine TSE tasks given at each of the two administrations. The tasks are each scored by two raters, but the data in Table 6 is based on scores from a single rater only. (We were unable to calibrate both raters' sets of scores together; see the conclusion section for further discussion.)

The IRT Models and Their Parameter Interpretation

Three IRT models were used in the study: the generalized partial credit model and the graded response model were used to calibrate the polytomously scored items, and the 3PL model was used to calibrate the dichotomously scored items.
### Table 3

**Testlet Score Frequencies for Reading Comprehension Sets in TOEFL/TWE Forms A-C**

*(N=1500)*

#### TOEFL/TWE Form A

<table>
<thead>
<tr>
<th>Score</th>
<th>Passage 1</th>
<th>Passage 2</th>
<th>Passage 3</th>
<th>Passage 4</th>
<th>Passage 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13</td>
<td>25</td>
<td>29</td>
<td>47</td>
<td>103</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>102</td>
<td>91</td>
<td>140</td>
<td>237</td>
</tr>
<tr>
<td>2</td>
<td>110</td>
<td>238</td>
<td>176</td>
<td>179</td>
<td>288</td>
</tr>
<tr>
<td>3</td>
<td>227</td>
<td>330</td>
<td>263</td>
<td>250</td>
<td>327</td>
</tr>
<tr>
<td>4</td>
<td>389</td>
<td>456</td>
<td>262</td>
<td>249</td>
<td>313</td>
</tr>
<tr>
<td>5</td>
<td>711</td>
<td>349</td>
<td>252</td>
<td>240</td>
<td>232</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>251</td>
<td>244</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>176</td>
<td>151</td>
<td></td>
</tr>
</tbody>
</table>

#### TOEFL/TWE Form B

<table>
<thead>
<tr>
<th>Score</th>
<th>Passage 1</th>
<th>Passage 2</th>
<th>Passage 3</th>
<th>Passage 4</th>
<th>Passage 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>28</td>
<td>31</td>
<td>52</td>
<td>131</td>
</tr>
<tr>
<td>1</td>
<td>45</td>
<td>58</td>
<td>111</td>
<td>115</td>
<td>260</td>
</tr>
<tr>
<td>2</td>
<td>113</td>
<td>140</td>
<td>195</td>
<td>252</td>
<td>325</td>
</tr>
<tr>
<td>3</td>
<td>243</td>
<td>309</td>
<td>293</td>
<td>233</td>
<td>286</td>
</tr>
<tr>
<td>4</td>
<td>420</td>
<td>563</td>
<td>273</td>
<td>221</td>
<td>259</td>
</tr>
<tr>
<td>5</td>
<td>440</td>
<td>402</td>
<td>312</td>
<td>218</td>
<td>239</td>
</tr>
<tr>
<td>6</td>
<td>233</td>
<td></td>
<td>285</td>
<td>214</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>195</td>
</tr>
</tbody>
</table>

#### TOEFL/TWE Form C

<table>
<thead>
<tr>
<th>Score</th>
<th>Passage 1</th>
<th>Passage 2</th>
<th>Passage 3</th>
<th>Passage 4</th>
<th>Passage 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
<td>12</td>
<td>29</td>
<td>62</td>
<td>93</td>
</tr>
<tr>
<td>1</td>
<td>44</td>
<td>64</td>
<td>102</td>
<td>158</td>
<td>147</td>
</tr>
<tr>
<td>2</td>
<td>111</td>
<td>122</td>
<td>176</td>
<td>358</td>
<td>291</td>
</tr>
<tr>
<td>3</td>
<td>277</td>
<td>166</td>
<td>237</td>
<td>390</td>
<td>368</td>
</tr>
<tr>
<td>4</td>
<td>543</td>
<td>252</td>
<td>301</td>
<td>315</td>
<td>355</td>
</tr>
<tr>
<td>5</td>
<td>516</td>
<td>336</td>
<td>283</td>
<td>217</td>
<td>246</td>
</tr>
<tr>
<td>6</td>
<td>323</td>
<td>251</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>225</td>
<td>121</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4

TWE Score Frequencies
TOEFL/TWE Forms A-C
(N=1500)

<table>
<thead>
<tr>
<th>Score</th>
<th>Form A</th>
<th>Form B</th>
<th>Form C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>4</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>1.5</td>
<td>4</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>2.0</td>
<td>55</td>
<td>62</td>
<td>38</td>
</tr>
<tr>
<td>2.5</td>
<td>63</td>
<td>94</td>
<td>76</td>
</tr>
<tr>
<td>3.0</td>
<td>326</td>
<td>378</td>
<td>368</td>
</tr>
<tr>
<td>3.5</td>
<td>252</td>
<td>261</td>
<td>276</td>
</tr>
<tr>
<td>4.0</td>
<td>523</td>
<td>467</td>
<td>514</td>
</tr>
<tr>
<td>4.5</td>
<td>133</td>
<td>102</td>
<td>112</td>
</tr>
<tr>
<td>5.0</td>
<td>94</td>
<td>97</td>
<td>77</td>
</tr>
<tr>
<td>5.5</td>
<td>29</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>6.0</td>
<td>17</td>
<td>11</td>
<td>5</td>
</tr>
</tbody>
</table>
Table 5

Testlet Scores Frequencies for Mini-talk Sets in TOEFL/TSE Forms X and Y

TOEFL/TSE Form X
(N=502)

<table>
<thead>
<tr>
<th>Score</th>
<th>Passage 1</th>
<th>Passage 2</th>
<th>Passage 3</th>
<th>Passage 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>25</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>1</td>
<td>21</td>
<td>107</td>
<td>26</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>62</td>
<td>208</td>
<td>96</td>
<td>151</td>
</tr>
<tr>
<td>3</td>
<td>159</td>
<td>162</td>
<td>180</td>
<td>296</td>
</tr>
<tr>
<td>4</td>
<td>254</td>
<td></td>
<td>195</td>
<td></td>
</tr>
</tbody>
</table>

TOEFL/TSE Form Y
(N=434)

<table>
<thead>
<tr>
<th>Score</th>
<th>Passage 1</th>
<th>Passage 2</th>
<th>Passage 3</th>
<th>Passage 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1¹</td>
<td>0</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>6</td>
<td>50</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>61</td>
<td>20</td>
<td>71</td>
<td>58</td>
</tr>
<tr>
<td>3</td>
<td>360</td>
<td>73</td>
<td>154</td>
<td>136</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>335</td>
<td>147</td>
<td>208</td>
</tr>
</tbody>
</table>

¹ 0,1 scores combined in the calibration.
Table 6
TSE Task Score Frequencies
TOEFL/TSE Forms X and Y
One Rater Only

TOEFL/TSE Form X
(N=502)

<table>
<thead>
<tr>
<th>Score</th>
<th>Q4&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Q5-1&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Q5-2&lt;sup&gt;2&lt;/sup&gt;</th>
<th>Q5-3&lt;sup&gt;2&lt;/sup&gt;</th>
<th>Q5-4&lt;sup&gt;2&lt;/sup&gt;</th>
<th>Q6-1&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Q6-2</th>
<th>Q6-3</th>
<th>Q7&lt;sup&gt;2&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td>5</td>
<td>9</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>45</td>
<td>22</td>
<td>49</td>
<td>46</td>
<td>51</td>
<td>41</td>
<td>51</td>
<td>84</td>
<td>59</td>
</tr>
<tr>
<td>2</td>
<td>251</td>
<td>163</td>
<td>197</td>
<td>191</td>
<td>228</td>
<td>234</td>
<td>237</td>
<td>219</td>
<td>246</td>
</tr>
<tr>
<td>3</td>
<td>206</td>
<td>317</td>
<td>253</td>
<td>261</td>
<td>221</td>
<td>227</td>
<td>209</td>
<td>190</td>
<td>196</td>
</tr>
</tbody>
</table>

<sup>1</sup> No 0 scores.
<sup>2</sup> 0, 1 scores combined in the calibration.

TOEFL/TSE Form Y
(N=434)

<table>
<thead>
<tr>
<th>Score</th>
<th>Q4&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Q5-1&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Q5-2&lt;sup&gt;2&lt;/sup&gt;</th>
<th>Q5-3&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Q5-4&lt;sup&gt;2&lt;/sup&gt;</th>
<th>Q6-1&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Q6-2</th>
<th>Q6-3</th>
<th>Q7&lt;sup&gt;1&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>12</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td>10</td>
<td>35</td>
<td>22</td>
<td>33</td>
<td>55</td>
<td>31</td>
<td>42</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>206</td>
<td>67</td>
<td>170</td>
<td>161</td>
<td>147</td>
<td>172</td>
<td>171</td>
<td>179</td>
<td>185</td>
</tr>
<tr>
<td>3</td>
<td>198</td>
<td>357</td>
<td>225</td>
<td>251</td>
<td>253</td>
<td>207</td>
<td>220</td>
<td>203</td>
<td>224</td>
</tr>
</tbody>
</table>

<sup>1</sup> No 0 scores.
<sup>2</sup> 0, 1 scores combined in the calibration.
The generalized partial credit model. The following description is based on Muraki's development of the model (Muraki, 1992). Suppose an item \( j \) has \( m_j \) ordered score categories. Let \( k \) denote the score categories for item \( j \), \( k = 1, 2, ..., m_j \). The generalized partial credit model assumes that each of two adjacent score categories \( (k \) and \( k - 1) \) in a polytomously scored item can be viewed as dichotomous categories and therefore the likelihood of a person with ability \( \theta \) reaching the score category \( k \) rather than \( k - 1 \) can be described by a dichotomous 2PL IRT model:

\[
P_{jk | k-1}(\theta) = \frac{P_{jk}(\theta)}{P_{jk-1}(\theta) + P_{jk}(\theta)} = \frac{\exp[a_j(\theta-b_{jk})]}{1+\exp[a_j(\theta-b_{jk})]} \tag{1}
\]

The generalized partial credit model can then be generalized from the dichotomous IRT model to describe the probability of reaching a particular score category \( k \) from all the possible \( m_j \) categories for an examinee with ability \( \theta \), and is expressed as:

\[
P_{jk}(\theta) = \frac{\exp[\sum_{v=1}^{t} a_j(\theta-b_{jv})]}{\sum_{c=1}^{m_j} \exp[\sum_{v=1}^{c} a_j(\theta-b_{jv})]} \tag{2}
\]

where \( b_{j1} = 0 \).

In equation (2), \( \theta \) is the ability parameter, \( a_j \) is the item discrimination parameter, and \( b_{jv} \) is the category threshold parameter. The item discrimination parameter describes how well the item can distinguish between individuals with different levels of ability. Including an item discrimination parameter in the model is what distinguishes the generalized partial credit model from the partial credit model (Masters, 1982), which assumes that the item discrimination is a constant for all the items in a test. The difference between these two models is similar to the difference between the Rasch model and the 2PL IRT model in the dichotomous case. The category threshold parameter in equation (2) can be interpreted in the following fashion: for an examinee with ability equal to the value of the \( k \)th step parameter, the probability of reaching score category \( k \) and score category \( k - 1 \) will be equal. For an examinee with ability less than the value of the \( k \)th step parameter, the probability of reaching score category \( k \) will be less than the probability of reaching score category \( k - 1 \). For an examinee with ability greater than the value of the \( k \)th step parameter, the probability of reaching score category \( k \) will be greater than the probability of reaching score category \( k - 1 \). Furthermore, \( b_{jv} \) can be decomposed additively as \( b_{jv} = b_{j1} + d_j \), where \( b_{j1} \) is the item location parameter, which indicates the difficulty for item \( j \) (labeled as threshold in PARSSCALE output and in the plots for this paper). The parameter \( d_j \) is the item category parameter, which is interpreted as the relative difficulty of an item step in comparison to other steps within item \( j \).

The graded response model. The graded response model is an extension of Thurstone's (1928) method of forming successive intervals to the analysis of graded responses on educational tests. The following description is based on
Samejima's development and Muraki's discussion of the model (Muraki, 1990; Samejima, 1972). The model dichotomizes the response categories into two overall categories: greater or equal to score category k or less than score category k. The probability of an item response greater or equal to score category k can then be described by a dichotomous IRT model (the 2PL model). Based on this assumption, an examinee's probability of scoring in score category k is described by the difference in probabilities for the person having scored greater or equal to k and having scored greater or equal to k + 1, as described in the following form:

\[
P_{jk}(\theta) = P_{jk}^*(\theta) - P_{jk+1}^*(\theta)
\]

\[
= \frac{\exp[a_j(\theta-b_{jk})]}{1+\exp[a_j(\theta-b_{jk})]} - \frac{\exp[a_j(\theta-b_{k+1,j})]}{1+\exp[a_j(\theta-b_{k+1,j})]},
\]

where \( k = 1,2,\ldots,m_j \), \( P_{j1}^*(\theta) = 1.0 \), and \( P_{j,m_j+1}^* = 0.0 \). Similar to the generalized partial credit model, \( a_j \) is the item discrimination parameter. In (3), the \( b_{jk} \) is the category threshold parameter that distinguishes the probabilities of scoring less than score category k and greater than or equal to score category k.

The 3PL model. The 3PL model was used to calibrate the dichotomously scored items and is described as:

\[
P_j(\theta) = c_j + \frac{1-c_j}{1+\exp[-1.7a_j(\theta-b_j)]},
\]

where \( a_j \) is the item discrimination parameter, \( b_j \) is the item difficulty parameter, and \( c_j \) is the guessing parameter (Lord, 1980).

Calibration Procedures

Check of the Unidimensionality Assumption. A principal component analysis of the Pearson item/testlet correlation matrix was used in this study to assess unidimensionality.\(^1\) This method was used by Huyhn and Ferrara (1994) to assess the dimensionality of MSPAP data. The same rule as was used in Huyhn and Ferrara's study, that the first factor should account for at least 20% of the test variance, was used in this study. The advantage of the principal component analysis procedure is that it does not require a large sample size. For example, for a 40 to 50 item test, data from 200 examinees would be sufficient (Stevens, 1986).

The principal component analysis used in this study is in many ways related to linear factor analytic procedures, although their purposes are somewhat different. McDonald (1981, 1982) discussed issues involved in attempting to apply linear factor analysis to dichotomously scored item-level data, where the relationship between response and underlying trait is nonlinear. The same sorts

\(^1\)Since the focus of this study was on fit of the various IRT models to the data, a quick and somewhat less rigorous approach to assessing dimensionality was employed. See Schedl, Gordon, Carey, and Tang (1996) for more formal approaches that might have been employed.
of issues exist for principal component analysis. In an attempt to circumvent
the nonlinearity problem at the item level when applying linear factor analysis,
Cook, Dorans, & Eignor (1988) formed item parcels or mini-tests of five or
more items, based on the work of Cattell (1974), and applied linear factor
analytic procedures to the parcel data. A similar procedure should work
effectively for principal component analysis. The testlet data formed in this
study is akin to the formation of parcels done by Cook et al. In addition to
using testlet data in the principal component analysis, data from the sets of
discrete items (i.e., the vocabulary items in the Vocabulary and Reading
Comprehension section and the statements and the dialogues in the Listening
Comprehension section) were each combined into a number of parcel scores for the
principal component analyses performed. The 29 discrete vocabulary items in each
of the three Vocabulary and Reading Comprehension sections were divided into 5
parcels of approximately equivalent difficulty, with each parcel containing
either 5 or 6 items. In a similar fashion, the 20 discrete statements items in
each of the two Listening Comprehension sections were divided into 4 parcels of
approximately equivalent difficulty, with each parcel containing 5 items.
Finally, the 15 discrete dialogue items in each of the two Listening
Comprehension sections were divided into 3 parcels of approximately equivalent
difficulty, with each parcel containing 5 items.

Because, as just mentioned, the scoring method used in this study grouped
the potentially locally dependent items into testlets, which is one of the
methods of managing local item dependence recommended by Yen (1993), the local
independence assumption was assumed to have been met and was not checked by
statistical methods.

Calibration Program. The PARSACLE PC computer program (Muraki & Bock, 1993) was
used for the item calibrations. This program calibrates polytomous items using
either the graded response model or the generalized partial credit model. The
program also concurrently calibrates both dichotomously scored and polytomously
scored items by assigning items to different "blocks." Items in a common block
have the same number of response categories.

The PARSACLE program uses a marginal maximum likelihood (MML) estimation
method with an expectation and maximization algorithm (MML-EM) to estimate item
parameters. The MML-EM algorithm consists of two steps: the expectation step (E
step) and the maximization step (M step). In the E step, the expected value of
the marginal likelihood function is computed, and in the M step, the marginal
likelihood function is maximized. The EM algorithm is applied in PARSACLE to
estimate the parameters item by item, and the algorithm continues through
multiple iterations of the EM process until the estimates for all items become
stable. Because the item parameters are estimated item by item, polytomous items
with differing numbers of response categories or dichotomous and polytomous items
with differing numbers of parameters can be calibrated concurrently.

Evaluation of the Calibration Results. The effectiveness of fit statistics
proposed by Mislevy and Bock (1990) and provided by the PARSACLE program were
used in combination with other methods to evaluate the model-data fit for each
polytomously or dichotomously scored item. To compute the effectiveness of fit
statistic, the $\theta$ scale is divided into $H$ intervals. $H$ is the number of intervals
left after neighboring intervals are merged, if necessary, to avoid expected
values $N_h P_h(\theta_h)$ less than five. An $m \times H$ contingency table for each item is then
constructed, where $m$ is the number of response categories for a given item. Let
$r_{kk}$ be the observed response frequency in the $k$th response category and in the
$h$th $\theta$ interval, and $N_h$ be the marginal response frequency over the $m$ response
categories in the hth \( \theta \) interval. For each \( \theta \) interval, the mean of \( \theta \), \( \theta_h \), is also computed. The corresponding \( P_k(\theta_h) \), \( k = 1, \ldots, m \), is computed using the item parameter estimates for each polytomously or dichotomously scored item. The likelihood-ratio chi-square statistic for each item is then computed using the following formula:

\[
G^2 = 2 \sum_{h=1}^{H} \sum_{k=1}^{m} r_{hk} \ln \frac{r_{hk}}{N_h P_k(\theta_h)} .
\]

(5)

The degree of freedom is \( H \times (m - 1) \).

It should be noted that the likelihood-ratio statistic defined above is only asymptotically distributed as a chi-square. Hence, performing statistical tests of fit of single models to item data should, strictly speaking, not be done. Usually competing models are formulated, where one model can be subsumed within the other, and the difference in likelihood ratio statistics for the competing models is evaluated via statistical hypothesis testing. Because this study was purely exploratory in nature, competing models were not generated and differences in likelihood ratio statistics were not computed and subjected to hypothesis testing. However, because the individual degrees of freedom associated with the fit statistics varied across items, making the magnitudes of the statistics by themselves difficult to evaluate, the associated probability levels were also included in the tables, but only for comparison purposes.

The graphic methods used to evaluate model-data fit for the polytomous models in NAEP were also used to evaluate the model-data fit for each polytomously scored item. The plot presents the observed versus estimated probability of scoring in the kth score category for 33 equally spaced \( \theta \) points, where \(-3.0 < \theta < 3.0\). The estimated probabilities are obtained using the item parameters for the generalized partial credit or graded response models. To obtain the observed probabilities, examinees were divided into 33 groups according to their estimated abilities (\( \theta \)s). The observed proportion of examinees scoring in each score category for the 33 examinee groups was computed. The observed probabilities were then overlaid with the estimated probabilities on the plot for each item. Fit can be visually assessed by observing the proximity of the estimated probabilities to the observed probabilities. The closer the proximity of the two sets of probabilities, the better the fit of the particular model to the data.

For the dichotomously scored items, the graphic method that was used to evaluate the model-data fit was the item-ability regression plot procedure currently used for TOEFL (Kingston & Dorans, 1985). This procedure is similar to the procedure that was used for the polytomously scored items, except that only the observed and predicted proportions of examinees answering the item correctly are plotted.

Comparison of Models. Root mean squared differences between the estimated true score and the actual observed score for the sum of the dichotomous and each of the polytomous items were calculated. This was done primarily to expedite comparisons of fit across the polytomous items calibrated using both the generalized partial credit model and the graded response model, although the overall average across all the dichotomously scored items was also formed to facilitate comparisons.
The average root mean squared difference between the estimated true scores and the observed scores across all the dichotomously scored items is defined as:

$$RMSD = \left[ \frac{1}{N} \sum_{s=1}^{N} \sum_{j=1}^{n} \left( \frac{1}{n} \sum_{j=1}^{n} \hat{p}_j(\delta_s) - \frac{1}{n} \sum_{j=1}^{n} X_{js} \right)^2 \right]^{1/2},$$ (6)

where $n$ is the number of dichotomously scored items, $N$ is the number of examinees, $\Sigma_{s=1}^{n} X_{js}$ is the total observed score on the dichotomously scored items for a particular examinee $s$, and $\Sigma_{s=1}^{n} \hat{p}_j(\delta_s)$ is the estimated true score for that same examinee ($s$) across the dichotomously scored items.

For a particular polytomously scored item $h$, the root mean squared difference between the estimated true score and the observed score on that item is defined as:

$$RMSD = \left[ \frac{1}{N} \sum_{s=1}^{N} \left( \frac{1}{m_h-1} \sum_{k=1}^{m_h-1} k_h \hat{p}_k(\delta_s) - \frac{1}{m_h-1} \Sigma k_h \cdot X_{hs} \right)^2 \right]^{1/2}$$ (7)

where $N$ is the number of examinees, $k$ is the number of response categories ($k = 1, 2, \ldots, m$), $X_{hs}$ is the observed score for examinee ($s$), and $\Sigma k_h \hat{p}_k(\delta_s)$ is the estimated true score for that same examinee ($s$). This was calculated first for items calibrated using the generalized partial credit model and then for the same items calibrated using the graded response model. It should be noted that the probability of scoring 0 is not included in the estimation of the true score for the polytomous item so the computation method is consistent with the conventional true score definition for the dichotomous items.

The average root mean squared difference between the estimated true score and the observed total score for the entire test is defined as:

$$RMSD = \left[ \frac{1}{N} \sum_{s=1}^{N} \left( \sum_{j=1}^{n} \hat{p}_j(\delta_s) + \sum_{k=1}^{m_h-1} k_h \hat{p}_k(\delta_s) - \left( \sum_{j=1}^{n} X_{js} + \sum_{k=1}^{m_h} X_{hs} \right) \right)^2 \right]^{1/2}$$ (8)

where $L$ is the total number of polytomous items on the test, and the other variables are defined in the previous equations.

In addition to the root mean squared difference for each of the total tests (TOEFL/TWE Forms A-C and TOEFL/TSE Forms X and Y), two total test observed-estimated true score correlations were calculated and two bivariate plots of observed versus estimated true scores for all examinees taking each of the test forms were produced. For one correlation and plot, the true scores were estimated using the generalized partial credit model and for the others, the graded response model was used.
Results

Dimensionality

As mentioned previously, a principal component analysis of the Pearson item/testlet correlation matrix was carried out to determine if there was a dominant first component in the data from the three TOEFL/TWE forms and the two TOEFL/TSE forms used in this study.

Table 7 lists the percentage of total test variance that was accounted for by the five largest eigenvalues for each of the five TOEFL/TWE and TOEFL/TSE form combinations. Figure 1 depicts the percentage of total variance of these five largest eigenvalues for each of the three TOEFL/TWE forms; Figure 2 depicts comparable information for the two TOEFL/TSE forms.

The data in the table and figures provides strong evidence that each of the five combined forms used in this study is dominated by a major first component. In both Figures 1 and 2, there is a sharp drop in percent total test variance from the first to second eigenvalue, followed by a much more gradual change for the remaining four eigenvalues. Also notable is that the first component for the TOEFL/TWE forms appear to be slightly more dominant than the first component for the TOEFL/TSE forms. In all cases, however, the percentage of variance accounted for by the first component is greater than 40%.

Because of the unanticipated level of dominance of the first component for all forms studied, it was decided that for TOEFL/TWE Form A, the dichotomously scored Vocabulary items would be unbundled and treated as 29 discrete items rather than five parcels in the principal component analysis. While this violates the linearity assumption underlying the principal component procedure, results in this case can be compared to published results by Reckase (1979). With the unbundling, the percentage of variance accounted for by the first component dropped from 48% to 31%, but this result appears to be very much in line with Reckase's results for dichotomously scored items.

In summary, all five forms in this study appeared to be sufficiently unidimensional that the combination of the 3PL model and the generalized partial credit or graded response models could safely be used for item calibration purposes.

Effectiveness of Fit

Table 8 contains the chi-square fit statistics for the polytomous items from TOEFL/TWE Form A calibrated using both the generalized partial credit and graded response models, along with associated degrees of freedom and the probability level. Table 9 contains comparable data for the polytomous items in TOEFL/TSE Form X. Chi-square fit statistics were calibrated for all 105 items in the three TOEFL/TWE combined forms and all 96 items in the two TOEFL/TSE combined forms, once using the generalized partial credit/3PL combination and then using the graded response/3PL combination. Forms A and X were chosen because they are exemplary in each case of the complete set.
A decision was made to only report data for the polytomous items in these tables because the data for the dichotomous items closely paralleled that seen when these items were calibrated operationally for IRT equating purposes using the 3PL model.

Table 7

Percentage of Test Variance for the Five Principal Components with Largest Eigenvalues
TOEFL/TWE and TOEFL/TSE

<table>
<thead>
<tr>
<th>Form</th>
<th>Principal Component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>TOEFL/TWE Form A</td>
<td>48.2%</td>
</tr>
<tr>
<td>TOEFL/TWE Form B</td>
<td>50.6</td>
</tr>
<tr>
<td>TOEFL/TWE Form C</td>
<td>46.0</td>
</tr>
<tr>
<td>TOEFL/TSE Form X</td>
<td>44.4</td>
</tr>
<tr>
<td>TOEFL/TSE Form Y</td>
<td>40.9</td>
</tr>
</tbody>
</table>
Figure 1: Plots of eigenvalues for TOEFL/TWE Forms A-C
Figure 2: Plots of eigenvalues for TOEFL/TSE Forms X-Y
Table 8

Item Fit Statistics for the Polytomous Items
TOEFL/TWE Form A

<table>
<thead>
<tr>
<th>Item</th>
<th>Partial Credit Model</th>
<th></th>
<th></th>
<th>Graded Response Model</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Passage 1</td>
<td>48.61</td>
<td>59</td>
<td>0.83</td>
<td>55.86</td>
<td>57</td>
<td>0.52</td>
</tr>
<tr>
<td>Passage 2</td>
<td>94.51</td>
<td>82</td>
<td>0.16</td>
<td>78.88</td>
<td>81</td>
<td>0.55</td>
</tr>
<tr>
<td>Passage 3</td>
<td>138.94</td>
<td>99</td>
<td>&lt; 0.10</td>
<td>111.95</td>
<td>101</td>
<td>0.21</td>
</tr>
<tr>
<td>Passage 4</td>
<td>166.84</td>
<td>114</td>
<td>&lt; 0.10</td>
<td>171.32</td>
<td>115</td>
<td>&lt; 0.10</td>
</tr>
<tr>
<td>Passage 5</td>
<td>121.87</td>
<td>104</td>
<td>0.11</td>
<td>119.96</td>
<td>103</td>
<td>0.12</td>
</tr>
<tr>
<td>TWE</td>
<td>154.94</td>
<td>121</td>
<td>&lt; 0.10</td>
<td>154.97</td>
<td>118</td>
<td>&lt; 0.10</td>
</tr>
</tbody>
</table>
### Table 9

**Item Fit Statistics for the Polytomous Items**

**TOEFL/TSE X**

<table>
<thead>
<tr>
<th>Item</th>
<th>Partial Credit Model</th>
<th></th>
<th></th>
<th>Graded Response Model</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL Passage 1</td>
<td>30.11</td>
<td>25</td>
<td>0.22</td>
<td>37.04</td>
<td>22</td>
<td>&lt; 0.10</td>
</tr>
<tr>
<td>TOEFL Passage 2</td>
<td>43.94</td>
<td>40</td>
<td>0.31</td>
<td>42.14</td>
<td>37</td>
<td>0.26</td>
</tr>
<tr>
<td>TOEFL Passage 3</td>
<td>35.57</td>
<td>30</td>
<td>0.22</td>
<td>44.54</td>
<td>31</td>
<td>&lt; 0.10</td>
</tr>
<tr>
<td>TOEFL Passage 4</td>
<td>15.73</td>
<td>17</td>
<td>0.54</td>
<td>15.53</td>
<td>18</td>
<td>0.63</td>
</tr>
<tr>
<td>TSE Question 4</td>
<td>23.70</td>
<td>17</td>
<td>0.13</td>
<td>28.59</td>
<td>20</td>
<td>0.10</td>
</tr>
<tr>
<td>TSE Question 5-1</td>
<td>6.95</td>
<td>16</td>
<td>0.97</td>
<td>11.02</td>
<td>14</td>
<td>0.69</td>
</tr>
<tr>
<td>TSE Question 5-2</td>
<td>6.10</td>
<td>19</td>
<td>1.00</td>
<td>14.31</td>
<td>18</td>
<td>0.71</td>
</tr>
<tr>
<td>TSE Question 5-3</td>
<td>14.19</td>
<td>18</td>
<td>0.72</td>
<td>16.74</td>
<td>17</td>
<td>0.47</td>
</tr>
<tr>
<td>TSE Question 5-4</td>
<td>20.60</td>
<td>21</td>
<td>0.48</td>
<td>30.35</td>
<td>20</td>
<td>&lt; 0.10</td>
</tr>
<tr>
<td>TSE Question 6-1</td>
<td>26.26</td>
<td>16</td>
<td>&lt; 0.10</td>
<td>15.53</td>
<td>19</td>
<td>0.69</td>
</tr>
<tr>
<td>TSE Question 6-2</td>
<td>19.33</td>
<td>20</td>
<td>0.50</td>
<td>21.69</td>
<td>20</td>
<td>0.36</td>
</tr>
<tr>
<td>TSE Question 6-3</td>
<td>17.56</td>
<td>25</td>
<td>0.86</td>
<td>20.05</td>
<td>23</td>
<td>0.64</td>
</tr>
<tr>
<td>TSE Question 7</td>
<td>19.82</td>
<td>20</td>
<td>0.47</td>
<td>22.51</td>
<td>22</td>
<td>0.43</td>
</tr>
</tbody>
</table>
Looking at the chi-square fit statistics reported in Table 8, a number of the polytomously scored testlets and the TWE items in TOEFL/TWE Form A appear to not be particularly well fit by the polytomous IRT models. On the other hand, as can be seen from the chi-square fit statistics reported in Table 9, a much smaller number of the polytomously scored items on TOEFL/TSE Form X appear to be poorly fit by the models. One problem with chi-square fit statistics is their susceptibility to sample size (Hambleton & Swaminathan, 1985). Given a large enough sample size, fit of the particular model to the data will always be rejected. Since the sample sizes for TOEFL/TWE Form A were at least three times as large as the sample sizes for TOEFL/TSE Form X, a decision was made to make the plots for the items the primary mechanism for evaluating individual item fit, and not the fit statistics. These plots are discussed in the following paragraphs.

Two plots were produced for each of the 105 TOEFL/TWE items and each of the 96 TOEFL/TSE items used in the study. One set of plots was produced for the calibration, making use of the generalized partial credit/3PL combination, and another set produced from the calibration based on the graded response/3PL combination. For this paper, certain plots for TOEFL/TWE Form A and TOEFL/TSE Form X were chosen for discussion purposes. Plots for selected dichotomously scored items on TOEFL/TWE Form A and TOEFL/TSE Form X are contained in Appendix B. All plots for the polytomously scored items for these two forms are contained in Appendices C and D of the paper. Each item in Appendices C and D has two plots, one based on the generalized partial credit model and the other based on the graded response model.

The first figure in Appendix B contains item-ability regression plots for three dichotomously scored Vocabulary items from TOEFL/TWE Form A. There is a plot based on the 3PL calibration of each item when the generalized partial credit model was used with the polytomous items and another plot when the graded response model was used. There are 58 of these plots in total, each for each of the 29 Vocabulary items. Observed proportions (the boxes) and predicted proportions (the solid curve) correspond reasonably closely for all three items. Parameter estimates for the parameters of the 3PL model used also correspond closely when the items are calibrated along with either the generalized partial credit or the graded response models. The second figure in Appendix B contains item-ability regression plots exactly like those depicted in the first figure, but for three of the dichotomously scored items from TOEFL/TSE Form X. Again, fit is reasonably good and parameter estimates closely correspond.

The first figure in Appendix C contains the overall generalized partial credit model and graded response model plots and the fit plots for each response category for the first reading comprehension item set in TOEFL/TWE Form A. This set contains five items related to the same reading passage and was scored as a single testlet having six response categories (0 to 5 items correct). The fit plots for each response category, where the boxes are again observed proportions and the solid curves the predicted proportions, demonstrate good model-data fit, with fit the poorest for the middle score categories of 2 and 3. Fit is comparable for the generalized partial credit and graded response models, although the groupings by estimated ability are somewhat different.

The last figure in Appendix C contains the overall generalized partial credit and graded response model plots and fit plots for each response category for the TOEFL/TWE Form A TWE essay. There are 11 response categories, corresponding to average TWE scores from 0 to 6 in increments of .5. Noteworthy here is a difference between results for the generalized partial credit and
graded response models seen for this item, and for a number of the other polytomously scored items. In order to get the PARSCALE computer program to run with the graded response model, adjacent response categories having relatively small frequencies had to be collapsed. In the case of this TWE essay, score categories of 1.0 and 1.5 had to be collapsed. This was not the case for the generalized partial credit model. While having to collapse categories will no longer provide a completely accurate accounting of the data, some justification for doing this with the graded response model can be found in Andrich (1995). Andrich discusses the fact that, with the graded response model, adjacent categories may be combined and the probability of response in the new combined category is the sum of the probabilities of the responses in the original categories. This means that fit of the model to the data will not be altered if categories are combined. (Andrich goes on to point out that the combining of categories will perturb fit when using the partial credit model.) Finally, for both models, looking at the fit plots for each response category, reasonable model-data fit can be observed, but as was the case for the reading comprehension testlet previously discussed, fit is poorest for the middle score categories, this time 3, 3.5, and 4.

The first figure in Appendix D contains the overall generalized partial credit and graded response plots and the fit plots for each response category for the first mini-talk item set in TOEFL/TSE Form X. This set contains four items related to the same mini-talk passage and was scored as a single testlet having five response categories (0 to 4 items correct). As was the case for the TWE essay previously discussed, response categories had to be collapsed in order to fit the graded response model; in this case, scores of 0 and 1 were collapsed. The fit plots for each response category for each model demonstrate reasonable data fit. The calibration sample size was small (N = 502) for TOEFL/TSE Form X, however, and this clearly has a noticeable effect on model-data fit.

The last figure in Appendix D contains the overall generalized partial credit and graded response model plots and fit plots for each response category for TOEFL/TSE Form X TSE Question 7. (Question 7 is the question for Section 7 of the TSE; see Appendix A for a general description of Section 7.) As mentioned earlier, these plots are based on only the scores from a single rater, and in this case, scores from 0 to 3 were assigned by the rater (scores of 0 and 1 were combined for calibration purposes). For both models, the fit plots for each response category demonstrate good model-data fit even with the small calibration sample size. This finding was also true for the other TOEFL/TSE Form X TSE questions.

To summarize, model-data fit in the plots for the dichotomously scored items in the TOEFL/TWE and TOEFL/TSE forms was much like the level of model-data fit observed in plots for operational calibrations of TOEFL forms for score equating purposes. Model-data fit for the polytomously scored items, calibrated using both the generalized partial credit and graded response IRT models, varied somewhat depending on the specific sort of item studied. For the TOEFL/TWE Form A combination, fit from both models was reasonably good, although for middle response categories, fit was not as good as for the response categories at the upper and lower ends. In addition, for certain of the polytomous items on TOEFL/TWE Form A, adjacent response categories having small frequencies had to be collapsed when using the graded response model. For TOEFL/TSE Form X, fit was not as good for the testlet-based TOEFL mini-talks as it was for the TSE tasks.
Again, in some cases, adjacent response categories had to be collapsed to fit the graded response model. However, given the small calibration sample size, the fit of the generalized partial credit model and the graded response model to the TSE task data was remarkably good.

**Comparison of Models**

Table 10 contains the average root mean squared differences (RMSDs) between the estimated true scores and the observed scores for the sum of the dichotomous items, each polytomous item calibrated using either the generalized partial credit or graded response models, and all items collectively, for the three combined forms of TOEFL/TWE. Also contained in Table 10 are the observed score/estimated true score correlations, which are based on all items collectively. Table 11 contains comparable information for the two combined forms of TOEFL/TSE used in the study.

On average, the data in Table 10 and 11 indicate somewhat better performance of the generalized partial credit model over the graded response model, that is, RMSDs are generally smaller and correlations are generally higher for the generalized partial credit model. In particular, for a number of the TOEFL/TWE polytomous items where exactly the same number of response categories were used in both calibrations, RMSDs were clearly smaller for the generalized partial credit model.

Figure 3 contains bivariate plots of observed versus estimated true scores for the three combined forms of TOEFL/TWE, with plots for the generalized partial credit/3PL combination on the left and graded response/3PL on the right. Figure 4 contains comparable data for the two combined forms of TOEFL/TSE. Noteworthy here is the fact that there are somewhat more outlying points with the graded response model than the generalized partial credit model. Outside of this observation, plots are consistent across models for each combined form.
Table 10

Average Root Mean Squared Difference (RMSD) between the Estimated True Scores and the Observed Scores and the Correlations between the Estimated True Scores and the Observed Scores
TOEFL/TWE

<table>
<thead>
<tr>
<th></th>
<th>Form A</th>
<th></th>
<th>Form B</th>
<th></th>
<th>Form C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Categories in Calibration</td>
<td>RMSD</td>
<td></td>
<td># Categories in Calibration</td>
<td>RMSD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Partial</td>
<td>Graded</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dichotomous</td>
<td>2</td>
<td>0.0676</td>
<td>0.0666</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage 1</td>
<td>6</td>
<td>0.1699</td>
<td>0.1729</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage 2</td>
<td>6</td>
<td>0.1983</td>
<td>0.1992</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage 3</td>
<td>8</td>
<td>0.1583</td>
<td>0.1605</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage 4</td>
<td>8</td>
<td>0.1869</td>
<td>0.1894</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage 5</td>
<td>6</td>
<td>0.2449</td>
<td>0.2460</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TWE</td>
<td>11/10</td>
<td>0.1279</td>
<td>0.1421</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Items</td>
<td></td>
<td>1.7287</td>
<td>1.9152</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td></td>
<td>0.9908</td>
<td>0.9886</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0670</td>
<td>0.0652</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1588</td>
<td>0.1608</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1645</td>
<td>0.1678</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1827</td>
<td>0.1850</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2027</td>
<td>0.2051</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2353</td>
<td>0.2371</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1314</td>
<td>0.1456</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.7541</td>
<td>1.9681</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.9915</td>
<td>0.9893</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0622</td>
<td>0.0639</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1582</td>
<td>0.1602</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1896</td>
<td>0.1900</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2143</td>
<td>0.2151</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2397</td>
<td>0.2396</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1158</td>
<td>0.1603</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.7460</td>
<td>1.8544</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.9902</td>
<td>0.9889</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1Score categories 1.0 and 1.5 were collapsed for the Graded Response Model.
2Score categories 1.0, 1.5, and 2.0 were collapsed for the Graded Response Model. Score categories 5.5 and 6.0 were also collapsed for the Graded Response Model.
Table 11

Average Root Mean Squared Difference (RMSD) between the Estimated True Scores and the Observed Scores and the Correlations between the Estimated True Scores and the Observed Scores
TOEFL/TSE

<table>
<thead>
<tr>
<th></th>
<th>Form X</th>
<th></th>
<th>Form Y</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Categories in Calibration</td>
<td>RMSD</td>
<td># Categories in Calibration</td>
<td>RMSD</td>
</tr>
<tr>
<td></td>
<td>Partial</td>
<td>Graded</td>
<td>Partial</td>
<td>Graded</td>
</tr>
<tr>
<td>Dichotomous</td>
<td>2</td>
<td>2</td>
<td>0.0807</td>
<td>0.0833</td>
</tr>
<tr>
<td>Passage 1</td>
<td>5</td>
<td>4'</td>
<td>0.1949</td>
<td>0.2534</td>
</tr>
<tr>
<td>Passage 2</td>
<td>4</td>
<td>4</td>
<td>0.2641</td>
<td>0.2653</td>
</tr>
<tr>
<td>Passage 3</td>
<td>5</td>
<td>4'</td>
<td>0.1905</td>
<td>0.2489</td>
</tr>
<tr>
<td>Passage 4</td>
<td>4</td>
<td>4</td>
<td>0.2262</td>
<td>0.2278</td>
</tr>
<tr>
<td>Q4</td>
<td>3</td>
<td>3</td>
<td>0.1964</td>
<td>0.1941</td>
</tr>
<tr>
<td>Q5-1</td>
<td>3</td>
<td>3</td>
<td>0.2401</td>
<td>0.2400</td>
</tr>
<tr>
<td>Q5-2</td>
<td>3'</td>
<td>3'</td>
<td>0.2406</td>
<td>0.2387</td>
</tr>
<tr>
<td>Q5-3</td>
<td>3'</td>
<td>3'</td>
<td>0.2370</td>
<td>0.2350</td>
</tr>
<tr>
<td>Q5-4</td>
<td>3'</td>
<td>3'</td>
<td>0.2353</td>
<td>0.2342</td>
</tr>
<tr>
<td>Q6-1</td>
<td>3</td>
<td>3</td>
<td>0.2069</td>
<td>0.2039</td>
</tr>
<tr>
<td>Q6-2</td>
<td>4</td>
<td>3'</td>
<td>0.1417</td>
<td>0.1933</td>
</tr>
<tr>
<td>Q6-3</td>
<td>4</td>
<td>4</td>
<td>0.1588</td>
<td>0.1570</td>
</tr>
<tr>
<td>Q7</td>
<td>3</td>
<td>3'</td>
<td>0.2046</td>
<td>0.2022</td>
</tr>
<tr>
<td>All Items</td>
<td></td>
<td></td>
<td>2.1032</td>
<td>2.4146</td>
</tr>
</tbody>
</table>

Correlation

0.9834 0.9791

0.9851 0.9757

1Score categories 0 and 1 were collapsed for the Graded Response Model.
2Score categories 0 and 1 were collapsed for the Generalized Partial Credit Model and the Graded Response Model.
Figure 3
Bivariate Plots of Observed vs Estimated True Scores
TOEFL/TWE

Form A

Generalized Partial Credit Model

Graded Response Model

Form B

Generalized Partial Credit Model

Graded Response Model
Figure 3 (con't.)

Bivariate Plots of Observed vs Estimated True Scores  
TOEFL/TWE

Form C

Generalized Partial Credit Model

Graded Response Model
Figure 4

Bivariate Plots of Observed vs. Estimated True Scores
TOEFL/TSE

Form X

Generalized Partial Credit Model

Graded Response Model

Form Y

Generalized Partial Credit Model

Graded Response Model
Discussion and Further Exploration

Based on the results of this study, it is clear that a combination of the 3PL and generalized partial credit models or 3PL and graded response models could be applied to data from a TOEFL 2000 test where its level of integration is comparable to the levels of integration demonstrated by the forms in this study. That is, data from a reading-writing combination made up of the TOEFL Vocabulary and Reading Comprehension section and the TWE were reasonably well fit by a unidimensional IRT model with items characterized by a combination of the 3PL and generalized partial credit models or 3PL and graded response models. In a similar fashion, data for a listening/speaking combination made up of the TOEFL Listening Comprehension section and selected tasks from the TSE were also reasonably well fit by a unidimensional IRT model and the 3PL/generalized partial credit and 3PL/graded response model combinations. Sample sizes for the TOEFL/TSE combination were extremely small and undoubtedly fit (as assessed by the plot procedure) would have been better had the calibration sample sizes been somewhat larger.

A variety of comparisons across the generalized partial credit and graded response models seem to indicate some preference for using the generalized partial credit model if PARSCALE is used as the calibration program. While differences were small across the two models, it was clear that more response categories could be retained when using the generalized partial credit model than with the graded response model. Hence, the generalized partial credit model is better able to capture the exact patterns in the data. Whether this collapsing of response categories for the graded response model would be necessary if another computer program such as MULTILG (Thissen, 1988) were used for calibration purposes remains an open question.

Finally, to date, we have been able to get the PARSCALE computer program to run with the TOEFL/TSE data only if we submit a single rater's scores on the TSE tasks. That is, the program has run with only one rater's scores, but not the combination of the two. (Raters' scores on individual tasks are not averaged in TSE. Only the overall score is averaged.) One possible explanation for the failure of PARSCALE to run on the combined rater data is that there may be too many parameters to be estimated (the number of score categories will be almost doubled) in relation to the small calibration sample size. Further investigation into this problem is needed before PARSCALE could be routinely used to calibrate TSE tasks.
References


APPENDIX A

Description of Pre-July 1995 TSE Sections 4-7 from Bulletin of Information for TSE
Section 4

In this section, you will see a series of pictures that tells a continuous story. You will be asked to tell the story that the pictures show.

During the actual test, you will be told how to begin your story. You will have one minute to tell the story. Speak as accurately and in as much detail as you can.

Now study pictures 1-6 below for exactly one minute and then tell the story in exactly one minute.
Section 5

In this section, you will be asked four questions about a picture. There are many different ways each question can be answered correctly. You will have 30 seconds to study the picture silently before you hear the questions.

First, study the picture silently for 30 seconds. Then, answer each of the following sample questions. Be sure to say as much as you can in the time allowed for each question. You will be given approximately 12 seconds to answer each question.

1. Where is this scene taking place?
2. What has just happened?
3. What will the boy probably do after this?
4. How could this situation have been avoided?

In the actual test, the questions in Section Five will NOT be printed in the test booklet.

SECTION FIVE PICTURE
Section 6

In this section you will be asked to describe certain things and to give your opinions on topics of general interest. Be sure to say as much as you can in responding to each question.

Remember, this is simply a test of spoken English. When it is scored, the graders will be interested in how you express your ideas, not the actual ideas. You will be given 15 seconds to prepare your answer to each question and approximately 45 seconds to answer the question.

1. Describe the things that make a perfect day.

2. Describe a telephone in detail.

3. The number of automobiles being manufactured in the world increases yearly. As a consequence, air pollution has also increased. How do you think the problem of automobile pollution should be handled?

The actual questions in Section Six will NOT be printed in the test book.
Section 7

In this section of the test, you will see a schedule or a notice describing activities of a club, conference, class, etc. For example, the club schedule below describes the club's activities. You will be asked to explain the schedule or notice. Be sure to include all important details in your description.

For the sample below, imagine that you are the club president talking to the club members. First, study the schedule silently; allow yourself only one minute.

Remember to include all important details in your description of the schedule to the club. In your presentation, do not just read the information printed, but present it as if you were talking to a group of people. Now, begin your description of the club schedule. Allow yourself exactly one minute.

MOUNTAINVILLE NATURE CLUB

Quarterly Meeting: April 15, 7:30 p.m.
Mountainville Nature Club
58 Fairview Drive

Speakers: Professor Alice Welton, Biology Department,
State University
"The American Bald Eagle: An Endangered Species"
Mr. Kenneth Shelby, Author, Focus on Nature,
"Photographing Animals Close Up"

Future Travel Plans: Florida Everglades
October 3-5
Total Cost: $500
Contact: Peter Jenkins, Tour Director
APPENDIX B

Selected 3PL Item-Ability Regression Plots

TOEFL/TWE Form A

TOEFL/TSE Form X

-38-
3-PL Item-Ability Regression Plots for Three Dichotomously-Scored Vocabulary Items in TOEFL/TWE Form A When Items were Calibrated with the Generalized Partial Credit and Graded Response Models

Generalized Partial Credit Model

\[ A = 1.26 \quad B = -1.80 \quad C = 0.19 \]

Graded Response Model

\[ A = 1.22 \quad B = -1.83 \quad C = 0.19 \]

\[ A = 0.68 \quad B = -0.86 \quad C = 0.33 \]

\[ A = 0.67 \quad B = -0.88 \quad C = 0.33 \]
Generalized Partial Credit Model
A=1.02  B=-1.34  C=0.18

Graded Response Model
A=1.00  B=-1.36  C=0.18
3-PL Item-Ability Regression Plots for Three
Dichotomously-Scored Items in TOEFL/TSE Form X
When Items were Calibrated with the Generalized Partial Credit
and Graded Response Models

Generalized Partial Credit Model
$A=1.47 \quad B=-1.96 \quad C=0.19$

Graded Response Model
$A=1.31 \quad B=-2.07 \quad C=0.19$

$A=0.79 \quad B=0.21 \quad C=0.16$

$A=0.74 \quad B=0.28 \quad C=0.16$
Generalized Partial Credit Model
A=0.65 B=-1.70 C=0.19

Graded Response Model
A=0.61 B=-1.78 C=0.19
Appendix C
Overall Generalized Partial Credit Model and Graded Response Model Plots and Fit Plots for Each Response Category

TOEFL/TWE Form A
Reading Comprehension Passages 1-5
TWE Essay
TOEFL/TWE Form A
Passage 1

Generalized Partial Credit Model

slope = 0.564, threshold = -1.590,
b_1 = 0.000, b_2 = -3.000, b_3 = -2.047, b_4 = -1.568, b_5 = -0.913, b_6 = -0.425

Graded Response Model

slope = 0.964, threshold = -1.735
b_1 = -3.616, b_2 = -2.540, b_3 = -1.733, b = -0.883, b_5 = 0.098
TOEFL/TWE Form A
Passage 1

Generalized Partial Credit Model

SCORE=0

SCORE=1

SCORE=2

Graded Response Model

SCORE=0

SCORE=1

SCORE=2

-45-
Generalized Partial Credit Model

slope = 0.455, threshold = -1.096,
b₁ = 0.000, b₂ = -3.071, b₃ = -1.959, b₄ = -0.875, b₅ = -0.427, b₆ = 0.851

Graded Response Model

slope = 0.843, threshold = -1.133
b₁ = -3.460, b₂ = -2.140, b₃ = -1.058, b₄ = -0.139, b₅ = 1.129
Generalized Partial Credit Model

slope = 0.595, threshold = -0.528, 
b_1 = 0.000, b_2 = -2.560, b_3 = -1.702, b_4 = -1.069, 
b_5 = -0.300, b_6 = 0.106 b_7 = 0.485, b_8 = 1.346

Graded Response Model

slope = 1.278, threshold = -0.511 
b_1 = -2.611, b_2 = -1.787, b_3 = -1.117, b_4 = -0.451, b_5 = 0.124, 
b_6 = 0.711, b_7 = 1.556
TOEFL/TWE Form A
Passage 3

Generalized Partial Credit Model

```
SCORE=0
```

```
SCORE=1
```

```
SCORE=2
```

Graded Response Model

```
SCORE=0
```

```
SCORE=1
```

```
SCORE=2
```

-51-
TOEFL/TWE Form A
Passage 3

Generalized Partial Credit Model

Graded Response Model

SCORE=3

SCORE=3

SCORE=4

SCORE=4

SCORE=5

SCORE=5
TOEFL/TWE Form A
Passage 3

Generalized Partial Credit Model

SCORE=6

Graded Response Model

SCORE=6

SCORE=7

SCORE=7
Generalized Partial Credit Model

slope = 0.428, threshold = -0.395,
b_1 = 0.000, b_2 = -2.685, b_3 = -1.193, b_4 = -0.988,
b_5 = -0.198, b_6 = 0.186 b_7 = 0.490, b_8 = 1.620

Graded Response Model

slope = 1.031, threshold = -0.383
b_1 = -2.635, b_2 = -1.631, b_3 = -1.006, b_4 = -0.360
b_5 = 0.236, b_6 = 0.877, b_7 = 1.834
TOEFL/TWE Form A
Passage 4

Generalized Partial Credit Model

SCORE=3

SCORE=4

SCORE=5

Graded Response Model

SCORE=3

SCORE=4

SCORE=5
Generalized Partial Credit Model

slope = 0.316, threshold = -0.413,
b_1 = 0.000, b_2 = -2.357, b_3 = -0.841, b_4 = -0.373, b_5 = 0.310, b_6 = 1.195

Graded Response Model

slope = 0.707, threshold = -0.401
b_1 = -2.626, b_2 = -1.318, b_3 = -0.392, b_4 = 0.564, b_5 = 1.766
TOEFL/TWE Form A
Passage 5

Generalized Partial Credit Model

Graded Response Model

SCORE=0

SCORE=0

SCORE=1

SCORE=1

SCORE=2

SCORE=2
TOEFL/TWE Form A
Passage 5

Generalized Partial Credit Model

Graded Response Model

SCORE=3

SCORE=4

SCORE=5

SCORE=3

SCORE=4

SCORE=5
Generalized Partial Credit Model

slope = 0.307, threshold = -0.360,
b_1 = 0.000, b_2 = -1.538, b_3 = -6.294, b_4 = -1.231,
b_5 = -3.813, b_6 = 0.160, b_7 = -1.394, b_8 = 2.998
b_9 = 1.431, b_{10} = 3.442, b_{11} = 2.635

Graded Response Model

slope = 0.724, threshold = 0.017
b_1 = -4.785, b_2 = -3.028, b_3 = -2.386, b_4 = -0.912, b_5 = -0.176
b_6 = 1.502, b_7 = 2.274, b_8 = 3.384, b_9 = 4.284
TCEFL/TWE Form A
TWE Essay

Generalized Partial Credit Model

SCORE=4.0

Grade d Response Model

SCORE=4.0

SCORE=4.5

SCORE=4.5

SCORE=5.0

SCORE=5.0
TOEFL/TWE Form A
TWE Essay

Generalized Partial Credit Model
SCORE=5.5

Graded Response Model
SCORE=5.5

SCORE=6.0

SCORE=6.0
Appendix D

Overall Generalized Partial Credit Model and Graded Response Model Plots and Fit Plots for Each Response Category

TOEFL/TSE Form X
TOEFL Minitalk Passages 1-4
TSE Tasks for Sections 4-7
Generalized Partial Credit Model

slope = 0.431, threshold = -1.866

$b_1 = 0.000, b_2 = -2.890, b_3 = -2.315, b_4 = -1.719, b_5 = -0.538$

Graded Response Model

slope = 0.605, threshold = -1.605

$b_1 = -3.165, b_2 = -1.702, b_3 = 0.051$
TOEFL/TSE Form X
TOEFL Passage 2

Generalized Partial Credit Model

slope = 0.297, threshold = -1.434
$b_1 = 0.000, b_2 = -3.472, b_3 = -1.544, b_4 = 0.715$

Graded Response Model

slope = 0.396, threshold = -1.639
$b_1 = -4.628, b_2 = -1.609, b_3 = 1.320$
TOEFL/TSE Form X
TOEFL Passage 2

Generalized Partial Credit Model

SCORE=0

SCORE=1

SCORE=2

Graded Response Model

SCORE=0

SCORE=1

SCORE=2
TOEFL/TSE Form X
TOEFL Passage 2

Generalized Partial Credit Model

SCORE=3

Graded Response Model

SCORE=3
Generalized Partial Credit Model

slope = 0.555, threshold = -1.535
\( b_1 = 0.000, b_2 = -3.000, b_3 = -2.241, b_4 = -1.058, b_5 = 0.159 \)

Graded Response Model

slope = 0.751, threshold = -1.021
\( b_1 = -2.566, b_2 = -1.039, b_3 = 0.542 \)
TOEFL/TSE Form X
TOEFL Passage 3

Generalized Partial Credit Model

SCORE=0

Graded Response Model

SCORE=0 or 1

SCORE=1

SCORE=2

SCORE=2
Generalized Partial Credit Model

slope = 0.486, threshold = -1.764
b_1 = 0.000, b_2 = -2.218, b_3 = -2.230, b_4 = -0.844

Graded Response Model

slope = 0.573, threshold = -2.234
b_1 = -3.936, b_2 = -2.401, b_3 = -0.364
Generalized Partial Credit Model

SCORE=0

SCORE=1

SCORE=2

Graded Response Model

SCORE=0

SCORE=1

SCORE=2
Generalized Partial Credit Model

SCORE=3

Graded Response Model

SCORE=3
TOEFL/TSE Form X
TSE Question 4

Generalized Partial Credit Model

slope = 1.687, threshold = -0.641
b_1 = 0.000, b_2 = -1.488, b_3 = 0.206

Graded Response Model

slope = 1.716, threshold = -0.603
b_1 = -1.493, b_2 = 0.286
TOEFL/TSER Form X
TSE Question 4

Generalized Partial Credit Model

SCORE=1

SCORE=2

SCORE=3

Graded Response Model

SCORE=1

SCORE=2

SCORE=3
Generalized Partial Credit Model

slope = 0.793, threshold = -1.496
\( b_1 = 0.000, \ b_2 = -2.365, \ b_3 = -0.628 \)

Graded Response Model

slope = 0.842, threshold = -1.578
\( b_1 = -2.662, \ b_2 = -0.493 \)
TOEFL/TSE Form X
TSE Question 5-1

Generalized Partial Credit Model

SCORE=1

SCORE=2

SCORE=3

Graded Response Model

SCORE=1

SCORE=2

SCORE=3
Generalized Partial Credit Model

slope = 1.154, threshold = -0.813  
$b_1 = 0.000, b_2 = -1.514, b_3 = -0.112$

Graded Response Model

slope = 1.256, threshold = -0.786  
$b_1 = -1.568, b_2 = -0.005$
TOEFL/TSE Form X
TSE Question 5-2

Generalized Partial Credit Model

SCORE=0 or 1

SCORE=2

SCORE=3

Graded Response Model

SCORE=0 or 1

SCORE=2

SCORE=3
Generalized Partial Credit Model

slope = 1.196, threshold = -0.838
b₁ = 0.000, b₂ = -1.519, b₃ = -0.156

Graded Response Model

slope = 1.281, threshold = -0.816
b₁ = -1.585, b₂ = -0.048
TOEFL/TSE Form X
TSE Question 5-3

Generalized Partial Credit Model

SCORE=0 or 1

SCORE=2

SCORE=3

Graded Response Model

SCORE=0 or 1

SCORE=2

SCORE=3
Generalized Partial Credit Model

slope = 1.150, threshold = -0.707
\( b_1 = 0.000, b_2 = -1.540, b_3 = 0.125 \)

Graded Response Model

slope = 1.215, threshold = -0.673
\( b_1 = -1.575, b_2 = 0.228 \)
TOEFL/TSE Form X
TSE Question 5-4

Generalized Partial Credit Model

SCORE=0 or 1

Score=2

Score=3

Graded Response Model

SCORE=0 or 1

Score=2

Score=3

-88-
Generalized Partial Credit Model

slope = 1.495, threshold = -0.754
b₁ = 0.000, b₂ = -1.586, b₃ = 0.077

Graded Response Model

slope = 1.550, threshold = -0.723
b₁ = -1.598, b₂ = 0.151
TOEFL/TSE Form X
TSE Question 6-1

Generalized Partial Credit Model

Score=1

![Graph 1](image1)

Score=2

![Graph 2](image2)

Score=3

![Graph 3](image3)

Graded Response Model

Score=1

![Graph 4](image4)

Score=2

![Graph 5](image5)

Score=3

![Graph 6](image6)
Generalized Partial Credit Model

slope = 1.643, threshold = -1.193  
b_1 = 0.000, b_2 = -2.386, b_3 = -1.378, b_4 = 0.184

Graded Response Model

slope = 1.841, threshold = -0.538  
b_1 = -1.334, b_2 = 0.258
Generalized Partial Credit Model

SCORE=0

SCORE=1

SCORE=2

Graded Response Model

SCORE=0 or 1

SCORE=2
Generalized Partial Credit Model

Graded Response Model

SCORE=3
Generalized Partial Credit Model

slope = 1.440, threshold = -1.008
b_1 = 0.000, b_2 = -2.260, b_3 = -1.067, b_4 = 0.302

Graded Response Model

slope = 1.587, threshold = -1.015
b_1 = -2.388, b_2 = -1.053, b_3 = 0.397
TOEFL/TSE Form X
TSE Question 6-3

Generalized Partial Credit Model

Graded Response Model

SCORE=0

SCORE=0

SCORE=1

SCORE=1

SCORE=2

SCORE=2

-95-
TOEFL/TSE Form X
TSE Question 6-3

Generalized Partial Credit Model

Graded Response Model

SCORE=3

SCORE=3
TOEFL/TSE Form X
TSE Question 7

Generalized Partial Credit Model

slope = 1.629, threshold = -0.530
b_1 = 0.000, b_2 = -1.329, b_3 = 0.296

Graded Response Model

slope = 1.675, threshold = -0.489
b_1 = -1.331, b_2 = 0.353
TOEFL/TSE Form X
TSE Question 7

Generalized Partial Credit Model

Graded Response Model

SCORE=0 or 1

SCORE=0 or 1

SCORE=2

SCORE=2

SCORE=3

SCORE=3

-98-