TOEFL®

Monograph Series

MS - 2
SEPTEMBER 1996

Polytomous Item Response Theory Models and Their Applications in Large-Scale Testing Programs: Review of Literature

K. Linda Tang

Educational Testing Service

Polytomous Item Response Theory (IRT) Models and Their Applications in Large-Scale Testing Programs: Review of Literature

K. Linda Tang

Educational Testing Service
Princeton, New Jersey
RM-96-8

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

Copyright © 1996 by Educational Testing Service. All rights reserved.

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

EDUCATIONAL TESTING SERVICE, ETS, the ETS logo, TOEFL, the TOEFL logo, and TSE are registered trademarks of Educational Testing Service.

Foreword

The TOEFL® Monograph Series features commissioned papers and reports for TOEFL 2000 and other Test of English as a Foreign Language program development efforts. As part of the foundation for the TOEFL 2000 project, a number of papers and reports were commissioned from experts within the fields of measurement and language teaching and testing. The resulting critical reviews and expert opinions were invited to inform TOEFL program development efforts with respect to test construct, test user needs, and test delivery. Opinions expressed in these papers are those of the authors and do not necessarily reflect the views or intentions of the TOEFL program.

These monographs are also of general scholarly interest, and the TOEFL program is pleased to make them available to colleagues in the fields of language teaching and testing and international student admissions in higher education.

The TOEFL 2000 project is a broad effort under which language testing at ETS will evolve into the 21st century. As a first step in the evolution of TOEFL language testing, the TOEFL program recently revised the Test of Spoken English (TSE®) test and announced plans to introduce a TOEFL computer-based test (TOEFL CBT) in 1998. The revised TSE, introduced in July 1995, is based on an underlying construct of communicative language ability and represents a process approach to test validation. The TOEFL CBT will take advantage of the new forms of assessments and improved services made possible by computer-based testing while also moving the program toward its longer-range goals, which include

- the development of a conceptual framework that takes into account models of communicative competence
- a research agenda that informs and supports this emerging framework
- a better understanding of the kinds of information test users need and want from the TOEFL test
- a better understanding of the technological capabilities for delivery of TOEFL tests into the next century

It is expected that the TOEFL 2000 efforts will continue to produce a set of improved language tests that recognize the dynamic, evolutionary nature of assessment practices and that promote responsiveness to test user needs. As future papers and projects are completed, monographs will continue to be released to the public in this new TOEFL research publication series.

TOEFL Program Office
Educational Testing Service

Table of Contents

A Brief Outline of This Review

This review discusses the following topics:

(1) Two commonly used polytomous IRT models and their assumptions

(2) The research results of comparing an equating procedure using a particular polytomous IRT model with conventional equipercentile equating

(3) The PARSCALE™ computer program, which allows the concurrent calibration of dichotomously and polytomously scored items

(4) A linking procedure that sets the parameter estimates obtained from different dichotomous and polytomous items concurrent calibrations to a common scale

(5) An application of the polytomous IRT model in a large-scale testing program: How does the NAEP conduct the concurrent calibrations for the dichotomously and polytomously scored items?

The Polytomous Models and Their Assumptions

Items that are scored in two categories — right or wrong — are referred to as dichotomously scored items. Items that are scored in multiple-ordered categories are referred to as polytomously scored items. Multiple-choice and short constructed-response items are dichotomously scored. The extended constructed-response items often used in performance assessments usually are polytomously scored. For the dichotomously scored items, the probability of a correct response for an examinee can be described by one of the logistic IRT models, most typically the three-parameter logistic (3PL) IRT model if the items are multiple choice. For the polytomously scored items, the probability of an examinee reaching a specific score category can be described by one of the polytomous IRT models, among which are the partial credit model (Masters, 1982) and its generalized version — the generalized partial credit model (Muraki, 1992) — and the graded response model (Samejima, 1969, 1972). These polytomous models are generalized from the dichotomous IRT models and reduced to the dichotomous IRT models when only two response categories exist. 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 categories is two.

The Generalized Partial Credit Model and Its Parameter Interpretation

Both the partial credit model and the generalized partial credit model assume that each of two adjacent categories (*k* and *k-1*) in a polytomously scored item can be viewed as dichotomous categories, and therefore the likelihood of a person with a certain ability level reaching the score category *k* rather than *k-1* can be described by a dichotomous IRT model. The models were thus generalized from the dichotomous IRT models to describe the probability of selecting a particular score category from all the possible score categories for an examinee.

The major difference between the partial credit model and the generalized partial credit model is that the partial credit model assumes that the item discrimination is a constant for all the items in a test. The generalized partial credit model, on the other hand, assumes that item discrimination can be different across items, and has a parameter to model this. The difference between these two models is similar to the difference between the Rasch model and the two parameter (2PL) IRT model in the dichotomous case. The items in a TOEFL 2000 test probably will have different discriminations (this is the case for the dichotomous items in the current TOEFL test). Therefore, the generalized partial credit model would seem to be the model of choice for TOEFL 2000 and thus will be discussed in detail.

Suppose a polytomously scored item has *m* score categories. Based on the generalized partial credit model, the item has one item discrimination parameter, one location parameter, and a set of *m-1* threshold parameters. The one location and the *m-1* threshold parameters can be combined into a set of *m-1* step parameters. The item discrimination parameter describes how well the item can distinguish between individuals with different levels of ability. The location parameter indicates the item difficulty. The threshold parameter is interpreted as the relative difficulty of an item step compared with other steps within an item. The step parameter can be interpreted this way:

- For an examinee with ability equal to the value of the *kth* 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 *kth* 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 *kth* step parameter, the probability of reaching score category *k* will be greater than the probability of reaching score category *k-1*.

The Graded Response Model and Its Parameter Interpretation

The graded response model is an extension of Thurstone's (1928) method of successive intervals to the analysis of graded responses on educational tests. The model dichotomizes the response categories into two overall categories: (1) greater or equal to score category *k;* and (2) 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). On the basis of this assumption, an examinee's probability of choosing a 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*. In other words, the partial credit model describes the probability of reaching score category *k* by the difference of two probabilities, each of which can be expressed through the use of a dichotomous IRT model.

Under the graded response model, each item has a discrimination parameter and a set of *m-1* threshold parameters. The discrimination parameter has the same interpretation as that in the generalized partial credit model. Each of the *m-1* threshold parameters distinguishes the probabilities of scoring less than score category *k* and greater than or equal to score category *k*.

The Unidimensionality Assumption

As with the dichotomous IRT models, unidimensionality (i.e., a single latent variable fully explains task performance) is one of the major assumptions for the polytomous IRT models. Several studies have shown that, even if the cognitive processes required to answer constructed-response items are inherently complex, data from these items can meet the unidimensionality assumption (Carlson, 1993; Huynh & Ferrara, 1994; Thissen, Wainer, & Wang, 1994).

Carlson (1993) studied the factor structure of 1992 NAEP mathematics and reading tests. These tests included both dichotomously and polytomously scored items. He concluded that the mathematics tests were essentially unidimensional. For the reading items, he found that mixing item types, such as multiple-choice and constructed-response, did not introduce multidimensionality in the resulting data.

Before applying the partial credit model to the data obtained from the 1991 field test for the Maryland School Performance Assessment Program (MSPAP), Huynh and Ferrara (1994) investigated whether the data met the unidimensionality assumption. All MSPAP tasks required brief or extended responses to performance tasks, which were designed to elicit students' ability to apply knowledge, skills, and thinking processes. They concluded that the responses to these polytomously scored performance assessment items were dominated by one major factor, or that the data were essentially unidimensional.

Thissen, Wainer, and Wang (1994) investigated dimensionality issues for the Computer Science and Chemistry tests of the College Board's Advanced Placement program; both tests include both multiple-choice items and free-response items. Restricted factor analyses showed that, for the most part, the free-response sections measured the same underlying proficiency as the multiple-choice sections. There was also a significant, but relatively small, amount of local dependence among the free-response items that produced a small degree of multidimensionality for each test. Thissen et al. (1994) concluded that the degree of multidimensionality in these tests appeared to be sufficiently small (no larger than that which also existed among the multiple-choice items) that it could be ignored for practical purposes.

The Local Independence Assumption

Huynh and Ferrara (1994) showed that multiple clusters or sets of items (i.e., items based on a common passage or bundled into a multistep mathematics problem) in a test were locally independent. However, the items within each cluster might show some level of dependency. Yen (1993) pointed out that performance assessments seem to produce more local item dependence (LID) than do the traditional multiple-choice tests. In the same paper, Yen has suggested some strategies for managing LID in order to avoid negative measurement consequences.

Comparing Test Score Equating Using a Polytomous IRT Model with Conventional Equipercentile Equating

Both IRT and equipercentile equating can model a raw score to raw score relationship that is nonlinear. The equipercentile procedure equates two test scores if they share the same percentile rank in the same or equivalent groups. IRT equating, on the other hand, equates the true scores on the two tests that correspond to the same ability level. The standard equipercentile equating procedure requires that the abilities of the two equating samples be equivalent (equivalent samples are difficult to obtain when the test populations are not homogenous). If the two equating samples are not equivalent, the examinees in the samples have to take a set of common items (anchor test). The equipercentile equating then can be conducted using scores on the anchor test. IRT equating, on the other hand, requires good model-data fit. The two IRT assumptions discussed in the previous sections also need to be met.

Huynh and Ferrara (1994) compared a partial credit model IRT equating with a conventional equipercentile equating using the MSPAP data described previously. They found that the two equating procedures appear to produce similar results when the examinations are of moderate difficulty for equating samples with typical score distributions. The two procedures might not give equivalent results, however, when the equating sample score distributions are markedly skewed.

The Dichotomous-Polytomous Item Concurrent Calibrations in NAEP

The NAEP Scales

The major assessment areas of the NAEP tests are reading, mathematics, and writing. For reading and mathematics, the scaling was carried out separately within subareas of the overall assessment area. For example, there are three scales corresponding to three specific purposes of reading: (1) Reading for Literary Experience; (2) Reading to Gain Information; and (3) Reading to Perform a Task. This scaling within subareas was done because it was anticipated that different patterns of performance might exist for these essential subdivisions of the overall subject area. By creating a separate scale for each of these subareas, potential differences in subpopulation performance between the content areas can be maintained. Because of the small number of items in writing, the writing section was not divided into subscales.

The creation of a series of separate scales to describe performance within a subject area does not preclude the reporting of a composite as a single index of overall performance in the subject area. The overall composite is computed as the weighted average of scores on the content area scales, where the weights correspond to the relative importance given to each content area.

The Concurrent Calibrations

The NAEP tests contain four types of items: (1) multiple-choice; (2) short constructed-response; (3) extended constructed-response; and (4) testlet-type items. The testlet-type items are sets of highly correlated parts where the score assigned to a testlet was the number of constituent parts answered correctly. Three distinct models were used in the analyses. A three-parameter logistic (3PL) model was used for the multiple-choice items. A two-parameter logistic (2PL) model was used for the short constructed-response items, which were scored correct or incorrect, and where the correct answer could not be guessed. A generalized partial credit model was used for the extended constructed-response and testlet-type items. For reading and mathematics, the possible scores for the polytomous items range from 0 to 4. For the writing assessment, the possible testlet scores range from 0 to 6.

Estimates of item parameters were obtained by the NAEP BILOG/PARSCALE computer program, which combines Mislevy and Bock's (1982) BILOG program for dichotomous items with Muraki and Bock's (1991) PARSCALE computer program, used with polytomous items. The program concurrently estimates parameters for all items (dichotomous and polytomous).

Calibration was performed in two stages. In the first stage, the ability distribution was constrained to be normally distributed. The values of the item parameters estimated on the basis of this normal ability distribution were then used as starting values for a second stage estimation run in which the parameters of the ability distribution were estimated concurrently with the item parameters.

Evaluations of the fit of the IRT models were carried out for each of the items. These evaluations were based primarily on graphical analyses. First, model fit was evaluated by examining plots of nonmodel-based estimates of the expected proportion correct (conditioned on ability level) versus the proportion correct predicted by the estimated item response function. For extended constructed-response items, similar plots were produced for each item category response function. For most items, the model fit was extremely good. Items that clearly did not fit the model were not included in the final scales.

Using a generalization of the Stocking-Lord transformation procedure (Muraki & Chang, 1994), parameter estimates from different calibrations were linked to a common scale. The final step was to transform ability estimates on this common scale to scores on the reporting proficiency metric, which has a mean of 250 and a variance of 50.

Conclusion

In conclusion, the literature shows that the polytomous IRT models have been successfully used in many large-scale performance-based tests. Two such programs — MSPAP and NAEP — have been discussed in this paper. In addition, several major issues, such as dichotomous and polytomous item concurrent calibrations and the linking of parameter estimates from such concurrent calibrations, have been successfully addressed. Further study is necessary to determine the applicability of this methodology to a TOEFL test made up of polytomous or dichotomous and polytomous items.

References

Baker, F. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16*, 87-96.

Carlson, J. E. (1993, April). *Dimensionality of NAEP instruments that incorporate polytomously-scored items*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Huynh, H., & Ferrara, S. (1994). A comparison of equal percentile and partial credit equating for performance-based assessments composed of free-response items. *Journal of Educational Measurement, 31*, 125-141.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter scaling of rating data*. [Computer program]. Chicago, IL: Scientific Software, Inc.

Muraki, E., & Bock, R. D. (1993). *PARSCALE: IRT based test scoring and item analysis*. [Computer program]. Chicago, IL: Scientific Software International.

Muraki, E., & Chang, H. (1994). *Horizontal and vertical test equating methods based on the generalized partial credit model*. Educational Testing Service internal report.

The NAEP 1992 Technical Report. (1993, December). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, No. 17.*

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph, No. 18.*

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*, 113-123.

50% RECYCLED PAPER
10% Post Consumer Waste

Cover Printed on Recycled Paper

58701-13953 • Y96M.5 • 253703 • Printed in U.S.A.