Research Report

Technology-Enhanced Items and Model–Data Misfit

ETS RR–22-11

Carol Eckerly
Yue Jia
Paul Jewsbury

December 2022

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Technology-Enhanced Items and Model–Data Misfit

Carol Eckerly, Yue Jia, & Paul Jewsbury

ETS, Princeton, NJ

Testing programs have explored the use of technology-enhanced items alongside traditional item types (e.g., multiple-choice and constructed-response items) as measurement evidence of latent constructs modeled with item response theory (IRT). In this report, we discuss considerations in applying IRT models to a particular type of adaptive testlet referred to as a branching item. Under the branching format, all test takers are assigned to a common question, and the assignment of the next question relies on the response to the first question through deterministic rules. In addition, the items at both stages are scored together as one polytomous item. Real and simulated examples are provided to discuss challenges in applying IRT models to branching items. We find that model – data misfit is likely to occur when branching items are scored as polytomous items and modeled with the generalized partial credit model and that the relationship between the discrimination of the routing component and the discriminations of the subsequent components seemed to drive the misfit. We conclude with lessons learned and provide suggested guidelines and considerations for operationalizing the use of branching items in future assessments.

Keywords Technology-enhanced items (TEIs); item response theory (IRT); branching item; branching format; model – data misfit, digital format; adaptive assessment; psychometric problem-solving; speededness; motivation; estimation of proficiencies; scaffolding; general partial credit model (GPCM)

doi:10.1002/ets2.12353

Many established testing programs that have traditionally operated in a paper-and-pencil format have recently transitioned or are in the process of transitioning to a digital format. Some examples include the National Assessment of Educational Progress, the *GRE*® test, and the Medical College Admission Test. Other tests that had been available only in paper-and-pencil format are now available in both paper-and-pencil and digital formats, including large-scale international assessments like the Program for International Student Assessment and the Program for the International Assessment of Adult Competencies. The transition to a digital format allows for novel item types only possible in a digital environment, defined as *technology-enhanced items* (TEIs). Whereas early computer-based tests largely continued to utilize traditional item types, recent advancements in commercial computer-based testing software have enabled testing programs to implement and pilot TEIs with greater ease (Parshall & Guille, 2016).

The focus of this report is the fixed branching adaptive TEI introduced by Wainer and Kiely (1987). These item types, which we will refer to as "branching items," consist of multiple components in which test takers are routed to different components based on responses to the routing component(s). Branching items have several desirable properties. First, though branching items are adaptive in nature, the routing paths are much more controlled than in fully adaptive assessment (i.e., a computerized adaptive test). This trait allows test developers to limit some potential psychometric problems with fully adaptive assessments, including differential context effects related to item location, cross-information, or unbalanced content (Wainer & Kiely, 1987). Furthermore, branching items retain some of the advantages of adaptive assessment, including tailoring assessment content to better match test takers' abilities (e.g., test takers of higher ability will be less likely to receive the easiest content, and vice versa). Thus, these items have the potential to help with speededness of timed tests, motivation of test takers of lower ability, and more precise estimation of proficiencies of test takers of lower and higher ability.

Additionally, branching items have the potential to introduce dynamic scaffolding into test content. Scaffolding is a strategy that provides instructional supports (i.e., scaffolds) to students who may not be able to solve a task independently (Rodgers & Rodgers, 2004). Although scaffolding has traditionally been utilized in formative assessment and classroom settings, technological advancements enable dynamic assessments to make use of scaffolding within traditional

Corresponding author: C. Eckerly, E-mail: ceckerly@ets.org assessments as a strategy to elicit partial knowledge. In the branching item context, the routing component can be utilized to inform what type of scaffold will be presented to the test taker next (i.e., a more difficult component or an easier component). For example, if a student answers the routing component incorrectly, the prompt of the subsequent component can introduce context that may help the test taker answer the following component correctly. Such a design has the potential to provide test takers with opportunities to show knowledge and abilities beyond what can be shown with independent, single-score items (Wolf et al., 2016).

In practice, testing programs have been evaluating the use of TEIs like branching items alongside traditional item types (e.g., selected-response and constructed-response items) as measures of latent constructs with item response theory (IRT) methodology. The branching item format described was utilized in several pilot items as part of an interactive computer task for a pilot subject test in a large-scale educational assessment in which significant model–data misfit was observed for all branching items when standard operational scaling procedures were implemented. In this report, we describe our investigation to understand the source of the observed misfit, including (a) alternate analyses of the empirical data and (b) a simulation study designed to reproduce the observed effects. We learned that the relationship between the discrimination of the routing component and the discriminations of the subsequent components seemed to drive the misfit. We conclude with lessons learned from the pilot administration and provide suggested guidelines and considerations for operationalizing the use of branching items in future assessments.

Scoring of Branching Items

Wainer and Kiely (1987) described two ways in which branching items can be scored: (a) Score levels of an ordered polytomous IRT model are determined by various response patterns to all components, or (b) each component in the branching item is scored as a separate item. Descriptions and implications of each are described in the following sections.

Scoring Branching Items with an Ordered Polytomous Item Response Theory Model

Fully hierarchical adaptive branching items (Wainer & Kiely, 1987) in which the difficulties of the components are ordered have an ordered set of unique possible outcomes determined by the sequence of administered items and the response to the final item (Bolt, 2016). This ordered set of outcomes can be used to construct the scoring rubric. For example, the scoring categories of a branching item that includes one dichotomous routing component, one dichotomous high component (i.e., a component of higher difficulty than the router that will be presented to test takers who respond correctly to the routing component), and one dichotomous low component (i.e., a component of lower difficulty than the router that will be presented to test takers who respond incorrectly to the routing component) is shown in Figure 1. In the figure, a “1” indicates a correct response and a “0” an incorrect response. The score categories for the responses to both components are as follows: 0 points, incorrect routing component, incorrect low component; 1 point, incorrect routing component, correct low component; 2 points, correct routing component, incorrect high component; and 3 points, correct routing component, correct high component.

Branching items should be scored polytomously if conditional independence among the responses to individual components cannot be assumed. Conditional independence in the branching context is described in the following section.

Scoring Branching Item Components Separately

If branching item components are scored as separate items, it is necessary to ensure that the missing data in the high and low components do not lead to estimation problems for IRT model item parameters. When estimating item parameters

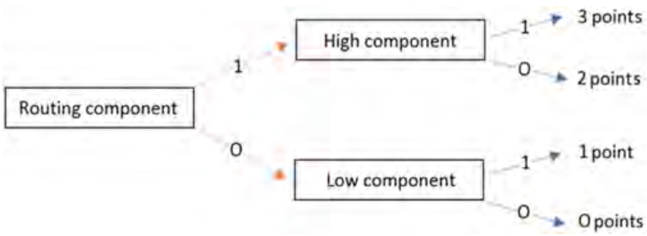

Figure 1 Example of a three-component, four-level branching item, with dichotomous components.

using marginal maximum likelihood (MML) estimation, missing data in the response matrix are considered ignorable when they are missing completely at random or missing at random (MAR; Mislevy, 2016). In the branching item context, the data missing by design (i.e., missing due to routing) in the low and high components (i.e., missing responses in the high component for students who were routed to the low component and missing responses in the low component for students who were routed to the high component) are MAR, because the probability of missingness is independent of the missing responses conditional on the routing responses. Because the missing data are MAR, the missingness can be ignored in marginal likelihood inference, so MML item parameter estimates using the likelihood are statistically consistent. However, sample sizes and ability ranges of test takers responding to each component should be monitored to ensure that item parameters of the high and low components can be estimated with sufficient precision.

It is also necessary to address potential violations of the local independence assumption that may arise because of an intended item development feature of constructing the components around a common stimulus. Marais and Andrich (2008) distinguished between two types of violations of the local independence assumption for standard unidimensional IRT models: (a) *response dependence*, in which the response function, conditional on ability, for one item depends on the observed response to another item, and (b) *trait dependence*, in which there is an additional trait dimension that is not modeled (e.g., Bolt, 2016; Li et al., 2006; Wainer et al., 2002; Wainer et al., 2007; Wainer & Kiely, 1987). DeMars (2012) noted that response dependence may occur when the response to one item, conditional on ability, depends directly on the correctness of another item, such as when the second item asks the test taker to explain their answer to the first item. Additionally, trait dependence may occur when multiple items are linked to the same stimulus but those items do not directly build off each other.

The first type of dependence, response dependence, for a simple branching item with three dichotomous components (i.e., a routing, low, and high component), can be described by the following joint probabilities representing all possible combinations of responses to the individual components for a three-component branching item with dichotomous components. Each of the probabilities on the right-hand sides of the equations can be determined by an appropriate IRT model, and the probability of response to the high or low component is conditional on the observed response to the routing component in addition to θ:

$$P\left(x_{\text{route}} = 1, x_{\text{high}} = 0 \,|\, \theta\right) = P\left(x_{\text{route}} = 1 \,|\, \theta\right) * P\left(x_{\text{high}} = 0 \,|\, x_{\text{route}} = 1, \theta\right), \tag{1}$$

$$P\left(x_{\text{route}} = 1, x_{\text{high}} = 1 \,|\, \theta\right) = P\left(x_{\text{route}} = 1 \,|\, \theta\right) * P\left(x_{\text{high}} = 1 \,|\, x_{\text{route}} = 1, \theta\right), \tag{2}$$

$$P\left(x_{\text{route}} = 0,\ \ x_{\text{low}} = 0 \,|\, \theta\right) = P\left(x_{\text{route}} = 0 \,|\, \theta\right) * P\left(x_{\text{low}} = 0 \,|\, x_{\text{route}} = 0, \theta\right), \tag{3}$$

$$P\left(x_{\text{route}} = 0, x_{\text{low}} = 1 \,|\, \theta\right) = P\left(x_{\text{route}} = 0 \,|\, \theta\right) * P\left(x_{\text{low}} = 1 \,|\, x_{\text{route}} = 0, \theta\right). \tag{4}$$

Owing to the branching in which test takers are presented either the high or low component, but not both, some combinations of components and associated responses will not be observed in the data, thus, the probability of observing any of the following response combinations is zero:

$$P\left(x_{\text{route}} = 0, x_{\text{high}} = 0 \,|\, \theta\right) = 0, \tag{5}$$

$$P\left(x_{\text{route}} = 0, x_{\text{high}} = 1 \,|\, \theta\right) = 0, \tag{6}$$

$$P\left(x_{\text{route}} = 1, x_{\text{low}} = 0 \,|\, \theta\right) = 0, \tag{7}$$

$$P\left(x_{\text{route}} = 1, x_{\text{low}} = 1 \,|\, \theta\right) = 0. \tag{8}$$

Because there are no response data to estimate the item parameters of the high component from test takers who answered the routing component incorrectly and there are no response data to estimate the item parameters of the low component from test takers who answered the routing component correctly (as reflected in Equations (5)–(8)), the following equations hold when item parameters of the IRT model are estimated using MML:

$$P\left(x_{\text{high}} = 0 \mid x_{\text{route}} = 1, \theta\right) = P\left(x_{\text{high}} = 0 \mid \theta\right), \tag{9}$$

$$P\left(x_{\text{high}} = 1 \mid x_{\text{route}} = 1, \theta\right) = P\left(x_{\text{high}} = 1 \mid \theta\right), \tag{10}$$

$$P\left(x_{\text{low}} = 0 \mid x_{\text{route}} = 0, \theta\right) = P\left(x_{\text{low}} = 0 \mid \theta\right), \tag{11}$$

$$P\left(x_{\text{low}} = 1 \mid x_{\text{route}} = 0, \theta\right) = P\left(x_{\text{low}} = 1 \mid \theta\right). \tag{12}$$

Thus, Equations (1)–(4) reduce to

$$P\left(x_{\text{route}} = 1, x_{\text{high}} = 0 \mid \theta\right) = P\left(x_{\text{route}} = 1 \mid \theta\right) * P\left(x_{\text{high}} = 0 \mid \theta\right), \tag{13}$$

$$P\left(x_{\text{route}} = 1, x_{\text{high}} = 1 \mid \theta\right) = P\left(x_{\text{route}} = 1 \mid \theta\right) * P\left(x_{\text{high}} = 1 \mid \theta\right), \tag{14}$$

$$P\left(x_{\text{route}} = 0, x_{\text{low}} = 0 \mid \theta\right) = P\left(x_{\text{route}} = 0 \mid \theta\right) * P\left(x_{\text{low}} = 0 \mid \theta\right), \tag{15}$$

$$P\left(x_{\text{route}} = 0, x_{\text{low}} = 1 \mid \theta\right) = P\left(x_{\text{route}} = 0 \mid \theta\right) * P\left(x_{\text{low}} = 1 \mid \theta\right), \tag{16}$$

indicating that response dependence cannot lead to violations of the local independence assumption of unidimensional IRT models in the branching context. It is important to note that this will only be true if all test takers who answer the routing component correctly are routed high and all test takers who answer the routing item incorrectly are routed low (i.e., deterministic rather than probabilistic routing is used).

Although we have shown that response dependence is not a problem in the branching context, the second type of dependence, trait dependence, can still lead to violations of the local independence assumption because dependencies could potentially exist between responses to the routing component and the low component or between responses to the routing component and the high component. Thus, if individual components are to be scored as separate items and the components are linked to a common stimulus, it is still necessary to check for dependencies before choosing to score each component as a separate item.

Assessing Model–Data Fit

Throughout this report, we use two strategies for assessing model–data fit: the chi-squared fit statistic (ChiSq) produced by the version of PARSCALE modified by ETS and plots of empirical versus model-predicted probabilities. Both the ChiSq statistic and the visual plots make use of the "pseudocounts" (e.g., Stone et al., 1994) at each quadrature point, which reflect the summed posterior densities for test takers responding in each score category at each quadrature point q. The pseudocounts r for score level j given θ_q are shown by

$$r_{jq} = \sum_{n=1}^{N} x_{jn} P\left(\boldsymbol{x}_n \mid \theta_q\right) A\left(\theta_q\right) / P\left(\boldsymbol{x}_n\right),$$

where N is the number of test takers; x_{jn} is 1 if the observed response of the nth test taker for the item equals j, and is 0 otherwise; $P\left(\boldsymbol{x}_n \mid \theta_q\right)$ is the conditional probability of the nth test taker's response pattern $\boldsymbol{x}_n$ given ability level θ_q; $A\left(\theta_q\right)$ is a weight corresponding to the normal density function at θ_q; and $P\left(\boldsymbol{x}_n\right)$ is the marginal probability of observing the response pattern $\boldsymbol{x}_n$. For an item, the pseudo-counts can populate a $J \times Q$ table of observed (or empirical) counts for each score category at each quadrature point, where J is the number of score categories and Q is the number of quadrature points. Expected (or theoretical) counts can be obtained from the estimated item response function. The ChiSq statistic is then calculated in the usual way given a set of observed and expected frequencies. However, owing to using pseudocounts in the construction of the observed frequencies, a test taker's contribution to the item fit table is distributed over multiple cells, so observations are not independent; thus, the statistic is used to provide a measure of relative item fit, but the sampling distribution cannot be assumed to be chi-square.

Pseudocounts are also used to calculate empirical probabilities of scoring in each score category at each quadrature point by dividing each pseudocount by the number of test takers. These empirical probabilities are compared to theoretical probabilities shown by the item characteristic curve(s) of each item. Empirical and theoretical probabilities are overlaid on the same plot for a visual indicator of model–data fit.

Methods

Real Data Example

For the branching items from the pilot assessment, all test takers were assigned to a common routing component, while the assignment of the next component was determined by the response to the routing component. The components at both stages were scored together as one polytomous item because each component of the branching item relied on a common stimulus, suggesting that treating each component as a separate item may violate the local independence assumption. Three branching items (B1, B2, and B3) that were part of the same interactive computer task were analyzed. Each of these items has an associated routing component (R1, R2, and R3), high component (H1, H2, and H3), and low component (L1, L2, and L3).

When the branching items were developed, each was intended to be a fully hierarchical adaptive branching item (Wainer & Kiely, 1987) in which the difficulties of the components were ordered, with the routing component having a higher difficulty than the low component and a lower difficulty than the high component. Details about the scoring of the branching items are included in Table 1, in which score categories for the polytomous score rubrics of each branching item are broken down based on the number of points received on the separate components (where, e.g., 0/1 R1 indicates that a test taker scored 0 out of 1 point on R1):

To investigate how these pilot branching items functioned, IRT scaling was conducted in which dichotomous items were modeled with the three-parameter logistic (3PL) model (Birnbaum, 1968) and polytomous items were modeled with the generalized partial credit model (GPCM; Muraki, 1992) following the convention of the assessment program. The pilot assessment included a total of 280 items (including branching items) with an average of 1,300 student responses per item. After observing significant misfit of the branching items following the usual procedures for analyzing pilot data, we followed up with additional calibrations to help better understand the sources of misfit. First, we conducted scaling procedures utilizing the response data for the 277 nonbranching items in the pilot to set the scale without the influence of the branching items. Second, three separate calibrations were conducted to project the branching items onto the scale. In these calibrations, the parameters of the nonbranching items were fixed, but the branching item parameters were estimated based on three different scoring methods.

1. *Calibration 1: Original rubric scoring*. Branching items were scored according to the score categories from the original deterministic rubric described in Table 1.
2. *Calibration 2: Collapsed categories rubric scoring*. Branching items were scored by collapsing some score categories to improve IRT model fit.
3. *Calibration 3: Separate items scoring*. Branching item components (i.e., routing, high, and low) were scored as separate items.

Calibration 1 was used as a baseline to evaluate model–data fit using the original scoring rubrics. Calibration 2 followed the usual procedure of the testing program to collapse score categories in which the group of students who responded in adjacent categories had similar overall ability as measured by their performance on the rest of the test. Calibration 3 was used as a tool to evaluate the performance of individual components to see if the overall misfit could be attributed to an individual component. Each of the calibrations was completed using expectation–maximization MML (EM-MML; Bock & Aitkin, 1981; Bock & Lieberman, 1970) in PARSCALE (Muraki & Bock, 1991).

Table 1 Original Scoring Rubrics of Branching Items

	Branching item		
Score category	1	2	3
0	0/1 R1, 0/1 L1	0/1 R2, 0/1 L2	0/2 R3, 0/1 L3
1	0/1 R1, 1/1 L1	0/1 R2, 1/1 L2	0/2 R3, 1/1 L3
2	1/1 R1, 0/1 H1	1/1 R2, 0/2 H2	1/2 R3, 0/1 L3
3	1/1 R1, 1/1 H1	1/1 R2, 1/2 H2	1/2 R3, 1/1 L3
4	–	1/1 R2, 2/2 H2	2/2 R3, 0/2 H3
5	–	–	2/2 R3, 1/2 H3
6	–	–	2/2 R3, 2/2 H3

Table 2 Pilot Exam: Branching Item Descriptions

	Branching item		
Item property	B1	B2	B3
No. of score categories			
Total	**4**	**5**	**7**
Routing	2	2	3
High	2	3	3
Low	2	2	2
Percentage correct: Routing component	83	52	80
IRT difficulty parameter (from Calibration 3)			
Routing	−2.2	0.7	−2.3
High	2.8	1.5	0.0
Low	0.5	−0.2	0.8
Percentage test takers in each score category			
0	10.9	12.3	3.9
1	5.8	18.3	1.7
2	80.1	19.1	9.1
3	3.2	23.1	6.2
4	–	27.2	32.3
5	–	–	7.4
6	–	–	39.5

Real Data Results

The proportions of test takers correctly responding to the routing components were .83, .52, and .80 for R1, R2, and R3, respectively, indicating that R1 and R3 were much easier than R2. When a routing component is very easy, it may be difficult to construct a subsequent component that is easier. To assess the difficulty of the high and low components of each branching item, we used Calibration 3 to evaluate each component of each branching item in an IRT framework. Although Calibration 3 was used as a tool to evaluate the performance of individual components of branching items, results should be interpreted with caution because of potential violations of the local independence assumption and, in some instances, small sample sizes for estimation. Small sample sizes were observed for low components that followed easy routing items (i.e., B1 and B3). For example, in Item B3, only 90 test takers were routed to and responded to L3 (a multiple-choice item with four options), and of those, only 25 test takers answered it correctly, slightly more than would be expected if those 90 test takers had responded to the item using random guessing. Parameter estimates for the separate components for each branching item and additional descriptive statistics are shown in Table 2.

On the basis of the preceding analysis of the three branching items B1, B2, and B3, we will focus on B2 because the IRT analysis of the individual components and the percentages of students responding correctly to the router suggest that it was the only branching item of the three that performed as intended, with a medium-difficulty routing component, lower difficulty low component, and higher difficulty high component. Although the individual components of Item B2 showed acceptable fit and generally exhibited difficulties in the targeted range, when all

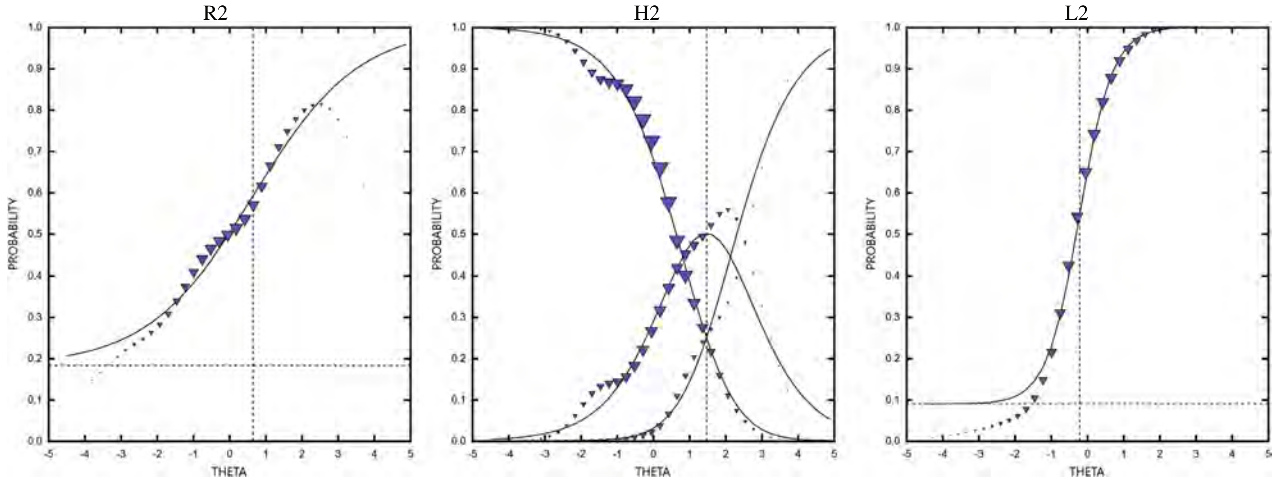

Figure 2 Model – data fit of individual components of Branching Item 2 (B2) from Calibration 3.

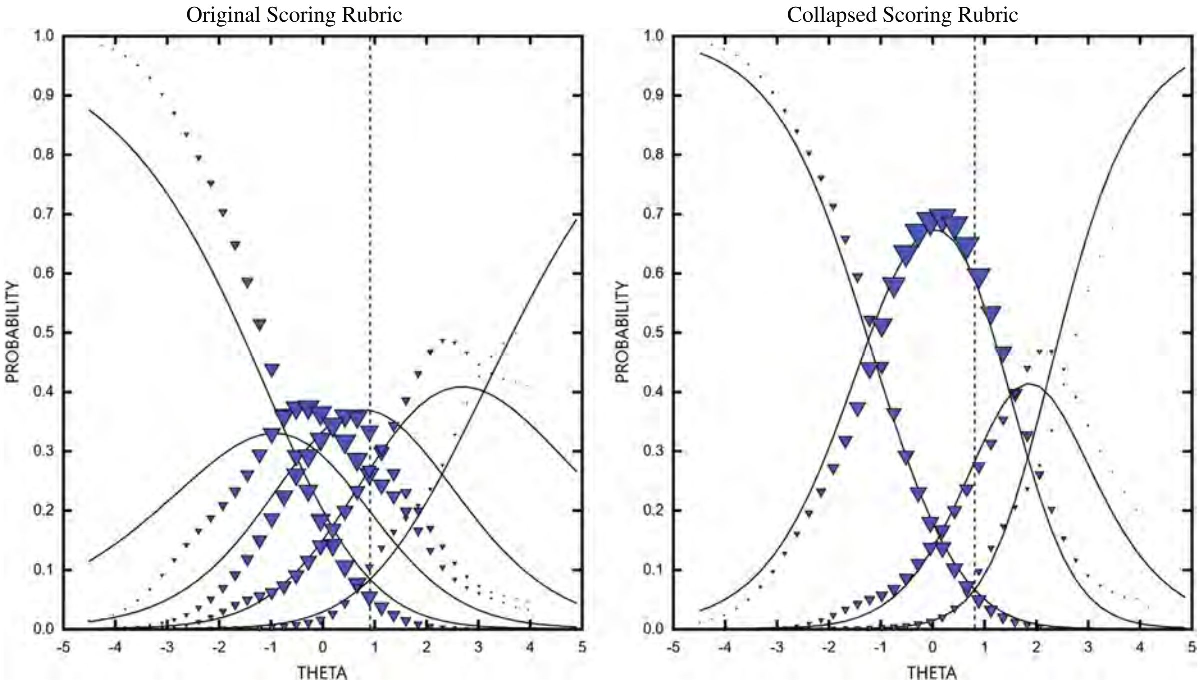

Figure 3 Model – data fit of Branching Item 2 (B2): Original scoring rubric (Calibration 1) and collapsed scoring rubric (Calibration 2).

components were scored as 1 five-category polytomous item using the GPCM, significant model misfit was observed. This can be seen in Figures 2 and 3, in which plots of empirical versus model-predicted probabilities are shown associated with Calibration 3 (Figure 2) and Calibrations 1 and 2 (Figure 3). The blue triangles indicate the empirical probabilities at each quadrature point, and the black lines represent the item characteristic curves describing model-predicted probabilities. Larger blue triangles are associated with higher sample sizes of test takers for the associated quadrature point.

By collapsing score categories associated with 1 and 2 points, overall fit of the model improved greatly. These two score categories were associated with test takers who answered the routing component incorrectly but the low component correctly and with test takers who answered the routing component correctly but received zero points on the high component, respectively. We will refer to collapsing these two categories as collapsing at the boundary of the router.

Simulation Study

To better understand the potential causes of the model–data misfit observed in the branching items, we designed a small simulation study to examine the effects of various levels of discrimination parameters of the routing and high/low components for a simple branching item consisting of dichotomous routing, high, and low components. We based the design of the simulation on the ideal scenario observed with the second branching item B2, where the routing component was of medium difficulty, the low component was of lower difficulty, and the high component was of higher difficulty. Because we were interested in evaluating the misfit that was observed when branching items were scored polytomously, we did not simulate any conditional dependence between responses to components of the branching item. While we anticipate that different types of misfit may occur in designs that deviate from this situation (e.g., an easy routing component that routes to different components, both of higher difficulty), we focused on this condition in which the branching item is designed as a best case scenario for branching as described by Wainer and Kiely (1987).

We first generated response data for 170 dichotomous nonbranching items in which responses fit a 3PL model with $\theta \sim N(0, 1)$ for 25,000 test takers, in which the generating item parameters were set to the estimates from the pilot exam. To simplify the analysis, we only simulated responses to the dichotomous items from the pilot assessment. We simulated large sample sizes of both test takers and items with the intent of introducing only negligible sampling error because we were not investigating misfit due to sampling. Additionally, response data were generated for dichotomous components of the branching items, in which the generating item parameters varied across two factors: the discrimination parameter of the routing component (i.e., a_{route}) and the common discrimination parameter of the high and low components (i.e., $a_{\text{high\_low}}$). The levels of each factor were set to the 10th, 30th, 50th, 70th, and 90th percentiles of the estimated discrimination parameters for the 3PL items in the pilot data (i.e., 0.48, 0.64, 0.80, 0.96, and 1.31). Both of these factors were fully crossed, leading to 25 separate branching items. We chose to vary the discrimination parameters of the components because of what was observed for Item B2 in the real data. When each component of B2 was scored separately (in Calibration 3), each of the components had a different estimated discrimination parameter, and acceptable model–data fit was observed for each component. However, when the item was scored polytomously (in Calibration 1), misfit was observed, leading us to question the role of the differing discrimination of the individual components in the misfit of the polytomously scored item.

The generating difficulty parameters of the low, routing, and high components of each branching item were −1.0, 0.0, and 1.0, respectively, to reflect the ideal ordered item difficulties of branching items. The generating guessing parameters for each component were set to the average guessing parameter estimate from the pilot data (i.e., .24). Response data for the branching items were generated by first simulating responses for the routing, high, and low components across all test takers. Then, responses to low items were treated as missing for test takers whose simulated router responses were correct, and responses to high items were treated as missing for test takers whose simulated router responses were incorrect. The item was then scored according to the rubric described in the example branching item shown in Figure 1.

To scale each of the branching items, a calibration was first conducted on the response data for the dichotomous nonbranching items using the 3PL model; item parameters were estimated using EM-MML with PARSCALE. Then, for each simulated branching item, a separate calibration was conducted in PARSCALE, fixing the item parameters to those estimated based only on the responses to the dichotomous items, while only estimating the parameters of the GPCM for the branching item. We chose to evaluate the model fit of the branching items in this context because we are evaluating how this particular TEI type functions when assuming that the branching items are created to measure a unidimensional construct that can be measured with traditional multiple-choice questions.

Parameter recovery was sufficient, where correlations between the generating values and the *a*, *b*, and *c* parameter estimates were .994, .998, and .936, respectively. Owing to the large sample size and sufficient item parameter recovery, little variability in the ChiSq fit statistics was expected across additional replications. To ensure that our expectation was accurate, we conducted one additional replication and observed that the distributions of ChiSq statistics for both replications had very similar shape and that the correlation between the fit values produced by the two replications was .996. The results presented in this report are from the first replication.

Simulation Results

Total test scores were calculated for each test taker by summing across all nonbranching item responses. Figure 4 shows mean total test score by each score category for each of the simulated branching items, separated by the various levels of $a_{\text{high\_low}}$. This figure helps to describe how well each score category of each simulated branching item separates test takers of different ability levels. By design of the branching item, the average scores for the rest of the items on the test should increase as the score category increases if the item is discriminating well. Figure 4 shows that this is generally not the case for Score Categories 1 and 2 (i.e., score categories at the boundary of the router). Conditions with high values of $a_{\text{high\_low}}$ relative to values of a_{route} show either minimal difference between average score of test takers scoring in Categories 1 and 2 or test takers in Category 1 actually exhibiting higher scores than those in Category 2.

These results are not unexpected because the routing item will inherently misclassify a number of test takers in both directions (high and low). Thus, we may not observe a meaningful difference between test takers who incorrectly responded to the router but correctly responded to the low item and test takers who correctly responded to the router but incorrectly responded to the high item. When model misfit is observed for a polytomous item and the groups of test takers that scored in each of two adjacent score categories have similar average performance on the rest of the test, it is common to collapse those adjacent categories into one score level to improve model–data fit (see, e.g., National Center for Education Statistics, 2008).

Informed by the real data analysis and results in Figure 4, we employed the strategy of collapsing score categories at the boundary of the router (i.e., Score Categories 1 and 2) for each of the branching items to observe the effect on model fit. In some instances, the patterns observed for simulated branching items with high levels of a_{route} and low levels of $a_{\text{high\_low}}$ showed higher levels of separation of test taker ability between Score Categories (1 and 2) than between Score Categories (0 and 1) and (2 and 3), we did not consider collapsing the latter two pairs of categories as a strategy to improve fit, because the resulting item would only reflect test taker responses to the routing component.

Figure 5 shows comparisons of empirical versus model-predicted probabilities for the simulated branching items with the highest and lowest values of ChiSq, respectively, scored using the original rubric. The item with the best fit (i.e., the lowest level of ChiSq) is associated with a highly discriminating router relative to the high and low components, and the item with the worst fit (i.e., the highest level of ChiSq) is associated with highly discriminating low and high components relative to the router.

The results of Figure 5 suggest that the difference between the discrimination parameter of the routing and high/low components may be a predictor of the ChiSq fit statistic, as this difference is large and negative for the item with the best fit and large and positive for the item with the worst fit. To further explore this relationship, Figure 6 shows a plot of $\left(a_{\text{high\_low}} - a_{\text{route}}\right)$ versus the ChiSq fit statistic for each simulated branching item, for which observations are color coded by the level of a_{route}. Results are shown separately for simulated branching items scored with the original rubric (with Score Categories 0, 1, 2, and 3) versus the collapsed rubric (with Score Categories 0, 1, and 2, for which Score Category 1 combines Score Categories 1 and 2 from the original rubric). For the original rubric, we see a pronounced quadratic relationship between $\left(a_{\text{high\_low}} - a_{\text{route}}\right)$ and the ChiSq fit statistic, with the highest levels of $\left(a_{\text{high\_low}} - a_{\text{route}}\right)$ being associated with the worst fit. The discrimination of the router also appears to have a modest effect, because for a given level of $\left(a_{\text{high\_low}} - a_{\text{route}}\right)$, lower discriminating routers are associated with better fit.

To quantify these relationships between the predictors $\left(a_{\text{high\_low}} - a_{\text{route}}\right)$ and a_{route} and the outcome variable ChiSq for simulated branching items using both the original and collapsed rubrics, we estimated the following regression models:

Model 1: $\text{ChiSq}_{\text{original}}$ regressed on $\left(a_{\text{high\_low}} - a_{\text{route}}\right)$, $\left(a_{\text{high\_low}} - a_{\text{route}}\right)^2$, and $\left(a_{\text{route}}\right)$.

Model 2: $\text{ChiSq}_{\text{original}}$ regressed on $\left(a_{\text{high\_low}} - a_{\text{route}}\right)$ and $\left(a_{\text{high\_low}} - a_{\text{route}}\right)^2$.

Model 3: $\text{ChiSq}_{\text{collapsed}}$ regressed on $\left(a_{\text{high\_low}} - a_{\text{route}}\right)$, $\left(a_{\text{high\_low}} - a_{\text{route}}\right)^2$, and $\left(a_{\text{route}}\right)$.

Model 4: $\text{ChiSq}_{\text{collapsed}}$ regressed on $\left(a_{\text{high\_low}} - a_{\text{route}}\right)$ and $\left(a_{\text{high\_low}} - a_{\text{route}}\right)^2$.

Model 2 is a reduced version of Model 1, and Model 4 is a reduced version of Model 3.

Results for Models 1 and 2 are shown in Table 3, and results for Models 3 and 4 are shown in Table 4.

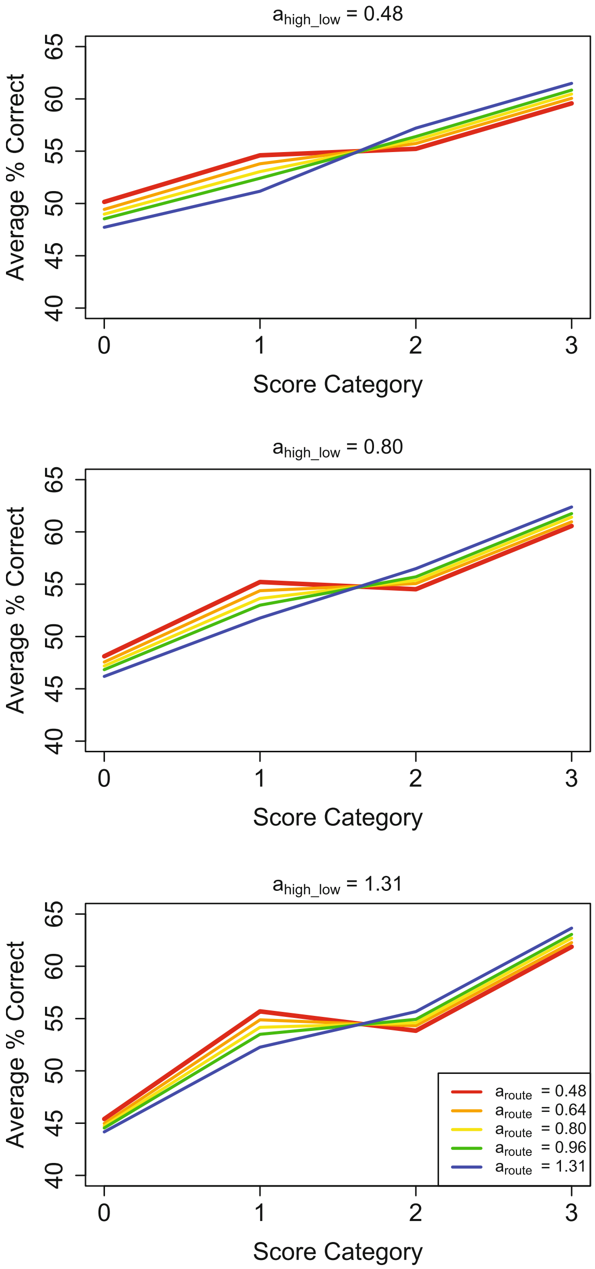

Figure 4 Average score by score category for simulated branching items.

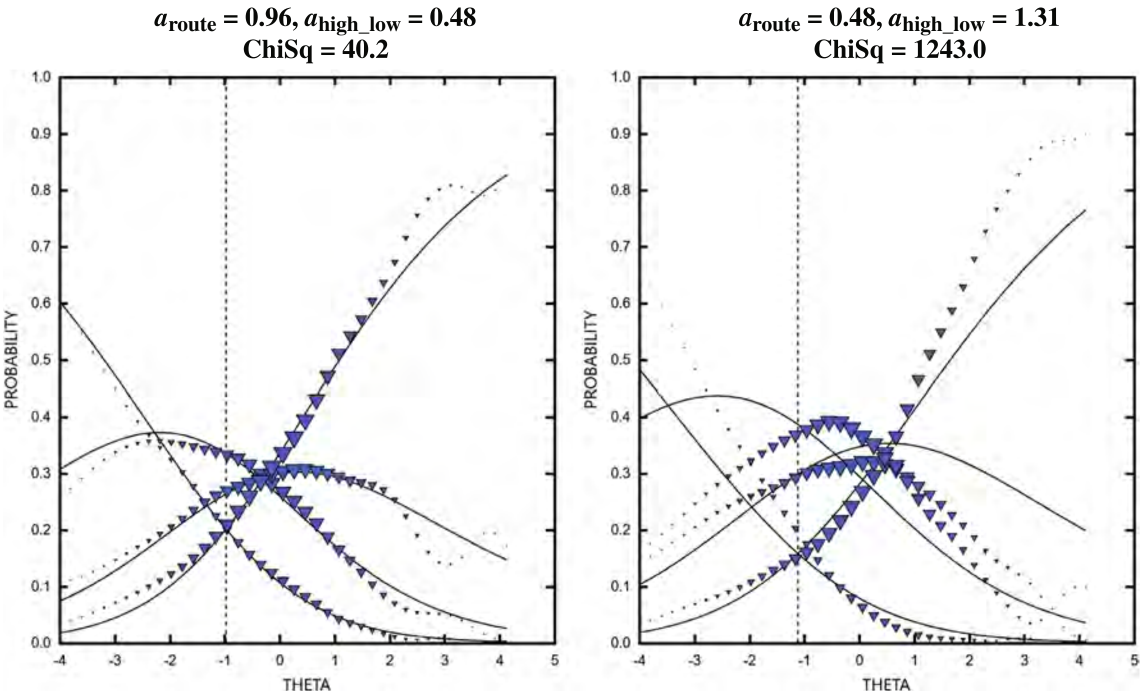

Figure 5 Model–data fit of simulated branching items with lowest and highest ChiSq: Original rubric.

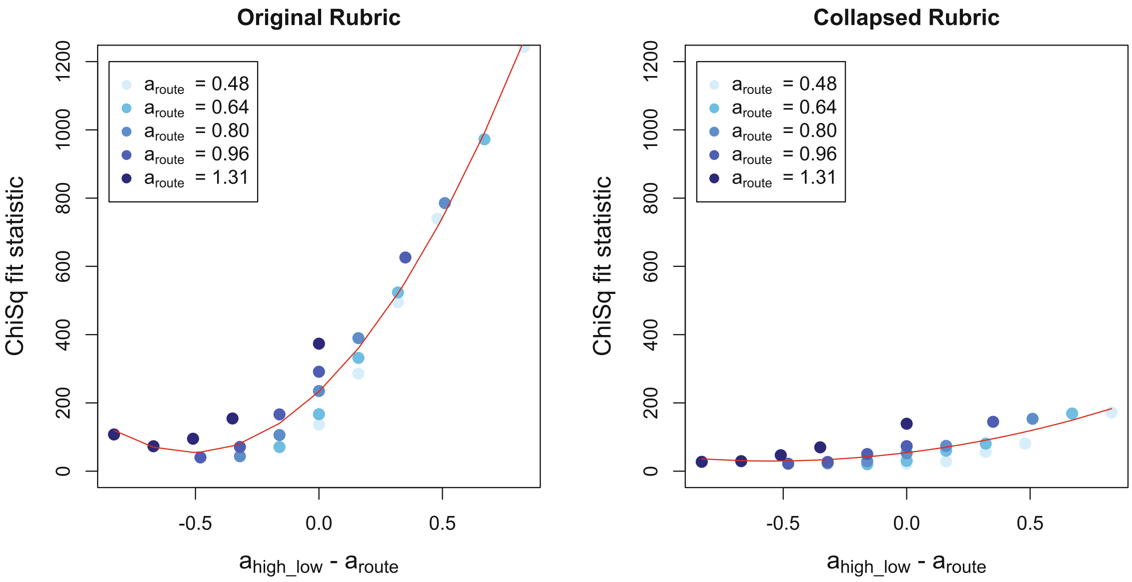

Figure 6 ChiSq fit statistic of simulated branching items: Original versus collapsed rubrics.

Table 3 Regression Results: $\text{ChiSq}_{\text{original}}$

	Model 1		Model 2	
Variable	*B*	*SE B*	*B*	*SE B*
$(a_{\text{high\_low}} - a_{\text{route}})$	823	15.3	687.2	14.3
$(a_{\text{high\_low}} - a_{\text{route}})^2$	609	12.8	662.9	27.7
a_{route}	255.7	18.4		
R^2	0.997		0.972	
N	25		25	

Table 4 Regression results: $\text{ChiSq}_{\text{collapsed}}$

	Model 3		Model 4	
Variable	*B*	*SE B*	*B*	*SE B*
$(a_{\text{high\_low}} - a_{\text{route}})$	159.8	6.8	88.66	15.0
$(a_{\text{high\_low}} - a_{\text{route}})^2$	44.4	9.9	80.23	30.2
a_{route}	140.4	9.8		
R^2	0.971		0.651	
N	25		25	

While both Models 1 and 3 explain most of the variation in their respective outcome variables $\text{ChiSq}_{\text{original}}$ and $\text{ChiSq}_{\text{collapsed}}$, each of the effects for Model 3 is smaller than its counterpart for Model 1 mainly because $\text{ChiSq}_{\text{collapsed}}$ has less variation than $\text{ChiSq}_{\text{original}}$. However, the unique explanatory contribution of a_{route} is much larger for Model 3 than for Model 1, which can be observed by removing a_{route} as a predictor in both regressions (see Models 2 and 4). In doing so, R^2 decreases from 0.997 to 0.972 for Model 2 and from 0.971 to 0.651 for Model 4.

The narrower spread of values of $\text{ChiSq}_{\text{collapsed}}$ indicates that collapsing score categories at the boundary of the router can be used as a strategy to improve model fit. For each of the simulated branching items, collapsing these two score categories decreases the ChiSq fit statistic. While we would expect some decrease in ChiSq for the collapsed versus original version of an item even in the absence of improved fit due to summing over a smaller number of cells (see Stone et al., 1994), the magnitude of the decrease in the range of $\text{ChiSq}_{\text{collapsed}}$ versus $\text{ChiSq}_{\text{original}}$ for the simulated items is telling. These results are consistent with what was observed in Figure 4, where the average performance of students who scored in Categories 1 and 2 was similar in most instances, suggesting that it is not appropriate to model two separate scoring categories.

The GPCM has a slope parameter α_j that is constant for all adjacent category characteristic curves of an item j, as shown by

$$p\left(x_j = k_j \mid \theta\right) = \frac{\exp\left[\sum_{h=1}^{k_j} \alpha_j\left(\theta - \delta_{jh}\right)\right]}{\sum_{c=1}^{m_j} \exp\left[\sum_{h=1}^{c} \alpha_j\left(\theta - \delta_{jh}\right)\right]},$$

where θ is the latent trait, δ_{jh} is the transition location parameter between categories h and $h-1$, and m_j is the number of categories. This common slope parameter α_j indicates the assumption of the model that responses in each of two adjacent categories (h and $h-1$) can be modeled using dichotomous 2PL IRT models with different locations δ_{jh} but the same slope α_j. See the appendix for a detailed description and evaluation of this assumption.

The results of the simulation indicate that this assumption will likely be violated in the branching context. Even in the conditions in which individual components (i.e., routing, high, and low) had the same discrimination parameters when treated as separate items, the observed model misfit suggests that a common discrimination parameter cannot be assumed to model the transition from each category to an adjacent category. The results shown in Figure 6 and Table 3 suggest that the best-fitting branching items prior to any collapsing of score categories will be those that have a routing component that is more discriminating than the high and low components. However, as the difference gets very large, misfit will again start to increase. This is illustrated well with results in the conditions that have the lowest level of $a_{\text{high\_low}}$ (i.e., .48). As

a_{route} increases in these conditions, the ChiSq statistic steadily decreases until $a_{\text{route}} = .96$, but when a_{route} further increases to 1.31, the ChiSq statistic increases again, illustrating the quadratic relationship between ChiSq and $\left(a_{\text{high\_low}} - a_{\text{route}}\right)$.

Discussion

The results shown in this report suggest that constructing and modeling branching items that exhibit acceptable model–data fit is a complex task. Even if individual components of branching items fit well as separate items, scoring the branching item as one polytomous item according to Wainer and Kiely's (1987) framework will likely result in misfit when the GPCM is used to model student responses, as the overall model–data fit will be determined by several factors related to individual component properties. Thus we offer several recommendations to practitioners who plan to employ branching items scored using a polytomous IRT model.

First, it may be helpful to pilot multiple versions of individual components to have multiple potential combinations of components available to choose from for the operational branching item. For example, if two or more routing components are piloted, it is possible to choose the best-performing routing component (e.g., with difficulty in the targeted medium range) for the operational branching item to improve its performance. This strategy is similar to that employed in a multistage testing context in which blocks of items are constructed from an item pool to target-specific ability levels based on known statistical properties of the items. To better understand the properties of the individual components, a strategy of routing test takers to the low and high component probabilistically rather than deterministically in the pilot stage may prove useful as well.

Second, when the GPCM is used as the scoring model, we recommend that practitioners carefully check model fit and be prepared to collapse score categories at the boundary of the router as was illustrated in the real and simulated data examples. The simulated branching items showed that for a range of plausible combinations of routing and high/low component discrimination parameters, average simulated total test scores for the group of test takers who responded to the routing component correctly but the high component incorrectly and the group of test takers who responded to the routing component incorrectly but the low component correctly were similar when individual components were simulated to have difficulty levels in the targeted range. Intuitively, it makes sense that if the routing item is not very discriminating, misclassification at the routing stage will lead to groups of test takers in adjacent score categories that cannot be distinguished well in terms of ability. However, improving the classification of the router will not necessarily lead to acceptable model–data fit, as the GPCM's assumption of a common discrimination parameter (at the item level) can still be violated in the branching context even with highly discriminating routers, depending on the properties of the high and low components.

Last, modeling item responses with the nominal response model (NRM; Bock, 1972) rather than the GPCM may lead to improved model–data fit if sample sizes are sufficient to recover the parameters of the NRM. The GPCM is a constrained version of the NRM, in which the NRM relaxes the assumption of the GPCM that a common discrimination parameter can describe the relationships between probabilities of scoring in each pair of adjacent categories. Thus, the NRM can handle situations in which the groups of test takers scoring in different categories are very similar to each other and/or the assumed ordinal pattern of the scoring rubric does not hold. Although the NRM does not require that score categories coincide with the ordering in the scoring rubric, if an underlying ordinal relationship among the categories is present, the parameter estimates of the NRM should reflect that relationship when sample size in each category is sufficient and adjacent categories are sufficiently discriminative (García-Pérez, 2018).

These recommendations are based on the premise that branching items should be scored polytomously because the branching items in the real data example were specifically constructed with a common stimulus, which may lead to violations of the local independence assumption to some extent if the components were assumed to be separate items. However, it may be preferable to score components separately if desired model–data fit cannot be obtained by scoring polytomously and if it can be shown that assumptions for modeling the components as separate items are reasonably satisfied.

This study had several limitations. First, only dichotomous components were included in the simulation study. Second, the simulation study only examined the ideal conditions in which the routing component was medium difficulty, the low component was lower difficulty, and the high component was higher difficulty. Last, we only investigated the fit of the GPCM in this study. We hypothesize that the NRM may alleviate some of the concerns with model–data fit and suggest an investigation into the utility of the NRM for future research.

As assessment programs continue to transition to digital formats, TEIs will likely become increasingly ubiquitous. We have described several considerations for including one such item type, the branching item, on assessments while continuing to utilize traditional measurement models like the 3PL and the GPCM. Some of the discussion in this report focused on issues that are not unique to digital formats, such as violations of the local independence assumption for groups of items that share a common stimulus, while others focused on issues unique to a digital format, such as IRT model misfit, which may occur when a branching item is scored using the GPCM.

The results and discussion presented in this report highlight the need to better understand sources of misfit specific to the new item types to potentially modify either the items or the modeling strategy. In the context of the branching item, we showed that misfit of branching items that were scored polytomously with the GPCM could be improved by adjusting the scoring rubric to collapse score categories at the boundary of the router, and we discussed how it may be possible to model each component of the branching item as a separate item. While the branching item is just one of many TEI types, the analysis of potential causes and solutions of misfit presented in this report can serve as an example of the type of investigation that may need to be done to understand why TEIs may function in unexpected ways when we rely on traditional psychometric modeling techniques.

Acknowledgments

The research reported here was supported by the National Assessment of Educational Progress (ED-IES-13-C-0017) contract to ETS as administered by the Institute of Education Sciences, U.S. Department of Education. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring a test taker's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51. https://doi.org/10.1007/BF02291411

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of the EM algorithm. *Psychometrika*, *46*, 443–459. https://doi.org/10.1007/BF02293801

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for dichotomously scored items. *Psychometrika*, *35*, 179–197. https://doi.org/10.1007/BF02291262

Bolt, D. M. (2016). Item response models for CBT. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 305–322). Routledge.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, *36*(2), 104–121. https://doi.org/10.1177/0146621612437403

García-Pérez, M. A. (2018). Order-constrained estimation of nominal response model parameters to assess the empirical order of categories. *Educational and Psychological Measurement*, *78*(5), 826–856. https://doi.org/10.1177/0013164417714296

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3–21. https://doi.org/10.1177/0146621605275414

Marais, I. D., & Andrich, D. (2008). Formalising dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*(3), 200–215.

Mislevy, R. J. (2016). Missing responses in item response modeling. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 2. Statistical tools* (pp. 171–194). Taylor and Francis. https://doi.org/10.1201/b19166-10

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

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

National Center for Education Statistics. (2008). Combining the categories of constructed-response items. https://nces.ed.gov/nationsreportcard/tdw/analysis/2000_2001/scaling_checks_compar_poor_comb.aspx

Parshall, C. G., & Guille, R. A. (2016). Managing ongoing changes to the test: Agile strategies for continuous innovation. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 305–322). Routledge.

Rodgers, E. M., & Rodgers, A. (2004). The role of scaffolding in teaching. In A. Rodgers & E. M. Rodgers (Eds.), *Scaffolding literacy instruction: Strategies for K–4 classrooms* (pp. 1–10). Heinemann.

Stone, C. A., Mislevy, R. J., & Mazzeo, J. (1994, April 4–8). *Classification error and goodness-of-fit in IRT models* [Paper presentation]. American Educational Research Association annual meeting, New Orleans, Louisiana, USA.

Wainer, H., Bradlow, E. T., & Du, Z. (2002). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Kluwer Academic. https://doi.org/10.1007/0-306-47531-6_13

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511618765

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185–201. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wolf, M. K., Guzman-Orth, D., Lopez, A., Castellano, K., Himelfarb, I., & Tsutagawa, F. S. (2016). Integrating scaffolding strategies into technology enhanced assessments of English learners: Task types and measurement models. *Educational Assessment*, *21*(3), 157–175. https://doi.org/10.1080/10627197.2016.1202107

Appendix A

The GPCM models the probability of a test taker responding in the hth score category of item j as

$$p\left(x_j = k_j \mid \theta\right) = \frac{\exp\left[\sum_{h=1}^{k_j} \alpha_j\left(\theta - \delta_{jh}\right)\right]}{\sum_{c=1}^{m_j} \exp\left[\sum_{h=1}^{c} \alpha_j\left(\theta - \delta_{jh}\right)\right]},$$

where θ is the latent trait, α_j is the slope parameter, δ_{jh} is the transition location parameter between the hth category and the $h-1$ category, m_j is the number of categories, and $k_j = \{1, \ldots, m_j\}$ is associated with score levels $\{0, \ldots, m_j - 1\}$. In the GPCM, the probability of a test taker scoring in category h rather than category $h-1$ in item j is modeled as a 2PL model,

$$P\left(x_j = k_j \mid x_j \in \left\{k_j k_j - 1\right\}, \theta, \alpha_j, \delta_{jh}\right) = \frac{\exp\left[\alpha_j\left(\theta - \delta_{jh}\right)\right]}{1 + \exp\left[\alpha_j\left(\theta - \delta_{jh}\right)\right]},$$

for $k_j > 1$.

The set of such 2PL models for a polytomous item describes the response process for each of $m_j - 1$ pseudoitems underlying the GPCM. The four category branching items from the simulation study in this report each has three pseudoitems associated with score levels (0,1), (1,2), and (2,3). Decomposing the GPCM into its associated pseudoitems helps with conceptual understanding of the common slope parameter α_j. Figures A1 and A2 illustrate model–data fit for each of the pseudoitems associated with the simulated branching items with the lowest ChiSq fit statistic and the highest ChiSq fit statistic, respectively. Parameters for each of the pseudoitems were fixed to values obtained from the calibration of the GPCM. To observe the empirical probabilities, data were recoded from the original score levels of the branching item to

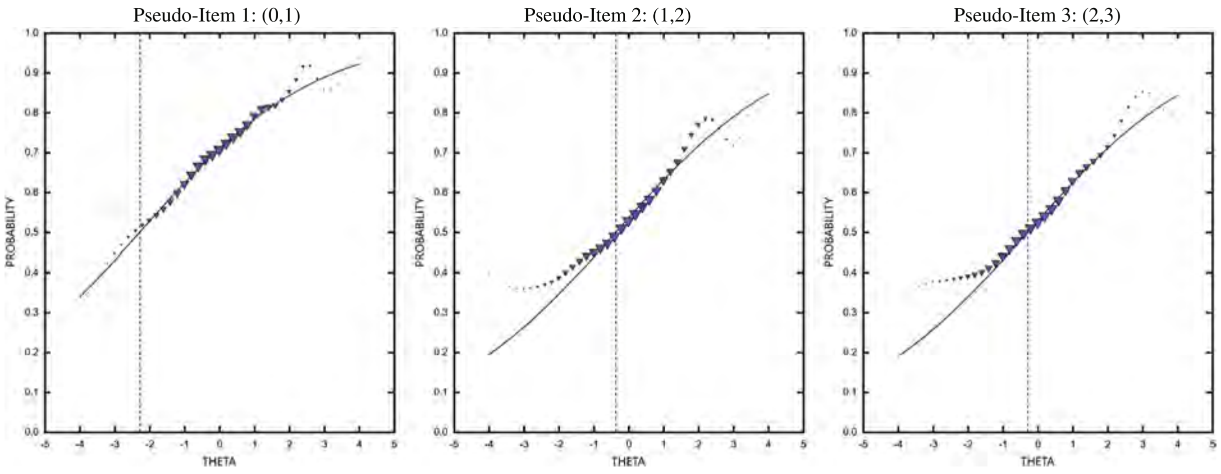

Figure A1 Model–data fit of generalized partial credit model pseudoitems for simulated branching item with lowest ChiSq fit statistic.

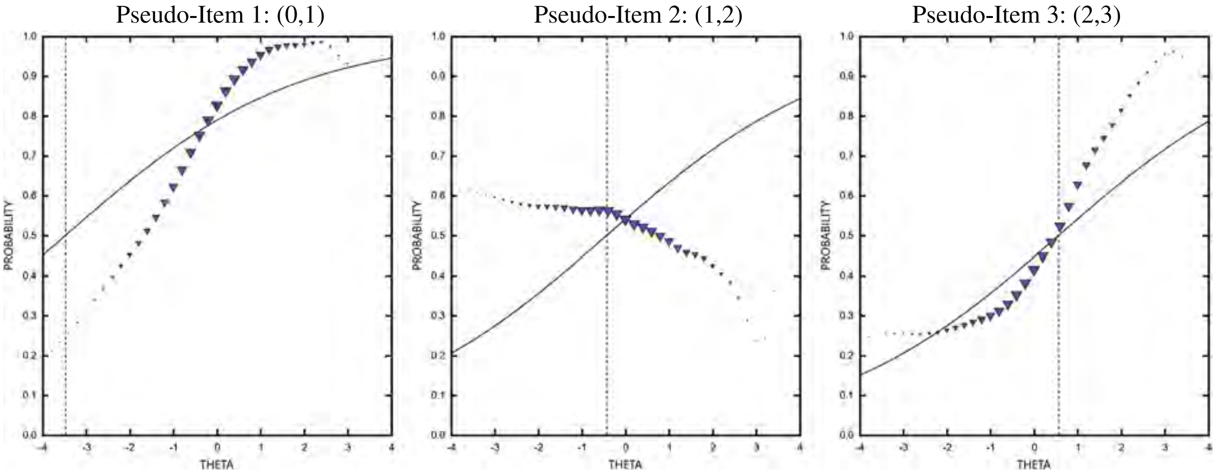

Figure A2 Model–data fit of generalized partial credit model pseudoitems for simulated branching item with highest ChiSq fit statistic.

1 for the highest score level in the pair and 0 for the lowest score level in the pair (for Pseudoitem 1, this recoding was not necessary, because the response data already followed this pattern). Item responses that were different from the adjacent pair of score levels associated with each pseudoitem were recoded as missing.

We observe that the item characteristic curves for each pseudoitem associated with the same branching item are parallel, illustrating the common slope parameter α_j. For the branching item with the lowest ChiSq fit statistic shown in Figure A1, empirical and model-based probabilities of correct response correspond closely. However, for the branching item with the highest ChiSq fit statistic shown in Figure A2, significant misfit is observed between empirical and model-based probabilities, especially for Pseudoitem 2. This indicates that the assumption of a common slope parameter α_j for the pseudoitems is not reasonable.

Suggested citation:

Eckerly, C., Jia, Y., & Jewsbury, P. (2022). *Technology-enhanced items and model–data misfit* (Research Report No. RR-22-11). ETS. https://doi.org/10.1002/ets2.12353

Action Editor: James Carlson

Reviewers: John Donoghue and Adrienne Sgammato

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/