Modeling Answer Change Behavior: An Application of a Generalized Item Response Tree Model

Minjeong Jeon
University of California, Los Angeles

Paul De Boeck
Ohio State University
KU Leuven

Wim van der Linden
Pacific Metrics

We present a novel application of a generalized item response tree model to investigate test takers’ answer change behavior. The model allows us to simultaneously model the observed patterns of the initial and final responses after an answer change as a function of a set of latent traits and item parameters. The proposed application is illustrated with large-scale mathematics test items. We also describe how the estimated results can be used to study the benefits of answer change and to further detect potential academic cheating.

Keywords: answer change; item response tree; answer change gains; academic cheating

1. Introduction

Multiple-choice assessment is a practical device to evaluate test takers’ performance and progress in educational measurement. A common multiple-choice format requires test takers to select the best answer from a number of alternatives. After having made an initial decision, test takers may subsequently identify an alternative option as the best answer. Hence, test takers may either retain their original answer or switch to an alternative answer (Milia, 2007).

A test taker’s answer change behavior has intrigued educational researchers and practitioners for quite some time. Originally, researchers have been motivated by the question of whether changing initial answers would increase students’ test scores (van der Linden & Jeon, 2012). According to conventional belief, changing original answers is likely to lower students’ overall test scores. However, a number of empirical studies have contradicted this popular belief, showing that most answer changes can be categorized as being from wrong to
right (WR) and that the majority of examinees improve their test scores via answer change behavior (e.g., Benjamin, Cavell, & Shallenberger, 1984; Crocker & Benson, 1980; Foote & Belinky, 1972; Mathews, 1929; McMorris et al., 1991; Papanastasiou & Reckase, 2008; Vispoel, 1998).

Numerous studies have also been conducted to develop a more in-depth understanding of test takers’ answer change behavior. For instance, researchers have shown that answer change gains depend on the characteristics of individual test takers; for example, test takers with moderate to high ability levels earned minor to moderate gains from answer change, while low-ability test takers earned nominal gain (McMorris, DeMers, & Schwarz, 1987). Other researchers discussed how answer change behavior is also related to test takers’ differential confidence level (e.g., McMorris et al., 1987), attitudes (e.g., Friedman & Cook, 1995), or cognitive styles (e.g., Bjorklund, 1989; Millman, Bishop, & Ebel, 1975). For example, field independent examinees are better able to isolate parts from their whole in order to analyze these parts and they are more effective in answer changes than field-dependent examinees who perceive the parts from the whole (Friedman & Cook, 1995). Reflective examinees might expend more time to answer items initially but be better at correctly changing answers when compared to impulsive examinees who minimize their time to answer items (Friedman & Cook, 1995).

It is noteworthy that many of these earlier studies on answer change are based on classical test theory. That is, simple number-correct scores and marginal counts (or proportions) of wrong to write (WR) or right to wrong (RW) changes have commonly been utilized. This classical approach, however, overlooks the issue that examinees can have different ability levels and test items can have different item properties (van der Linden, Jeon & Ferrera, 2011). Hence, an improved approach should take into account the examinees’ ability levels and individual item properties.

Recently, van der Linden, Jeon, and Ferrera (2011) and van der Linden and Jeon (2012) proposed a sophisticated item response theory (IRT) approach to model test takers’ item review and answer change process. Their model was formulated based on the assumption that examinees first produce initial answers on all items; once completing a first pass, examinees review the initial answers to either confirm or change them. To model the two-stage item review process, van der Linden et al. utilized a regular IRT model for the initial answers and employed a fixed-ability logistic regression model for the final answers, given the initial responses.

In the current study, we propose a new modeling approach to investigate examinees’ answer change behavior. The main idea is to model a sequential answer change process based on a tree structure; the process of making an initial answer is represented as a top node and the process of making a change (or no change) is represented with two intermediate nodes. The outcomes of the answer change process are modeled as the leaves of the tree (or as terminal nodes). Note that our proposed approach is based on a generalized item response tree model.
Jeon et al.

(e.g., Jeon & De Boeck, 2015). Item response tree models have received increased attention in educational measurement and have been utilized in a variety of situations, for example, for modeling response styles (Plieninger & Meiser, 2014), missing/omitted item responses (Debeer, Janssen, & De Boeck, 2014), sensitive survey questions (Boeckenholt, 2013), and slow/fast intelligence (DiTrapani, Jeon, De Boeck, & Partchev, 2016; Partchev & De Boeck, 2012). Our contribution is to provide a unique application of the tree approach to explicate the answer change process.

Our approach is similar to van der Linden et al.’s (2011, 2012) method in that the examinees’ answer change behavior is modeled as a sequence of test-taking activities: (1) giving initial answers and (2) changing or keeping these initial answers. Unlike van der Linden et al.’s method, which is based on one fixed ability, we allow that other, somewhat different latent traits are involved in the answer change process than in the initial response. In addition, we posit that the two series of activities (making initial answers and changing/keeping the initial answers) are organically related to each other. Hence, the two activities and their latent traits should be modeled and estimated simultaneously, in one single step.

For empirical illustration purposes, we will apply the proposed model to analyze a large-scale assessment data set that includes answer change. We will demonstrate that the proposed tree model is a reasonable representation of test takers’ answer change behavior. We will further show that the data analysis with the proposed approach provides new information that can improve our understanding of the answer change process. Finally, we will illustrate how the proposed approach can be utilized (1) to investigate whether or not changing an answer is beneficial to the examinees’ score and (2) to detect irregular answer change patterns (from WR) that may indicate academic cheating.

2. Modeling Item Response-and-Change Behavior

Test takers may change their initial answers during an exam. For paper tests, optical scanners can be set to detect answers to multiple-choice items that have been erased and replaced by new answers. For computerized tests, such information is readily accessible by checking the log-files. The proposed methodology can therefore be applied to both paper and computerized testing.

To model the item response-and-change behavior in multiple-choice assessments, we first defined initial responses as the first responses a subject provides and final responses as the responses a subject submits. If more than one change is made on an item, multiple intermediate responses exist between the initial and final answers. Here we concentrate on modeling only the final answer to illustrate the proposed approach more clearly within a simplified setting. An extension of the methodology that allows for multiple answer changes is discussed in Section 4.
Modeling Answer Change Behavior

If there is a change, the final responses are different from the initial responses. If there is no change, the final responses are the same as the initial responses. Response changes necessarily imply that the initial responses are reevaluated. Item reevaluation may or may not have occurred if there is no change. The possible scenarios of no change are not identifiable from the data and thus collapsed into one category in the proposed approach. We focus on making inferences based on what is observable (initial and final responses) without making assumptions about what may have happened in between. We can speculate and propose interpretations, but the model formulation itself relies only on the observed responses being correct or incorrect.

2.1 An Extended Item Response Tree Model

Suppose all initial and final responses are coded as correct or incorrect. We then observe four possible combinations of the initial and final responses:

1. initial wrong response and final wrong response (WW),
2. initial wrong response and final right response (WR),
3. initial right response and final wrong response (RW), and
4. initial right response and final right response (RR).

Note that the WW outcome can imply three potentially different scenarios: (a) The earlier response is not reviewed, (b) the earlier response is reviewed and confirmed, and (c) the earlier response is reviewed and changed to a different wrong answer. Also, the RR outcome can have two possible cases: (a) The earlier response is not reviewed and (b) the earlier response is reviewed and confirmed. Since these potential cases are either unidentifiable (e.g., [a] and [b]) or has no impact on correctness (e.g., [c]), we collapse them into the WW and RR categories, respectively.

We view these four observed patterns as the outcomes of the item response-and-change behavior and model them utilizing a tree structure. This is illustrated in Figure 1.

The tree consists of three nodes (circles) with each node having two branches (one-directional arrows). In principle, the tree diverges through branches until it reaches leaves (four observed outcomes). The tree can be expanded if multiple response options are used and if more than one change is incorporated. Extensions to multiple responses options and multiple changes are discussed in Section 4.

In the most general model formation, the probability of the left or right branch for a node depends on a latent trait and item parameters. The latent trait for Node 1 can be seen as the ability to get the item right when the item is first reviewed. The latent trait for Node 2 can be seen as the ability to make a correct change when the initial response is wrong. The latent trait for Node 3 can be seen as the ability to make no change when the initial response is right.
In this most general case, all three latent variables are assumed to be differentiable. For example, information processing can be automated for the initial responses but controlled for the final responses. It is important to note that respondents do not need to be aware of whether their earlier responses are correct or incorrect to further differentiate Nodes 2 and 3. Latent variables for Nodes 2 and 3 can still be differentiable because different state of minds or abilities can play a role. For example, in rectifying an incorrect response (Node 2), respondents’ ability in a reflective state (which may be different from an initial state) and/or self-criticism ability may be involved, whereas in keeping or changing a correct response (Node 3), respondents’ nervousness about the responses being final and/or uncertainty about their earlier answers may be involved. These are of course speculative interpretations and other interpretations are certainly possible. These speculations fulfill only the role of explaining why the three latent variables do not need to be the same. Whether the latent variables can be differentiated is an empirical decision.

In contrast to the most general case, not all three latent variables may be distinguishable from each other (or perfectly correlated with each other). It is also possible that the latent variable for Node 1 is distinguishable from the latent variables for Nodes 2 and 3, while Nodes 2 and 3 are not differentiable because they are based on the same final response ability. The above two simpler cases can be formulated as constrained versions of the most general model with three node-specific latent traits. The item parameters can also be postulated to be equal across nodes similar to the latent traits for these simpler cases. An optimal model structure can be empirically determined based on model fit comparisons (which will be illustrated in Section 3.2). Hence, from now on, we will concentrate on describing the most general case with a latent trait and a set of item parameters per node.
2.2 Mathematical Formulation

The outcome of Node 1, denoted by $Y_{pi}^{\Delta P}$, represents the initial response of person $p \neq 1; \ldots; N$ to item $i \neq 1; \ldots; I$. The initial item response (for Node 1) can be explained by a two-parameter item response model as follows:

$$
\Pr[Y_{pi}^{\Delta P} = 1 | Y_{pi}^{\Delta P} = 1; X_{pi}^{\Delta P}; a_i, b_i] = F_{a_i, b_i}(Y_{pi}^{\Delta P})
$$

where superscript (1) stands for the first node, $Y_{pi}^{\Delta P} = 1$ means that the initial response is correct ($Y_{pi}^{\Delta P} = 0$ means that the initial response is incorrect), $Y_{pi}^{\Delta P}$; is the ability to make a correct response when the person solves the item initially, $f_{a_i, b_i}$ is the inverse of the link function (which is typically a logit or probit link for binary responses, leading to a two-parameter logistic or a probit model, respectively), and $a_i$ and $b_i$ are the slope (or loading) and the intercept parameter for item $i$ (the difficulty parameter can be obtained as $b_i^{\Delta P} = b_i^{\Delta P}$). If guessing is suspected to occur during the initial answer process, a three-parameter formulation may be employed by including guessing parameters in Equation 1. However, it should be noted that a huge sample size would be needed for a reliable estimation of the model parameters.

For Node 2, the outcome variable $Y_{pi}^{\Delta P}$ indicates whether the initial incorrect response is rectified. For Node 3, the outcome is $Y_{pi}^{\Delta P}$ that indicates whether the initial right response is maintained. The second and third nodes are modeled with respective two-parameter item response models as follows:

$$
\Pr[Y_{pi}^{\Delta P} = 1 | Y_{pi}^{\Delta P} = 1; X_{pi}^{\Delta P}; a_i, b_i] = F_{a_i, b_i}(Y_{pi}^{\Delta P})
$$

$$
\Pr[Y_{pi}^{\Delta P} = 1 | Y_{pi}^{\Delta P} = 1; X_{pi}^{\Delta P}; a_i, b_i] = F_{a_i, b_i}(Y_{pi}^{\Delta P})
$$

where superscripts (2) and (3) stand for the second and third nodes conditional on the initial wrong and right responses, respectively. Specifically, $Y_{pi}^{\Delta P} = 1$ indicates that an incorrect answer is changed to a correct answer, $Y_{pi}^{\Delta P} = 1$ indicates that a correct answer is not changed, $Y_{pi}^{\Delta P}$ is the person $p$’s ability to make a correct response when the initial response is correct and $Y_{pi}^{\Delta P}$ is the ability to stay with an initial correct response, and $a_i$, $b_i$ and $a_i^{\Delta P}$, $b_i^{\Delta P}$ are the item slope and intercept parameters for the second and third nodes, respectively. Equations 2 and 3 indicate the probabilities of the conditional responses, given the outcomes of the first node ($Y_{pi}^{\Delta P} = 0$ or $Y_{pi}^{\Delta P} = 1$).

We note that the issue of guessing can be somewhat different depending on the node. Node 1 outcomes represent correct and incorrect answers to multiple-choice items; hence, it may be reasonable to consider guessing. For the other nodes, however, one may assume that subjects have good reasons for changing or not changing an answer, so that guessing may be unlikely to happen. The
possibility of including guessing parameters in all or some of the nodes may also be empirically tested. Obviously, the sample size requirement should also be taken into account when the inclusion of guessing parameters is considered.

To model the probability of the tree outcomes (WW, WR, RW, and RR), denoted by \( Z_{pi} \) hereafter, we use the fact that \( Z_{pi} \) can be recoded with a set of node-specific outcomes \( (Y_{pi}, Y_{pi}^{ unpl}, Y_{pi}^{ repl}) \). The following \( T \) matrix shows how \( Z_{pi} \) is related to \( Y_{pi}^{ unpl}, Y_{pi}^{ repl}, \) and \( Y_{pi}^{ repl} \):

\[
\begin{array}{ccc}
Y_{pi}^{ repl} & Y_{pi}^{ repl} & Y_{pi}^{ repl} \\
Z_{pi} & 1/4 & 1 & 0 & 0 & \text{NA} \\
Z_{pi} & 1/4 & 2 & 0 & 0 & \text{NA} \\
Z_{pi} & 1/4 & 3 & 1 & 0 & \text{T} \\
Z_{pi} & 1/4 & 4 & 1 & 1 & \text{NA} \\
\end{array}
\]

For instance, \( Z_{pi} \) \( 1/4 \) (WW) corresponds to \( Y_{pi}^{ repl}, Y_{pi}^{ repl}, Y_{pi}^{ repl} \) \( 1/4 \); \( Y_{pi}^{ repl} \) \( 0 \); \( \text{NA} \) where \( \text{NA} \) represents a missing observation. Now, the probabilities of observing \( Z_{pi} \) can be computed as the product of the node-specific probabilities as follows:

\[
\begin{align*}
\text{Pr} \{ Z_{pi} \} & = \text{Pr} \{ Y_{pi}^{ repl} \} \cdot \text{Pr} \{ Y_{pi}^{ repl} \} \cdot \text{Pr} \{ Y_{pi}^{ repl} \} \\
\end{align*}
\]

Observe that the node-specific probability with a missing observation (NA) does not contribute to the probability of observing \( Z_{pi} \).

Based on Equations 5 through 8, the likelihood function for the general model can be written with \( Z_{pi} \) \( 1/4 \) \( m \) (\( m \) \( 1/4 \) \( 1, 2, 3, 4 \)):

\[
I \quad X \quad X \quad Z \\
\rho^{k1} \quad 1/\lambda^{1/1} \quad \text{Pr} \{ Z_{pi} \} \quad m \quad j \quad p \quad \text{Pr} \{ Y_{pi}^{ repl} \} \quad \text{Pr} \{ Y_{pi}^{ repl} \} \quad \text{Pr} \{ Y_{pi}^{ repl} \} \\
\]

where \( \rho^{k1} \) is the distribution of the three latent variables (will be specified later). Equation 9 can be rewritten using the node-specific subresponses \( Y_{pi}^{ repl} (k \quad 1/4 \) \( 1, 2, 3 \) ) as:

\[
I \quad X \quad X \quad Z \\
\rho^{k1} \quad 1/\lambda^{1/1} \quad \text{Pr} \{ Y_{pi}^{ repl} \} \quad 4 \quad m \quad j \quad p \quad \text{Pr} \{ Y_{pi}^{ repl} \} \quad \text{Pr} \{ Y_{pi}^{ repl} \} \quad \text{Pr} \{ Y_{pi}^{ repl} \} \\
\]

where \( T_{mk} \) is the \( k \)th node response that is related to \( Z_{pi} \) \( 1/4 \) \( m \) \( 1/4 \) \( 1 \); \( \ldots \); \( 4 \) and \( t_{mk} \) \( 1/4 \) if \( T_{mk} \) \( 1/4 \) \( 1 \) or \( 0 \) and \( t_{mk} \) \( 1/4 \) \( 0 \) if \( T_{mk} \) \( 1/4 \) \( \text{NA} \).

To specify the latent distribution \( \rho^{k1} \), we assume that the three latent traits \( Y_{pi}^{ repl}, Y_{pi}^{ repl}, \) and \( Y_{pi}^{ repl} \) are normally distributed with mean 0 and covariance \( S \). That is,
Note that the correlations between the three latent variables provide insights on the latent traits involved in the item response-and-change behavior. Null correlations suggest that answer-change-related abilities are unrelated to the initial problem-solving ability. That is, how well individuals solve items correctly upon an initial evaluation does not necessarily predict whether they are able to make correct changes. High correlations between the three latent traits indicate that the initial problem-solving ability is a good predictor of the outcomes for answer change activities. For example, high correlations between $y_{p}^{\Delta p}$ and $y_{p}^{\Delta p}$ indicate that people who are able to solve items correctly upon an initial evaluation are also able to rectify their mistakes if any.

Simpler models that we considered earlier can be formulated by considering relevant constraints on $a_k$, $b_{p_k}$, and $b_i$ parameters for the general model. For instance, the linear predictor of the model with one ability $y_p$ (but a different set of item parameters) across nodes can be formulated in matrix form as follows:

$$
\begin{bmatrix}
2 & 2 \\
2 & 2 \\
0 & 0 \\
0 & 0 \\
3 & 0 \\
3 & 0 \\
0 & 0 \\
0 & 0 \\
2 & 2 \\
2 & 2 \\
\end{bmatrix}
\begin{bmatrix}
\delta_i \\
\delta_i \\
\delta_i \\
\delta_i \\
y_p \\
y_p \\
y_p \\
y_p \\
y_p \\
y_p \\
\end{bmatrix}
$$

where the linear predictor $y_{p_i}^{\Delta p} \frac{1}{4} g_{p_i} \theta_i Y_{p_i} \frac{1}{4} 1j_{p_i}^{\Delta p}$. Additionally, the item parameters can be assumed equal by constraining that $a_{i}^{\Delta p} \frac{1}{4} a_{i}^{\Delta p} \frac{1}{4} a_{i}^{\Delta p}$ and $b_{i}^{\Delta p} \frac{1}{4} b_{i}^{\Delta p} \frac{1}{4} b_{i}^{\Delta p}$.

### 2.3 Identification and Estimation of the General Model

Since the proposed model can be parameterized as a simple-structure multidimensional model, the model is identified by complying with conventional identification constraints for simple-structure multidimensional models. Specifically, we fix the mean and variance of the latent distribution $\mu_{\Delta p}$ to 0 and 1, respectively. That is, we specify $p(q_{j}^{\Delta p} N(0;1) ; S \in \mathcal{P}$ with the diagonal elements of $S$ fixed to 1.

The node-specific response vector $Y_{p_i}^{\Delta p} \frac{1}{4} \delta_{p_i}^{\Delta p} ; \theta_{p_i}^{\Delta p} ; \gamma_{p_i}^{\Delta p} \in \mathcal{P}$ includes a missing observation as shown in matrix (4) if the $k$th node does not contribute to generating a particular answer change outcome ($Z_{p_i} \frac{1}{4} 1 ; 2 ; 3 ; 4$). The missing data are missing at random (MAR), given the observed responses. Therefore, the proposed model can be readily estimated with standard item response model software packages that allow for multidimensionality and missing data under MAR. We employed the R package flirt (Jeon, Rijmen, & Rabe-Hesketh, 2014) for all data analyses.

As indicated earlier, the proposed model described in Section 2.2 is formulated based on generalized item response tree modeling. For a detailed discussion and other applications of the item response tree approach, we refer readers to Jeon and De Boeck (2015).
3. Empirical Application

Here we provide an empirical example to show that the proposed model is a reasonable representation of the item response-and-change behavior. We will illustrate two potential uses of the proposed model, for (1) computing answer change gains and (2) detecting irregular WR changes that may indicate academic cheating.

The described model was applied to a large-scale, paper-and-pencil mathematics assessment that was created and administered by a large testing company. The test consisted of 43 test items and was applied to 3,900 eleventh-grade students. Students’ original answers and changed answers were detected by using a high-precision optical scanner. Only the final changes (if any) on individual items were utilized to apply the proposed modeling approach (as discussed in Section 2). We begin by providing descriptive results on answer change patterns in the data.

3.1 Descriptive Results

Among the total of 3,900 examinees, 2,528 students changed their initial answers for at least 1 item, while 1,318 students kept their original answer for all items. For those who changed their initial answers, the average sum score of the original answers (with standard deviation in parenthesis) was 17.92 (7.33), while the average sum score of the final answers was 18.63 (7.66). The average gain in the sum scores was 0.70 (1.63). For those who kept their initial answers, the average sum score of their answers was 15.74 (7.13).

In addition, the observed counts and proportions of original and changed responses in the data can be found in Table 1. Based on the table, it is clear that only a small amount (about 1.7%) of answer changes were made in this assessment. There were more WR changes (about 1.2%) than RW changes (about 0.5%). Figure 2 depicts the average marginal proportions of WR and RW changes for each of the 43 items and 50 random students.

<table>
<thead>
<tr>
<th>Initial</th>
<th>Changed</th>
<th>W</th>
<th>R</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>100,760</td>
<td>2,083</td>
<td>102,843</td>
<td></td>
</tr>
<tr>
<td>(0.60)</td>
<td>(0.012)</td>
<td></td>
<td>(0.61)</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>968</td>
<td>63,889</td>
<td>64,857</td>
<td></td>
</tr>
<tr>
<td>(0.0057)</td>
<td>(0.38)</td>
<td></td>
<td>(0.39)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>101,728</td>
<td>65,972</td>
<td>167,700</td>
<td></td>
</tr>
<tr>
<td>(0.61)</td>
<td>(0.39)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. W ¼ incorrect; R ¼ correct responses.
The results suggest that the proportions of WR and RW changes were quite small and more WR changes were made than RW changes across items and across persons. The small proportion of answer change seems to suggest that applying a two-parameter model to Nodes 2 and 3 outcomes may be challenging. However, as shown in the simulation study (Section 3.3), obtaining relatively accurate and reliable discrimination parameter estimates was nonproblematic.

3.2 Estimated Results

Now we fit the proposed model (the most general case presented in Equation 10) to the data which assume that each of the three nodes involves a latent trait and a set of item parameters. The proposed model converged and yielded a log likelihood of 111,861.7; Akaike information criterion (AIC) of 224,245.5; and Bayesian information criterion (BIC) of 225,881.6 with 261 parameters. This model fit was clearly better than a model with two latent variables (Nodes 2 and 3 are combined; log likelihood of 111,164, AIC of 224,846.1, and BIC of 226,469.7 with 259 parameters) and a model with a unidimensional latent structure that assumes only a single latent variable for the three nodes (log likelihood of 112,612.7, AIC of 225,741.5, and BIC of 227,358.8 with 258 parameters). This result suggests that the three latent traits are indeed needed to represent the item response-and-change behavior in the data.

Based on the general model, the estimated correlations between the three latent variables are as follows:

\[ \begin{align*}
    y_{\Delta pi}^2 & \text{ 1} & y_{\Delta pi}^3 & \text{ 0.57} & y_{\Delta pi}^4 & \text{ 0.35} \\
    y_{\Delta pi}^3 & \text{ 1} & y_{\Delta pi}^1 & \text{ 0.49} & y_{\Delta pi}^4 & \text{ 1} \\
    y_{\Delta pi}^4 & \text{ 0.57} & y_{\Delta pi}^2 & \text{ 1} & y_{\Delta pi}^3 & \text{ 0.49} \\
    y_{\Delta pi}^1 & \text{ 0.35} & y_{\Delta pi}^3 & \text{ 0.49} & y_{\Delta pi}^2 & \text{ 1} \\
\end{align*} \]
The estimated correlation matrix shows that the latent trait to initially solve items correctly (y_{\Delta I}) was positively correlated with the latent trait to correct initial mistakes (y_{\Delta P}) as well as the latent trait to not change the initial right answers into incorrect ones (y_{\Delta R}). We also observe that the latent trait to make initial correct responses (y_{\Delta I}) is more highly correlated with the latent trait to correct an initial incorrect answer (y_{\Delta P}) than with the latent trait to stay with an initial correct answer (y_{\Delta R}).

In addition, we found that allowing for a different set of item parameters per node led to a better fit than assuming equal item parameters across nodes (log likelihood of $-118,512.9$, AIC of $237,203.8$, and BIC of $237,761.7$ with 89 parameters). These results support our assumption that each tree node is associated with a different set of latent trait and item parameters. Figure 3 depicts the item parameter estimates for the three nodes. The estimated item slope parameters for Node 2 tend to be greater than those for Nodes 3 and 1, while the relationship between the Node 1 and Node 3 slope parameters appears unstructured. The estimated item intercept parameters for Node 3 tend to be greater than those for Nodes 1 and 2 in that order. These results tell us that the perceived item properties can be different when test takers revisit their earlier correct or incorrect answers. Specifically, items tend to be harder when initially wrong responses were corrected (Node 2) and easier when initially correct responses were maintained (Node 3) than when the items were initially solved (Node 1).

The standard errors for the item parameter estimates tend to be smaller for the initial responses (Node 1) than for the final responses (Nodes 2 and 3). Specifically, for the item slope estimates, the standard errors ranged from .03 to .07 for Node 1, from .13 to .28 for Node 2, and from .12 to .27 for Node 3. For the item
intercept parameter estimates, the standard errors ranged from .03 to .06 for Node 1, from .14 to .45 for Node 2, and from .22 to .89 for Node 3. The smaller standard errors for Node 1 can be explained by the (by definition) larger number of observations (see matrix 4).

In summary, for Node 2, the items seem to be more difficult (based on the intercept estimates) and also more informative for the underlying Node 2 latent trait than they are for the underlying Node 3 latent trait (based on the discrimination estimates). This result suggests that answer changes from incorrect to correct (i.e., Node 2 outcomes) may provide relatively more interesting information.

3.3 Simulated Results

A simulation study was conducted to validate the utilization of our proposed model for the empirical analysis of answer change data. We first generated 100 data sets based on the proposed model with the parameter estimates reported in Section 3.2 as data generating values (with \( N = 3,900, J = 43 \)). Figure 4 summarizes the item parameters’ estimated bias for each of the three nodes. Each point in the figure represents the bias for the corresponding parameter.

The result suggests that the proposed model shows generally good parameter recovery. For the \( a \) parameters, the bias ranged from \( .010 \) to \( .010 \) for Node 1, from \( .044 \) to \( .026 \) for Node 2, and from \( .072 \) to \( .056 \) for Node 3. For the \( b \) parameters, the bias ranged from \( .008 \) to \( .009 \) for Node 1, from \( .012 \) to \( .084 \) for Node 2, and from \( .189 \) (one of the two outliers) to \( .005 \) for Node 3. For the covariance parameters (three nonredundant off-diagonal elements of the covariance matrix), the bias was estimated as \( .009, .004, \) and \( .036 \) for the \((2, 1)\), \((3, 1)\), and \((3, 2)\)th (row-column) elements, respectively. Although there seems a systematic bias for the \( b \) parameters, the size of that bias is minimal.
To further validate the use of our proposed model for the empirical data, we evaluated whether the simulated data, generated based on the estimated parameter values and latent trait estimates from the proposed model, can accurately recover some key answer change characteristics of the original data. Specifically, we evaluated the characteristics of the simulated data in terms of (1) the mean (standard deviation) of the sum scores of the initial and final answers for answer keepers, (2) the mean (standard deviation) of the sum scores of the initial answers for answer keepers, and (3) the mean (standard deviation) of the gain scores for answer changes. The original data’s answer change features were reported in Section 3.1.

Table 2 confirms that the simulated data satisfactorily mimic the main answer change features of the original data, suggesting the proposed model is a reasonable representation of the analyzed answer change data. Note that the standard deviations (but not the means) of the sum scores are slightly underestimated, which can be expected based on a small shrinkage of the latent trait estimates.

### 3.4 Directional Test Information

Next we evaluate the proposed model’s measurement precision for the three latent traits using item and test information. Information can be defined as an indicator of the degree to which the reported score from a test (or item) differentiates real differences in the latent trait of interest (Lord, 1980; Reckase, 2009). For a unidimensional two-parameter model, the item information can be expressed as a function of $\theta$ as follows:

\[
I(\theta) = \frac{\partial^2 \log P(\theta)}{\partial \theta^2} P(\theta)
\]

where $a^2 I(\theta) P(\theta)$ represents the differential variance of the item score at $\theta$. The expected information for the test is the sum of the information for each item.

---

**Table 2. Comparison of Answer Change Features Between the Original Data and the Simulated Data (Generated Based on the Proposed Model)**

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Original Data</th>
<th>Simulated Data $^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD) of sum score of the initial answers for answer changers</td>
<td>17.92 (7.33)</td>
<td>17.98 (7.23)</td>
</tr>
<tr>
<td>Mean (SD) of sum score of the final answers for answer changers</td>
<td>18.63 (7.66)</td>
<td>18.63 (7.57)</td>
</tr>
<tr>
<td>Mean (SD) of sum score of the initial answers for answer keepers</td>
<td>15.74 (7.13)</td>
<td>15.83 (6.86)</td>
</tr>
<tr>
<td>Mean (SD) of gain scores for answer changers</td>
<td>0.70 (1.63)</td>
<td>0.66 (1.45)</td>
</tr>
</tbody>
</table>

$^a$For the simulated data, the reported values are the means over 100 simulated data sets.
where \( P_i \) is the first derivative of the probability of a correct response with respect to \( y \). Test-level information \( I_\theta \) is obtained as the sum of the item-level information, that is, \( I_\theta = \sum_k I_k \). For the unidimensional case, \( I_\theta \) is equivalent to the Fisher information—the negative of expected value of the squared first derivative of the log-likelihood function \( \log P \) that is, \( I_\theta = E \frac{\partial^2 \log P}{\partial \theta^2} \), which is the reciprocal of the variance of the maximum likelihood estimate of \( y \).

For multidimensional models, we can define the item information matrix, following Segall (1996), with \( k \)th diagonal elements

\[
I_{ii} = \frac{q_i \cdot \partial^2 P}{P_i \cdot \partial \theta \partial \theta} ;
\]

and \((k, l)\) off-diagonal elements

\[
I_{il} = \frac{q_i q_l \cdot \partial^2 P}{P_i \cdot \partial \theta \partial \theta} ;
\]

Here we focus on the diagonal elements in Equation 15 and express them using directional derivatives, following Reckase (2009, p. 121):

\[
I_{ii} = \frac{\partial^2 \log P}{P_i \cdot \partial \theta \partial \theta} \frac{\partial \theta}{r \cdot \partial \theta} \cdot \cos n_k ;
\]

where \( r_n \) denotes the directional derivative in the direction of the vector of angles \((\frac{\partial \theta}{\partial \theta} \frac{\partial \theta}{\partial \theta} \cdots \frac{\partial \theta}{\partial \theta} \frac{\partial \theta}{\partial \theta})\) with the coordinate axes that define the direction taken from the \( \theta \) point. The directional derivative can be given as \( r_n \frac{\partial \theta}{\partial \theta} \frac{\partial \theta}{\partial \theta} \cos n_k \). Then, Equation 17 can be simplified as follows (Reckase, 2009, p. 122):

\[
I_{ii} = \frac{\partial^2 \log P}{P_i \cdot \partial \theta \partial \theta} \cdot \cos n_k ;
\]

where \( a_{ik} \) is the slope parameter for item \( i \) in dimension \( k \). The partial derivative with respect to \( y_k \) is equivalent to the derivative to the direction of the \( k \)th axis, in which we have \( \cos n_k = 1 \). That is, in the direction of axis \( k \), Equation 18 can be written as:

\[
I_{ii} = \frac{\partial^2 \log P}{P_i \cdot \partial \theta \partial \theta} \cdot \cos n_k ;
\]
and the directional test-level information can be defined as a sum over item specific directional information as follows:

$$I_{ik}^n = \sum_i I_{ik}^n$$

In effect, Equation 20 can be seen as the $k$th diagonal element of the Fisher information matrix defined in Equation 15 (which is the reciprocal of the variance of maximum likelihood estimate of $y_k$).

We plot a unidimensional slice of the test information function (20) for each $y_{ik}$ axis when the other $y_{ij}$ ($j \neq k$) are fixed to 0 and 1 and 0, respectively. Figure 5 shows the unidimensional information curve as a function of $y_{ik}$.

The figure suggests that each latent trait shows maximum information at somewhat different locations of the respective latent score continuum. For $y_{ik}$
(the latent trait for the initial responses), more information is observed around mid-latent score levels than for extreme latent score levels when the \( y_{\text{ebp}} \) and \( y_{\text{dp}} \) values are 0. The peak point of each curve changes depending on the \( y_{\text{ebp}} \) and \( y_{\text{dp}} \) values. For \( y_{\text{ebp}} \) (the latent trait to correctly change initially incorrect answers), all information curves increase exponentially and the peak is achieved for high latent trait levels. For \( y_{\text{ebp}} \) (the latent trait to maintain initial correct answers), information curves tend to decrease as the latent trait values increase.

3.5 Two Potential Uses of the Proposed Model

We apply the proposed model to compute the answer change benefits and to detect irregular answer change patterns. We compare the results from the general model (with three node-specific latent traits) with the simpler model that assumes a single latent trait (but a different set of item parameters) across nodes.

3.5.1 Computing answer change benefits. Based on the general model, the expected benefit of answer change can be defined by using expected scores for the initial responses \( \Delta y_{\text{ebp}} \) and expected scores for the final responses conditional on the initial responses \( E y_{\text{ebp}} \). The initial and final expected scores can be expressed as a function of latent traits, given the observed responses as follows:

\[
E y_{\text{ebp}} \quad y_{\text{ebp}} = \frac{\theta_1}{4} \text{Pr} \theta_{\text{ebp}} \quad \frac{1}{4} \text{Pr} \theta_{\text{ebp}}
\]

Then the expected benefit of answer change \( B_{\text{ebp}} \) can be defined as the difference between the initial and final (and conditional) expected scores:

\[
B_{\text{ebp}} = \frac{1}{4} \left( \text{Pr} \theta_{\text{ebp}} \quad \frac{1}{4} \text{Pr} \theta_{\text{ebp}} \right) - \left( \frac{1}{4} \text{Pr} \theta_{\text{ebp}} \quad \frac{1}{4} \text{Pr} \theta_{\text{ebp}} \right)
\]

This item-level expected benefit can be plotted as a surface function in a three-dimensional \( \theta \)-grid, which is hard to visualize. We instead plot the expected benefit as a function of each latent trait \( y_{\text{ebp}} \) given fixed values of other \( y_{\text{ebp}} \) \( k \theta \) where \( \theta \) were generated from a multivariate normal distribution with the mean of 0 and the estimated covariance matrix (12). Figure 6 shows
FIGURE 6. Item-level expected benefits as a function of $y_{\text{DP}}^k (k \notin 1, 2, 3)$ for the first 3 items (given other $y_{i,j,k}^\text{DP} (k \notin \{i, j, k\})$ fixed to values generated from a multivariate normal distribution).

A 3 × 3 plot in which the column represents Nodes 1, 2, and 3, respectively, and the row represents Items 1, 2, and 3, respectively.

In the first subplot for $y_{\text{DP}}^k$, a triangular pattern is observed, suggesting that low ability examinees derive no clear benefit from answer change, while examinees with higher abilities can but do not necessarily benefit from answer change. In the second subplot for $y_{\text{DP}}^k$, we can see that the expected benefits increase (sharply) at some point of the latent trait $y_{\text{DP}}^k$. In the third plot, the expected benefits increase (slowly) as $y_{\text{DP}}^k$ increases.

When the local independence assumption holds (a common assumption for IRT models), the test-level expected benefits $\theta_{\text{DP}}^P$ can be computed as the sum of the item-level expected benefits. Figure 7 shows the test-level
expected benefits as a function of $y^{k\Phi}(k \neq 1; 2; 3)$ when other $y^{k\Phi}(k^0 \Phi k)$ fixed to values generated from a multivariate normal distribution.

The test-level expected benefits show a similar pattern to that of the item-level expected benefits. Observe that a number of students do derive a negative benefit of answer change across large ranges of $y^{k\Phi}(k \neq 1; 2; 3)$.

We compute the estimated answer change gains by plugging the estimated latent scores $\hat{\theta}_p$ in Equation 23 ($p \neq 1; \ldots; N$ and $\hat{\theta}_p$ are estimated based on the general model 10). As a comparison, we also compute the estimated benefits from the simpler model assuming a single latent trait over the three nodes. Figure 8 compares the distributions of the estimated benefits over persons between the two models.

The estimated benefits $\hat{B}_{\Phi p}$ based on the general model showed a median of .1 and a standard deviation of .22. About 21% of the students showed negative benefits from changing their initial answers. The estimated benefits $\hat{B}_{\Phi p}$ based
on the simple model assuming a single latent trait showed a median of .26 and a
standard deviation of .29. Interestingly, the general model, which is the better
fitting model, shows less benefits than the one-dimensional model does.

3.5.2 Detecting irregular answer change patterns. van der Linden and Jeon
(2012) outlined a statistical procedure for detecting potential academic cheating.
Specifically, an unusually large number of WR answer change can be identified
comparing a theoretical distribution of a WR change with the actual occurrence of
WR changes. The theoretical distribution of WR can be constructed as independent
Bernoulli trials with a probability $\Pr(\omega|\theta) \propto \Pr(\theta|d_p) \propto \prod \Pr(\theta|d_{p_i}), \gamma(\delta|\gamma, \theta) \propto \gamma(0|\gamma)$. The
resulting distribution is a generalized or compound binomial (Lord, 1980, section
4.1). The critical value $c_p$ is then found satisfying:

$$\Pr(\omega|\theta) \leq \gamma(c_p)$$

where $\gamma$ is the predefined significance level (e.g., a $\gamma = .001, .01, .05$). The critical
value $c_p$ is the smallest value of $\theta$ for which the area in the upper tail of the
generalized binomial distribution is not larger than $\gamma$. Person $p$ is flagged as a
potential cheater if $\omega_p$ is greater than the critical value $c_p$ at the corresponding
significance level $\gamma$.

We applied this statistical test described above to detect irregular WR change
patterns. Based on the general model, no one was flagged as possibly cheating at
three different significance levels (a \(< .001, .01, .05\)). However, based on the simpler model assuming a single latent variable, we found that several students were flagged (77, 46, and 23 students at a \(< .05, .01, .001\), respectively). This result is not unexpected, given that a WR-specific latent variable \(y_{p}^{\text{WR}}\) accounts for the probability of a WR change in the general model, whereas in the simpler model, a global latent trait \(\theta_{p}\) is used. The result suggests that using a WR-specific latent variable \(y_{p}^{\text{WR}}\) gives a more conservative result on identifying potential cheating than using the global latent trait \(\theta_{p}\).

From a cheating detection perspective, \(y_{p}^{\text{WR}}\) could be a cheating latent variable, indicating to what extent subjects change incorrect responses to correct responses. It is impossible to test whether such changes stem from cheating or from a genuine latent trait. However, the positive correlation between \(y_{p}^{\text{WR}}\) and \(y_{p}^{\text{WR}}\) (57) suggests that \(y_{p}^{\text{WR}}\) is not a cheating trait.

This analysis shows that the identification of cheaters depends on the model that is used and also on the interpretation of a model component (in this case, the Node 2 latent variable). The former issue (model dependence) underlines the importance of model validity and the second issue (model interpretation) illustrates that one cannot automatically jump from a statistical result to a conclusion regarding the intention of a test taker.

4. Discussion

Examinees’ answer change behavior has intrigued educational researchers for quite some time. In recent years, the increased utilization of computerized adaptive testing, which provides easier access to answer change data, has motivated researchers to undertake a more rigorous analysis on answer change.

In this article, we proposed a novel statistical approach for modeling the examinees’ answer change process. The proposed approach was based on a generalized item response tree model, utilizing a tree structure to represent a sequence of activities that occur during a test taker’s answer change process. Specifically, the process of the initial answer is modeled as the top node in the tree, the process of changing (or not changing) one’s answer is represented with the intermediate nodes, and the outcomes of the sequential process are the leaves (or terminal nodes) of the tree. Each node is allowed to have its own set of item parameters and its own latent trait.

The results of an empirical application showed that each tree node involved a different set of item parameters and latent trait, confirming that the decision made at each node was governed by a distinct, decision-making process. For instance, the difficulty level of items was generally higher when initially wrong answers were revised to a correct answer (Node 2) than when initially correct answers were maintained (Node 3). The fact that the item parameter estimates
differ depending on the node suggests that the item properties can be perceived differently when the initial responses and the final responses (based on the initial responses) are produced. This has an important implication for the practice of item analysis; that is, the standard assumption of measurement invariance may not hold (within subjects) when answer change is possible and actually occurs during the test.

In addition, we found that the latent trait to correct initially wrong answers (Node 2) and the latent trait to correctly maintain initially right answers (Node 3) were only moderately correlated with the latent trait underlying the initial answers (with correlation of .57 and .35, respectively). The Nodes 2 and 3 latent traits were also moderately correlated with each other (correlation of .49). The Nodes 2 and 3 latent traits are the “change” traits that are involved in making changes to initial answers, while the Node 1 latent variable is the regular ability trait that underlies the initial answers.

It is noteworthy that existing methods that have been utilized for studying answer change have neglected the possibility that multiple latent traits can be involved if the possibility of answer change exists. By utilizing a tree approach that can model a sequential decision-making process, we were able to differentiate the change traits from the initial answer latent trait. Commonly, answer changes are ignored and the final responses are used to estimate the ability. Our results show that in the presence of answer change the final responses reflect multiple latent traits, and without the possibility of answer change a somewhat different ability is being measured. We also found that the ability underlying the initial answers is positively correlated with the ability to correct one’s incorrect responses and the ability to stay with one’s correct answers, while the latter two abilities are different from the former and codetermine the final response. This differentiation is a unique and highly relevant finding that has not been discussed previously in the literature. The correction ability and the ability of not being distracted by the possibility to change one’s initial answer are potentially broader abilities that may play a role in other situations as well. For example, being reflective about one’s initial understanding and a willingness to give a second thought to that earlier understanding may tell us about some aspects of students’ learning styles. This is of course only a hypothesis that can be further investigated.

Note that test takers’ answer change behavior has typically been ignored in analyzing item response data, although answer change is a natural phenomenon that commonly occurs in testing situations. Our study showed that focusing on the ability estimated from item response data (without considering answer change) may prevent us from finding interesting other sources of ability-related information contained in the answer change latent traits.

Another contribution of the present study to educational measurement is to provide a unique application of the item response tree approach (e.g.,
Boeckenholt, 2012, 2013; De Boeck & Partchev, 2012; Jeon & De Boeck, 2015). By taking advantage of the flexibility of the generalized item response tree modeling framework, we can extend the proposed model to deal with more complex answer change scenarios. For instance, it is possible to take into account multiple answer changes by including extra nodes for such items. In this case, depending on how many changes are made to each item, the size (or the number of nodes) of a tree will vary across items. We can also extend the model for polytomous item responses by allowing for the top node to have as many branches as the number of response options. The node outcomes can then be represented with a polytomous item response model, such as a partial credit model, a graded response model, or a nominal response model. Furthermore, person-level covariates can be incorporated in the tree model in order to test whether latent trait levels can be explained with relevant person characteristics (e.g., gender, family background, cognitive styles, confidence). For instance, the ability to correctly revise initial wrong answers (latent trait for Node 2) may depend on students’ field dependency, while the ability to keep initially correct answers (latent trait for Node 3) can be explained by their confidence level.

Finally, although the proposed approach was illustrated within the context of educational testing, the model can also be applied to explain answer change behavior in psychological and behavior assessments more generally. For instance, in answering sensitive survey items, some people try to revise their earlier answers to present themselves more positively. The proposed model can be applied to capture subjects’ answer change behavior in relation to their social desirability tendency.

Note

1. For example, with the empirical data utilized in Section 3, the guessing parameter estimates for Nodes 2 and 3 were close to zero for most items, while for Node 1, some items showed nonzero guessing parameter estimates although their standard errors were quite large. Based on further investigations, however, we concluded not to include guessing parameters for Node 1 either.

References


**Authors**

MINJEOG GEON is an assistant professor at University of California, Los Angeles, 457
Portola Avenue, Los Angeles, CA 90024; email: mijjeon@ucla.edu. Her research inter-
est include latent variable modeling, multilevel modeling, item response modeling,
and research methods.

PAUL DE BOECK is a professor at the Ohio State University, 232 Lazenby Hall, 1827
Neil Avenue, Columbus, OH 43210; email: deboeck.2@osu.edu. His research interests
are explanatory item response models and psychometric models for cognitive and non-
cognitive test data in general.

WIM VAN DER LINDEN is a Distinguished Research Scientist and Director of Research
& Innovation at Pacific Metrics Corporation, 1 Lower Ragsdale Bldg. 1 Ste 150,
Monterey, CA 93940; email: wvanderlinden@pacificmetrics.com. His research inter-
est include test theory, applied statistics, and research methods.

Manuscript received June 4, 2016
Revision received November 18, 2016
Accepted December 13, 2016