Behav Res (2016) 48:1070–1085
DOI 10.3758/s13428-015-0631-y

A generalized item response tree model for psychological assessments

Minjeong Jeon[1] · **Paul De Boeck**[1]

Published online: 25 July 2015
© Psychonomic Society, Inc. 2015

Abstract A new item response theory (IRT) model with a tree structure has been introduced for modeling item response processes with a tree structure. In this paper, we present a generalized item response tree model with a flexible parametric form, dimensionality, and choice of covariates. The utilities of the model are demonstrated with two applications in psychological assessments for investigating Likert scale item responses and for modeling omitted item responses. The proposed model is estimated with the freely available R package *flirt* (Jeon et al., 2014b).

Keywords IRT · Tree · IRTree · Item response process · Likert scale · Omitted responses · *flirt*

Introduction

Item response theory (IRT) models are widely used tools for analyzing categorical item responses in psychological and behavioral assessments. IRT models focus on understanding the terminal outcome of a person's choice among several discrete options. Mathematically, the probability of selecting a particular response category can be explained as a function of the person's latent trait and the item's properties.

In this study, we are concerned with a new type of IRT model that focuses not only on the outcome but also on the internal cognitive or psychological decision process. The model can describe a postulated internal decision process with a tree structure, which is composed of sub-trees and their corresponding internal nodes and branches. The tree continues to diverge through branches until it reaches leaves. The leaves are the terminal nodes that represent the observed categorical item responses. The model will be referred to as an item response tree model due to its utilization of the tree structure (e.g., Boeckenholt, 2012; De Boeck & Partchev, 2012).

Figure 1 illustrates four tree structures that can be used to represent different cognitive processes for an item with four response categories (numbered from 1 to 4).

In a tree structure, circles represent nodes, arrows represent branches, and leaves are item response outcomes (1 to 4). Tree (a) represents a sequential selection of the response in order from Categories 1 to 4, while Tree (b) describes a two-stage selection process where a group of two adjacent categories is first chosen (either Categories 1 and 2 or Categories 3 and 4) and then the final answer is selected within the pair of adjacent categories. In Tree (c), the selection of Category 1 is qualitatively differentiated from the other three Categories (2, 3, and 4) and not choosing Category 1 requires a follow-up decision. In Tree (d), Categories 1 and 2 are two qualitatively distinct options, which are also differentiated from Categories 3 and 4. The selection between 3 and 4 involves a second decision.

Tree (a) is denoted as a 'linear' tree in that at least one branch from each internal node leads to a terminal node, whereas Tree (b) is denoted as a 'nested' tree since branches from an internal node lead to another internal node. Trees (c) and (d) can be seen as a combination of linear and nested trees. Trees (a) and (b) are 'binary' trees because they involve a choice between two branches, whereas Trees

✉ Minjeong Jeon
jeon.117@osu.edu

Paul De Boeck
deboeck.2@osu.edu

[1] Department of Psychology, Ohio State University, 1827 Neil Avenue, Columbus, OH 43210, USA

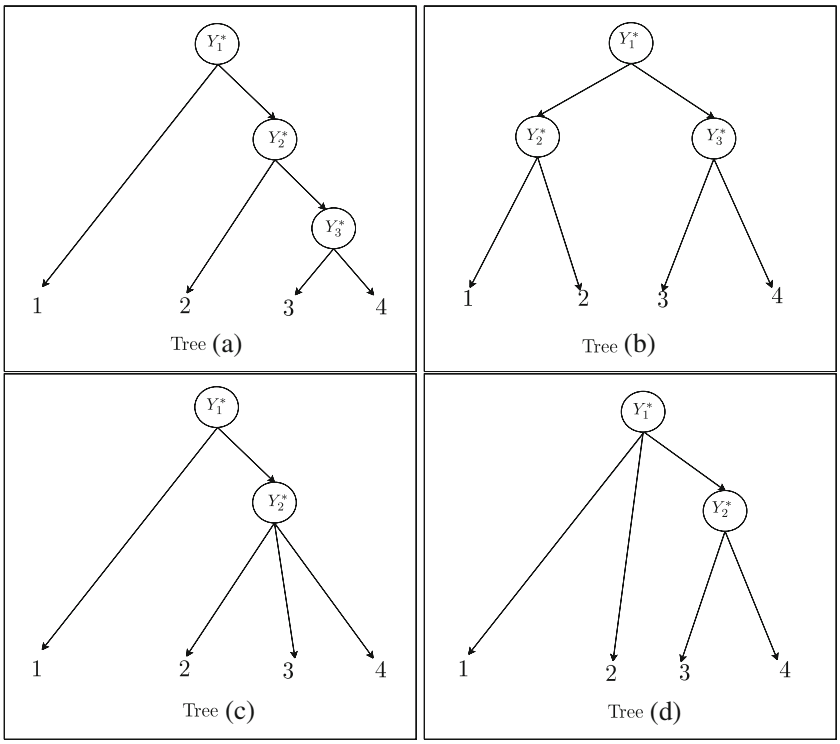


Fig. 1 Four tree structures (**a**) to (**d**) for four-category observed responses (1, 2, 3, and 4). Y_1^* to Y_3^* represent internal nodes 1 to 3, respectively

(c) and (d) are 'polytomous' trees in that more than two branches are involved.

Item response trees can also be utilized to describe distinctive features of item response categories. For instance, Tree (a) in Fig. 1 can be used to describe a unipolar scale that includes choices such as 'not at all sad', 'slightly sad', 'mostly sad', and 'completely sad' while Tree (b) is utilized to describe a bipolar scale with options such as 'completely sad', 'somewhat sad', 'somewhat joyful', and 'completely joyful'. To describe Likert scales with a middle response category (e.g., 'neither sad nor joyful', 'perhaps', 'not sure', 'undecided', '?', etc.), Tree (c) can be used where the first outcome 1 represents the undecided (middle) category while Categories 2 to 4 represent regular response categories. It should be noted that in the current practice of item analysis, the differences between various response/test formats are usually ignored and the responses are analyzed as if they were interval scale data or ordinal scale data (e.g., using ordinal factor/IRT models).

Differentiating slow and fast intelligence (Partchev & De Boeck, 2012), modeling motivated misreports to sensitive survey questions (Boeckenholt, 2013), examining content and response styles in multiple-choice items (Plieninger & Meiser, 2014), and modeling skipped and non-reached item responses (Debeer et al., submitted) are some of the examples of how item response tree models have been utilized. These prior approaches, however, have some limitations. For instance, the models proposed by De Boeck and Partchev (2012) do not allow for trees with more than two branches and rely on one-parameter logistic models, while Bockenholt (2012) is based on one-parameter probit models.

The purpose of this article is to present a generalized item response tree modeling framework that is flexible in several aspects: specifically, by utilizing nodes as building blocks, a node-specific, possibly different IRT model can be specified at each node. Latent variables can be unidimensional or multi-dimensional and can be node-specific or shared between nodes. Similarly, item parameters can also be node-specific or shared across nodes. Explanatory variables for persons and items can be incorporated to investigate sources of heterogeneity in a response scale. We will show that the proposed framework can be readily estimated with a standard IRT software package.

We will provide two examples to demonstrate the utilities of the proposed modeling framework: (1) for investigating the characteristics of three-point and four-point Likert scale items and (2) for examining response omission behaviors in psychological assessments. Note that not all the variants of the general model and all the aspects of the specific illustrated models will be discussed in this paper.

The outline of this paper is as follows: In the "Models" section, we present the general tree modeling framework with examples and extensions. In the "Estimation, software, and model selection" section, we describe a software package that is used for estimation and discuss model selection.

We provide a few case studies in the "Applications" section. We conclude with some discussion in the "Discussion" section.

Models

We will begin with an illustrative example of a generalized item response tree model. We then provide a general model formulation, describe its various extensions, and discuss several related models.

Example

Suppose there is a Likert scale with items that include three response categories: 'No', 'Perhaps', and 'Yes'. Let us assume that a particular response category is the outcome of the following two-stage decision process: (1) in Stage 1, the person decides on the certainty of the answer (i.e., perhaps), and (2) in Stage 2, the person decides on the direction of the answer (i.e., agree or disagree). The terminal nodes are the result of the sequence of the Stage 1 and Stage 2 decision process. Figure 2 illustrates the tree structure that represents this two-stage decision process.

The specified tree starts from the initial internal node Y_1^*, which is related to the certainty on item i for person p. If the person is uncertain, the left branch of the first sub-tree is chosen, which leads to the terminal outcome 'Perhaps'. If the person is certain, the right branch is chosen, which leads to the next sub-tree with the second internal node Y_2^*. The second internal node Y_2^* represents the direction of the decision, i.e., agree or disagree.

Within a sub-tree, choosing a branch can be parameterized with an IRT model; in other words, the probability of selecting a branch is expressed as a function of the person's latent trait (related to the choice of a branch) and the item parameters. The latent variable can also be seen as a source of heterogeneity between subjects in the corresponding decision (e.g., some people more often choose the right branch than the left branch while other people more often choose the left branch than the right branch).

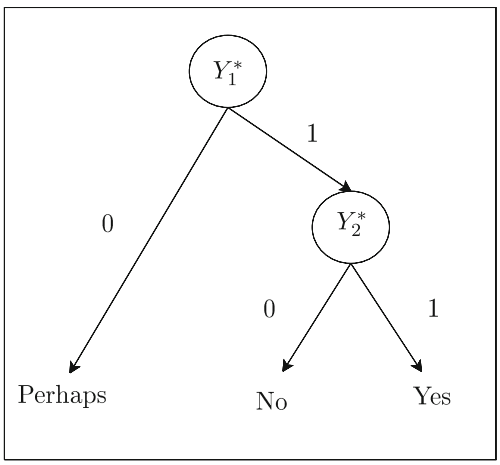


Fig. 2 Item response tree for person p to item i for three response categories ('Perhaps', 'No', and 'Yes')

Let Y_{pi1}^* denote the choice of a left or right branch at Node 1 for person p ($p = 1, ..., N$) to item i ($i = 1, ..., I$). Suppose the left and right branches are represented by 0 and 1, respectively. The probability of choosing the right branch ($Y_{pi1} = 1$) is then modeled with a regular two-parameter item response model as follows:

$$\Pr(Y_{pi1}^* = 1|\theta_{p1}) = g^{-1}\left(\alpha_{i1}\theta_{p1} + \beta_{i1}\right), \quad (1)$$

where θ_{p1} represents person p's latent trait that involves in choosing the right branch at Node 1 (indicating "certainty"), g^{-1} is the inverse of the link function (which is typically a logit or probit link for a binary choice), and α_{i1} and β_{i1} are the slope (or discrimination) and the intercept (or easiness) parameters, respectively, for item i at Node 1. The slope parameters α_{i1} can be interpreted as the sensitivity or relevance of item i to the probability of choosing the right branch at Node 1 and the intercept parameter β_{i1} can be interpreted as the degree of easiness in choosing the right branch for item i at Node 1.

At Node 2, Y_{pi2}^* represents whether person p chooses 'Yes' (right branch) or 'No' (left branch). Note that the decision at Node 2 (Y_{pi2}^*) is conditional on the decision at Node 1. That is, only when $Y_{pi1}^* = 1$, the outcome at Node 2 Y_{pi2}^* is observed. Suppose the left and right branches at Node 2 are represented with 0 and 1, respectively. The conditional probability of choosing the right branch given $Y_{pi1}^* = 1$ is then modeled with a two-parameter item response model as follows:

$$\Pr(Y_{pi2}^* = 1|\theta_{p2}) = g^{-1}\left(\alpha_{i2}\theta_{p2} + \beta_{i2}\right), \quad (2)$$

where θ_{p2} is the person p's latent trait that involves in choosing 'Yes' rather than 'No' (indicating admission or negation, respectively), and α_{i2} and β_{i2} are the item slope and intercept parameters for item i at Node 2. For notational simplicity, that the probability of $Y_{pi2}^* = 1$ is conditional on the earlier decision $Y_{pi1}^* = 1$ and the associated latent trait θ_{p1} at Node 1 is omitted in Eq. 2.

The model formulation with Eqs. 1 and 2 is based on two important assumptions: (1) internal node outcomes (conditional on the earlier decisions and on the latent variables involved) are independent of each other, and (2) only one path is allowed to be associated with an observed outcome. These two assumptions imply that each observed outcome ('Perhaps', 'No', 'Yes') is the result of a unique sequence of conditionally independent internal decisions.

The probability of an observed response ('Perhaps', 'No', 'Yes') can now be computed as the product of the conditional probabilities of the internal decisions that are involved in the path to the observed outcome. Suppose the

three observed outcomes Y_{pi} ('Perhaps', 'No', and 'Yes') are coded as 1, 2, and 3, respectively. The probability of observing each outcome can then be expressed as follows:

$$\Pr(Y_{pi} = 1|\theta_{p1}) = \Pr(Y^*_{pi1} = 0|\theta_{p1}), \quad (3)$$

$$\Pr(Y_{pi} = 2|\theta_{p1}, \theta_{p2}) = \Pr(Y^*_{pi1} = 1|\theta_{p1})\Pr(Y^*_{pi2} = 0|\theta_{p2}), \quad (4)$$

$$\Pr(Y_{pi} = 3|\theta_{p1}, \theta_{p2}) = \Pr(Y^*_{pi1} = 1|\theta_{p1})\Pr(Y^*_{pi2} = 1|\theta_{p2}), \quad (5)$$

where $\Pr(Y^*_{pi1} = 0|\theta_{p1}) = 1 - \Pr(Y^*_{pi1} = 1|\theta_{p1})$, $\Pr(Y^*_{pi2} = 0|\theta_{p2}) = 1 - \Pr(Y^*_{pi2} = 1|\theta_{p2})$, and $\Pr(Y^*_{pi1} = 1|\theta_{p1})$ and $\Pr(Y^*_{pi2} = 1|\theta_{p2})$ are specified in Eqs. 1 and 2, respectively.

The following mapping matrix T streamlines how each observed outcome Y_{pi} is related to the internal decisions Y^*_{pi1} and Y^*_{pi2} at Nodes 1 and 2:

	Y^*_{pi1}	Y^*_{pi2}
$Y_{pi} = 1$	1	NA
$Y_{pi} = 2$	0	0
$Y_{pi} = 3$	0	1

(6)

For instance, $Y_{pi} = 1$ corresponds to $(Y^*_{pi1}, Y^*_{pi2}) = (1, NA)$, where NA represents a missing observation (recall that $Y_{pi} = 1$ does not involve Node 2 in the path).

General model formulation

Here we provide a general formulation of the item response tree model with K nodes for M terminal observed outcomes. The mapping matrix T is of size $M \times K$ whose (m, k)-th element T_{mk} $(m = 1, ..., M, k = 1, ..., K)$ represents the outcome at the internal node k (that is associated with the mth observed outcome). When Node k includes L branches, T_{mk} take values 0, 1, 2, ..., $(L - 1)$ (and NA when Node k does not appear in the path to terminal observed outcome m).

The conditional probability of internal outcome T_{mk} at Node k (given the earlier internal outcomes and latent traits involved) can be formulated as follows:

$$\Pr(Y^*_{pik} = T_{mk}|\theta_{pk}) = g^{-1}\left(\alpha_{ik}\theta_{pk} + \beta_{ik}\right), \quad (7)$$

where θ_{pk} is the latent variable for person p at Node k, and α_{ik} and β_{ik} are the item slope and intercept parameters for item i at Node k. When Node k includes two branches, the link function $g(\cdot)$ can be a logit or probit function as discussed in the "Example" section (we utilize a logit link for our analysis later). When Node k involves more than two branches, a different link function can be specified, such as an adjacent logit or a cumulative link function (that are used to formulate a generalized partial credit model (Muraki, 1992) and a graded response model (Samejima, 1969), respectively). Equation 7 assumes a single latent trait θ_{pk} at Node k. In principle, however, a node can be multidimensional, that is, Node k can involve more than one latent variable. A multidimensional extension per node is straightforward, but we limit our discussion here to a unidimensional node (a single latent trait per node) for illustrative simplicity.

Using the conditional probabilities of internal outcomes $Y^*_{pik} = T_{mk}$ (7), the model for observed terminal outcome $Y_{pi} = m$ $(m = 1, ..., M)$ can be formulated as follows:

$$\begin{aligned}
&\Pr(Y_{pi} = m|\theta_{p1}, ..., \theta_{pK}) \\
&= \Pr(Y^*_{pi1} = T_{m1}, ..., Y^*_{piK} = T_{mK}|\theta_{p1}, ..., \theta_{pK}), \\
&= \prod_{k=1}^{K} \Pr(Y^*_{pik} = T_{mk}|\theta_{p1}, ..., \theta_{pK})^{t_{mk}}, \quad (8)
\end{aligned}$$

where $t_{mk} = T_{mk}$ if $T_{mk} = 0$ or 1 and $t_{mk} = 0$ if $T_{mk} =$ NA $(k = 1, .., K, m = 1, ..., M)$. The K latent variables $\boldsymbol{\theta}_p = (\theta_{p1}, ..., \theta_{pK})'$ are assumed to follow a multivariate normal distribution with $\boldsymbol{\theta}_p \sim N(\mathbf{0}, \Sigma)$, where Σ is a $K \times K$ covariance matrix. That is, the K node-specific latent traits are allowed to be correlated with each other.

It is important to note that model (8) is equivalent to a K-dimensional item response model fitted to the node-specific internal outcomes $(Y^*_{pi1}, Y^*_{pi2}, ..., Y^*_{piK})$ (of length K). For clarification, we re-write model (8) for $\nu_{pik} = g(\Pr(Y^*_{pik} = 1|\theta_{pk}))$ $(k = 1, 2, ..., K$ and $g(\cdot)$ is the link function) in a matrix form as follows:

$$\begin{bmatrix} \nu_{pi1} \\ \nu_{pi2} \\ \vdots \\ \nu_{piK} \end{bmatrix} = \begin{bmatrix} \alpha_{i1} & 0 & \cdots & 0 \\ 0 & \alpha_{i2} & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \alpha_{iK} \end{bmatrix} \begin{bmatrix} \theta_{p1} \\ \theta_{p2} \\ \vdots \\ \theta_{pK} \end{bmatrix} + \begin{bmatrix} \beta_{i1} \\ \beta_{i2} \\ \vdots \\ \beta_{iK} \end{bmatrix}. \quad (9)$$

For all item responses $(i = 1, ..., I)$, the model includes I items in each of the K dimensions (nodes) and each dimension involves latent trait θ_{pk}. There are no cross-loadings between dimensions; thus, model (6) is equivalent to a simple-structure K-dimensional item response model.

Hence, the model is identified by complying with conventional identification constraints for simple-structure multidimensional IRT models. Specifically, we apply constraints on the latent distribution, $\boldsymbol{\theta}_p \sim N(\mathbf{0}, \Sigma)$, where the means are fixed to 0 and the variances (diagonal elements of Σ) are fixed to 1 (so that all loading parameters can be freely estimated).

In the next four sub-sections ("A bifactor structure" to "Item and person covariates"), we will illustrate how the general model (8) can be modified with a bifactor structure, node-specific main effects, simplified latent structures, and item- and person- covariates.

A bifactor structure

The general model formulation (7) presumes that each internal node involves a node-specific latent trait, implying that

the response scale itself is assumed to be multidimensional. This creates the problem of how a score can be determined for the latent trait one intends to measure.

If the nodes are only partly specific and share a common latent variable, then a bifactor structure may help. For instance, in a depression scale item with four options, 'not at all sad', 'slightly sad', 'mostly sad', and 'completely sad', the common latent trait (or the general factor) would represent the degree of sadness (or depression), which is intended to be measured with this depression scale. If one is willing to impose an assumption that node-specific latent traits are independent of each other given the general latent trait, the model for internal outcome (7) can be modified with a *bifactor structure* as follows:

$$\Pr\left(Y^*_{pik}=1|\theta_{pk}\right) = g^{-1}\left(\alpha^g_{ik}\theta^g_p + \alpha_{ik}\theta_{pk} + \beta_{ik}\right), \quad (10)$$

where α^g_{ik} is the slope parameter for item i at Node k, which is associated with the general latent trait θ^g_p for person p, and α_{ik} is the node-specific slope parameter for item i at Node k for latent trait θ_{pk}. With this bifactor version, $K+1$ latent traits are specified and they are assumed to follow a multivariate standard normal distribution with a diagonal covariance matrix (that is, all latent variables are assumed to be independent of each other).

Node-measurement invariance and node-main effects

The general model formulation (7) allows for a set of node-specific item parameters (α_{ik} and β_{ik}). When node-specific item parameters are different across nodes, it means that items show different measurement properties depending on the nodes. That is, measurement is not invariant across nodes or 'node-measurement invariance' does not hold.

Node-measurement invariance is empirically testable. A simple way for testing node-measurement invariance is to exploit the 'main' effects for nodes. Specifically, we can decompose node-specific item parameters into common item parameters (equal across nodes) and node-main effects parameters (equal across items). The item parameters β_{ik} and α_{ik} for item i at Node k can be decomposed as follows:

$$\beta_{ik} = \beta_i + \delta_{\beta_k}, \quad (11)$$

$$\alpha_{ik} = \alpha_i + \delta_{\alpha_k}, \quad (12)$$

where β_i and α_i are the common parameters for item i, while δ_{β_k} and δ_{α_k} are the node-main effects for the item intercept and slope parameters, respectively, at Node k. If the node-specific item parameters can be reduced to the sum of the common item parameters and node-main effects, then the item parameters are the same across nodes and Node k creates an impact to the latent trait distribution. Specifically, significant values of δ_{β_k} and δ_{α_k} represent non-ignorable impacts to the mean and variance of the distribution of θ_{pk}, respectively. One can then conclude that node-measurement invariance does hold for item i (because differences are in the distribution, not in the item parameters). However, if the node-specific item parameters cannot be reduced into the main effects of the node (i.e., the model with node-specific item parameters (β_{ik}, α_{ik}) fits better than the simplified model with the node-main effect parameters and common item parameters (β_i, δ_{β_k}, α_i, δ_{α_k})), then measurement invariance does not hold with the response scale.

Item parameter decomposition shown in Eqs. 11 and 12 can also be used for parameter reduction. (e.g., $I \times K$ item slope parameters are reduced to $I + K$ parameters). A similar idea is used in formulating the rating scale model to simplify item response category threshold parameters in the partial credit model.

Simplifying latent structures

Suppose some (or all) of the node-specific latent traits are not differentiable from one another. We then can simplify the latent structure of the model by collapsing those latent variables. Suppose all K latent traits are perfectly correlated with each other; that is, a single general latent variable is sufficient to explain the mechanism of choosing different response categories. The general model (7) can then be simplified as

$$g(\Pr(Y^*_{pik}=1|\theta_p)) = \alpha_{ik}\theta_p + \beta_{ik}. \quad (13)$$

Note that the latent trait θ_p no longer includes subscript k, meaning that there is a single latent trait across nodes. The unidimensional latent trait θ_p is assumed to follow a standard normal distribution. In the modified model (13), the item parameters α_{ik} and β_{ik} are still node-specific; but they can also be simplified using the node-main effects parameters and common parameters as discussed in the "Node-measurement invariance and node-main effects" section.

Item and person covariates

If one is interested in explaining variability in the item parameters and latent traits across or within nodes, it is useful to introduce item and person covariates in the model. By incorporating node-specific explanatory variables for persons and items, model (7) can be modified as follows:

$$g(\Pr(Y^*_{pik}=1|\theta_{pk})) = \sum_q^Q \alpha_{kq}X_{ikq}\theta_{pk} + \sum_l^L \beta_{kl}W_{ikl}, \quad (14)$$

$$\theta_{pk} = \sum_m^M \gamma_{km}Z_{pkm} + \theta'_{pk},$$

where α_{kq} is the regression coefficient for the qth covariate X_{ikq} ($q = 1, ..., Q$) that explains the slope parameters in internal node k, β_{kl} is the regression coefficient for the lth covariate W_{ikl} ($l = 1, ..., L$) that explains the intercept parameters in Node k, and γ_{km} is the regression coefficient for the mth covariate Z_{pkm} ($m = 1, ..., M$) in Node k. Note that when person covariates are used, θ'_{pk} no longer indicates person p's latent trait at Node k; instead, θ'_{pk} is the residual of the person trait that is not explained by the covariates. Therefore, when the measurement of the latent trait is the main goal of the data analysis, person covariates may not be used.

Model (14) can be simplified by constraining (some or all) node-specific regression coefficients to be equal across nodes. For instance, shared covariate effects across nodes can be imposed for all parameters with the following constraints: $\alpha_{kq} = \alpha_{k'q}$, $\beta_{kl} = \beta_{k'l}$, and $\gamma_{km} = \gamma_{k'm}$ for $k \neq k'$.

Related models

Item response tree models are characterized with two main features: (1) the models are formulated based on a tree structure and (2) the models allow for multiple sources of individual differences for a response scale.

In cognitive psychology, multinomial processing tree models utilize a tree structure in model formulation. The models have been widely used for assessing the cognitive processes postulated in experimental settings (Riefer & Batchelder, 1988). Recently, multinomial processing tree models were applied to analyze personality assessments (Batchelder, 2009), to model binary choice data (Batchelder et al., 2009), and to capture individual differences (Klauer, 2010).

In IRT, a tree structure has been exploited in rather implicit ways for modeling categorical item responses. For example, sequential models (also called continuation ratio models, Tutz, 1990) are formulated based on a sequential choice rule, which assumes all options are reviewed in a sequential manner from the first option to the last option. Culpepper (2014) presented new item response models based on a sequential decision rule for analyzing partially ordered item responses and for investigating repeatedly attempted item responses.

A divide-by-total scoring rule, which assumes all possible options are considered at once before the final answer is chosen, is also applied to formulate the partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1992), and the rating scale model (Andrich, 1978). Recently, Revuelta (2010) and Suh and Bolt (2010) employed a divide-by-total type of scoring rule in order to capture different types of decision making strategies for multiple-choice items.

The models discussed above assume a single source of individual differences in a response scale. There are other types of IRT models that concentrate on capturing multiple sources of individual differences in categorical item responses. For instance, Johnson (2003) and Johnson (2007) incorporated multiple latent traits to capture individual differences in response styles. Bolt et al. (2012) presented a multidimensional version of the nested logit model that allows for different latent traits to be involved in choosing distractors for multiple-choice items.

De Boeck and Partchev (2012) and Boeckenholt (2012) presented item response models that are based on a tree structure and allow for multiple sources of individual differences. These item response tree models can be specified as special cases of the proposed generalized item response tree models. Furthermore, the models discussed earlier in this section, such as sequential models, partial credit models, and their extensions with multiple latent traits can also be formulated within the proposed tree modeling framework. However, the proposed modeling framework does not allow for multiple paths to an observed outcome; therefore, the model developed by Boeckenholt (2013), which includes multiple paths, would not be constructed with the current framework.

Estimation, software, and model selection

Estimation and software

As discussed in the "General model formulation" section, generalized item response tree models can be formulated as a simple-structure K-dimensional item response model. It is important to recall that (1) we use the K-dimensional internal outcomes $\mathbf{Y}^*_p = (\mathbf{Y}^*_{p1}, ..., \mathbf{Y}^*_{pK})$ as the response variable for estimation and (2) the response matrix $\mathbf{Y}^*_p$ can contain missing values if the kth node does not contribute to generating a particular observed outcome (e.g., see the mapping matrix T in Eq. 6 includes 'NA' for Y^*_{pi2} when $Y_{pi} = 1$). Hence, generalized item response tree models can be estimated with standard IRT software packages that allow for multidimensionality as well as missing data. If one wants to estimate the modified versions that were discussed in the "A bifactor structure" to "Item and person covariates" section, then only software that allows all these options can be considered.

Here we use the R package *flirt* (Jeon et al., 2014b) for this general purpose. The package *flirt*, which is based on a convenient modular approach, can be a useful tool for fitting various generalized item response tree models. For instance, one can specify a different type of IRT model per node in terms of parametric forms, dimensionality, item and person covariates, and link functions. The package *flirt* also provides an efficient maximum likelihood (ML) estimation option, i.e., a modern expectation-maximization (EM)

algorithm, which implements an efficient E-step combined with graphical model theory (Lauritzen, 1995; Rijmen et al., 2008). Specifically, the E-step replaces the numerical integration over the joint latent space by a sequence of integrations over smaller subsets of (i.e., low dimensional) latent variables. The gain for the efficient E-step is considerable, especially when estimating multidimensional models for which high-dimensional numerical integration is required over the joint space of all latent variables. For example, for a three-dimensional one-parameter logistic (1PL) model with 108 items (36 items per dimension) and 1,069 subjects, the Laplace approximation with the R package lme4 (Bates and Maechler, 2009) (which was adopted by Partchev and De Boeck (2012) for estimation of item response tree models) took nearly four hours (14,264 seconds) whereas *flirt* took 809 seconds on a Intel Pentium Dual-Core 2.5-GHz processor computer with 3.2 GB of memory. For details on the modeling framework and the estimation framework of *flirt*, see Jeon et al. (2014a) and the website http://faculty.psy.ohio-state.edu/jeon/lab/flirt.php.

Model selection

ML estimation of the generalized item response tree models allows for model selection using likelihood-based fit statistics, such as the likelihood-ratio (LR) statistics, the Akaike's information criterion (AIC), and the Bayesian Information criterion (BIC).

In this study, we utilize the LR statistics to compare nested tree models. Suppose L_0 and L_1 are the maximum value of the likelihood of the data for Model 0 (with p_0 number of parameters) and for Model 1 (with p_1 number of parameters) and Model 0 is nested within Model 1. Then, $\chi^2 = -2 \times (\log L_0 - \log L_1)$ follows a Chi-squared distribution with $p_1 - p_0$ degrees of freedom. The LR test rejects that the null hypothesis ($H_0 : L_0 = L_1$) if χ^2 is larger than a Chi-square percentile with $p_1 - p_0$ degrees of freedom.

LR tests can be used with a specified item response tree model, for instance, to test the significance of node-main effects for item parameters (by comparing models with and without the node-main effects parameters). It should be noted that LR tests are suitable only for comparing tree models of the same size (i.e., same number of internal nodes). Depending on the specified tree structure, the data size used for estimation (the vector of internal outcomes) changes. The likelihood values are not comparable when two models are applied to different size data. In addition, LR tests are conservative for a comparison of models with a different dimensionality because one dimension less implies variance of zero (for the dimension), which is a boundary value in the parameter space (e.g., Baayen et al., 2008).

Applications

Here we provide two applications of the generalized item response model for psychological assessments: (1) modeling Likert scale item responses and (2) investigating missing item responses. Specifically, we test whether the Likert scale response categories can be assumed to be ordinal and examine the nature of omitted item responses in the tree modeling framework. All data analyses were carried out with the R package *flirt*. We conducted a simulation study to show parameter recovery of the estimation. The procedure and results are provided in the "Appendix".

Modeling Likert scale item responses

Psychological constructs such as personality and attitudes are frequently measured with Likert scale items. The response categories in Likert scales are typically coded numerically and in ascending order (e.g., 'strongly agree', 'agree', 'disagree', and 'strongly agree'). Applied researchers often treat them as continuous variables (e.g., 0,1,2,3) or collapse them into dichotomized categories (e.g., 0,1). These common conventions, however, can result in biased parameter estimates, incorrect standard errors, and inaccurate model fit information (Johnson, 2003; Rhemtulla et al., 2012).

Alternatively, ordinal factor analysis models (Christoffersson, 1975) and ordinal IRT models (e.g., partial credit models (Masters, 1982), rating scale models (Andrich, 1978), and graded response models (Samejima, 1969)) are used for analyzing Likert scale item response data. It is important to note that these ordinal models assume that the response categories are ordinal. This assumption may not be true, however if the response categories show qualitative differences rather than ranking or intensity differences. For instance, a middle response category that is commonly used, such as "Neutral" or "Undecided", may not be part of a true order with the other response categories (e.g., agree or disagree). A tree will allow for a partial ordering of the response categories.

Here we apply a generalized tree approach for investigating two types of Likert scale item responses. A generalized item tree model is constructed to test the ordinality assumption of the response categories. Several modifications of the model are estimated and compared to determine an optimal structure of the model.

Three-point Likert scale

The first example uses the verbal aggression dataset from De Boeck and Wilson (2004) and Vansteelandt (2000). The data were obtained from 316 first-year psychology students (243 females and 73 males), presented with a verbal aggression inventory with 24 items. The inventory concerns the

source of verbal aggression (type of situation) and its inhibition (discrepancy between wanting and doing). Specifically, each item consists of one of four frustrating situations (bus, train, store, and operator), one of two situation types (other-to-blame, self-to-blame), one of three verbally aggressive behaviors (cursing, scolding, and shouting), and one of two behavioral modes (wanting and doing). An example item is "A bus fails to stop for me. I would want to curse.": This corresponds to the 'cursing' aggressive behavior with the 'wanting' behavior mode in the 'other-to-blame' situation related to the 'bus'. "A bus fails to stop for me. I would actually curse" corresponds to the 'cursing' aggressive behavior but with the 'doing' behavior mode in the 'other-to-blame' situation related to the 'bus'. For each item, a three-point Likert scale ('No', 'Perhaps', and 'Yes') was used indicating whether respondents would agree to give an aggressive verbal response to the scenario.

Such three response categories are often treated as ordinal (e.g., 'No', 'Perhaps', 'Yes'). Johnson (2003) re-analyzed the data for investigating heterogeneity in response styles of the ordered response categories. Here we will model the data with a two-stage decision process that is postulated as follows: (1) in Stage 1, the person decides on the direction of the response (admission vs. negation), and (2) in Stage 2, the person decides on the certainty of the answer given admission, 'Yes' or 'Perhaps'). Figure 3 illustrates the item tree that represents this two-stage decision process.

This tree contains two internal nodes: The first node (Y_1^*) is related to the direction of the answer, i.e., the decision on negation of verbal aggression ('No') vs. admission of verbal aggression ('Perhaps', 'Yes'); its left branch represents negation and its right branch represents admission. The second node (Y_2^*) represents the certainty of the admission; its left branch represents that the person is not certain and therefore chooses 'Perhaps'; its right branch represents that the person is certain about his/her admission and chooses 'Yes'. The terminal nodes (i.e., observed response categories) represent the result of these internal choices. Recall that in Fig. 2 the 'Perhaps' category is differentiated in Stage 1 instead of in Stage 2. The tree in Fig. 2 is useful for finding out whether the 'Perhaps' response stems from a preliminary feeling of certainty. The tree in Fig. 3 is more useful for testing whether the 'Perhaps' response is a way to admit verbal aggression. Based on the model fit, the tree in Fig. 3 (log-likelihood $= -6087.22$) appears to be a better representation of the item response process for the data than the tree in Fig. 2 (log-likelihood $= -6188.76$).

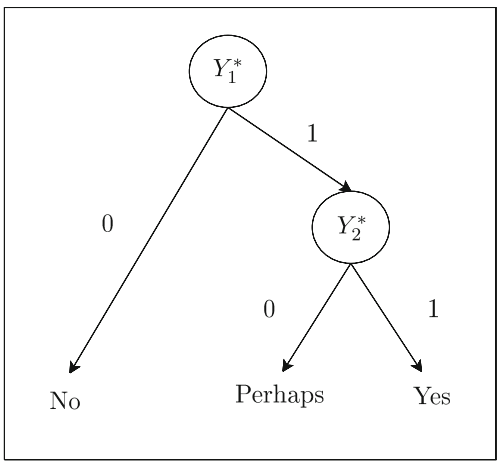


Fig. 3 Item response tree for the three-point Likert scale with the 'No', 'Perhaps', and 'Yes' categories

Based on the tree in Fig. 3, we fit a two-dimensional item response tree model. The correlation between the two dimensions is estimated as 0.38. This indicates that the two decision processes (direction and certainty) are only weakly correlated, implying that the resulting outcomes of the two internal processes are not closely related to each other. That is, the three response categories ('No', 'Perhaps','Yes') may not be really ordinal. The response scale seems to be multidimensional, meaning that a different latent variable is measured depending on the response option chosen by the person (also per item).

Additionally, the following three models are fit to test the dimensionality and internal measurement invariance of the tree (across nodes). All three are modifications of the initial model with node-specific latent trait and item parameters:

(1) Model 1: a reduced two-dimensional tree model that includes node-specific main effects parameters (but not node-specific item slopes and intercepts),
(2) Model 2: a unidimensional tree model that assumes a single latent variable (but with node-specific item slopes and intercepts), and
(3) Model 3: a bifactor tree model that assumes a general latent trait for all response categories in addition to the latent traits for the two internal nodes.

The results suggest the following: First, the full two-dimensional model fits better than the reduced, node-main effects model (Model 1) based on the LR test ($\chi^2 = 134.27$, $df = 46$, $p < 0.01$). The estimated node main effects are –2.69 for the slope parameter and –0.98 for the intercept parameter, which correspond to 0.068 and 0.37 in the odds, respectively. The two parameters are significant based on the LR test ($\chi^2 = 61.99$, $df = 2$, $p < 0.01$). Second, the full two-dimensional tree model fits better than the unidimensional model (Model 3) ($\chi^2 = 380.04$, $df = 1$, $p < 0.01$) but worse than the bifactor tree model ($\chi^2 = 247.05$, $df = 47$, $p < 0.01$). This analysis shows that (1) the measurement properties are not invariant across nodes in this response scale and (2) the scale is multidimensional but a general latent trait (that summarizes all dimensions) can be measured. In fact, the general latent trait would be the better

measurement than an ordinal model latent trait because it takes node specificity into account, whereas the ordinal model (which assumes a single latent trait) does not.

To illustrate this point, we fit the regular graded response model to the data and compare the latent trait estimates from the regular graded response model with the latent scores of the general factor from the bifactor item response tree model. We also compare them with the sum scores. The results show that (1) the latent trait estimates from the graded response model are nearly perfectly correlated with the sum scores (with the correlation of 0.97), and (2) the correlation of the sum scores with the latent trait estimates of the general factor from the bifactor tree model is 0.81. The former makes sense given that both the sum scores and the graded response model are based on the assumption that the response scale is ordinal. The latter implies that the response scale may not be perfectly ordinal. Furthermore, approximately 46 % of the people show more than 10 % difference in ranking when the bifactor item response tree model is applied than when the graded response model is applied. These results suggest that inference and decision based on subjects' latent trait scores can be biased if non-ordinality of a response scale is ignored in data analysis.

Four-point Likert scale

The second example uses the four-point Likert scale item responses from the British sample of the 1992 Eurobarometer Survey (Reif and Melich, 1992) provided by Bartholomew et al. (2008). The survey was administered to 392 people and seven questions in the survey measured people's perceptions of science and technology. An example question is 'Science and technology are making our lives healthier, easier, and more comfortable'. Each question provided the four response categories, 'strongly disagree', 'disagree to some extent', 'agree to some extent', and 'strongly agree'. The scale was comprised of two dimensions but for simplicity purposes we focus on one dimension containing four items (pertaining to comfort, work, future, and benefit).

For the four-point Likert scale, it is typically assumed that the response categories are ordinal from strongly disagree to strongly agree. To examine features of this response scale, we posit the following linear tree with three nodes as in Fig. 4: (1) Node 1: strongly disagree vs. the three higher categories, (2) Node 2: disagree vs. the two higher categories, and (3) Node 3: agree vs. strongly agree.

Here the first node (Y_1^*) represents choosing against strong negation ('strongly disagree'), the second node (Y_2^*) represents choosing against negation ('disagree'), and the third node (Y_3^*) represents choosing in favor of strong admission ('strongly agree'). The terminal nodes (four observed response categories) indicate the result of the internal decision processes.

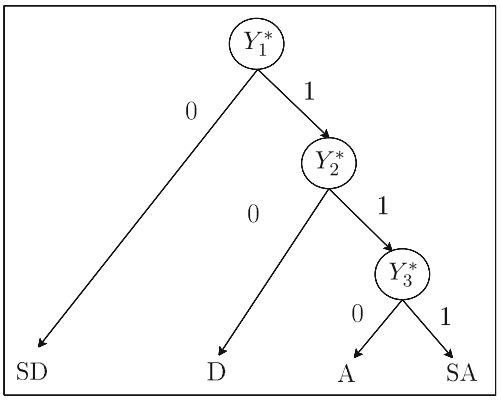


Fig. 4 Item response tree for four-point Likert item responses. SD is strongly disagree, D is disagree, A is agree, and SA is strongly agree

We first fit a three-dimensional tree model to the data. The estimated correlation is 0.54 between Dimensions 1 and 2, 0.13 between Dimensions 1 and 3, and 0.82 between Dimensions 2 and 3. This suggests that Dimension 1 is not correlated with Dimension 3 but somewhat correlated to Dimension 2, while Dimension 2 is highly correlated to Dimension 3. It shows that choosing against 'strongly disagree' (Dimension 1) is not really related to choosing in favor of 'strongly agree' (Dimension 3). It is important to note that if the scale were ordinal, one would expect high correlations between all decisions (choosing for the branch to the right). The low correlation between Dimension 1 and Dimension 3 indicates that this four-point Likert scale may not be perfectly ordinal.

Like for the previous example, we fit three additional models to test the dimensionality and internal measurement invariance of the tree:

(1) Model 4: a reduced three-dimensional model that includes the node-main effects parameters (but not node-specific item slopes and intercepts),
(2) Model 5: an unidimensional model that assumes a single latent variable across all nodes (with node-specific item slopes and intercepts), and
(3) Model 6: a bifactor tree model that assumes a general latent trait for all response categories in addition to the node-specific latent traits (with node-specific item slopes and intercepts).

The full three-dimensional model fits better than the reduced node-main effects model (Model 4) based on the LR test ($\chi^2 = 20.35$, $df = 12$, $p < 0.01$). Compared with Node 1 (as the reference node), the estimated node-main effects for the discrimination parameter are -0.91 and -0.92 for internal nodes 2 and 3, which correspond to about 0.4 (both) in the odds. The estimated node-main

effects for the intercept parameter are −2.79 and −6.34 for internal nodes 2 and 3, which correspond to about 0.006 and 0.002 in the odds, respectively. Jointly, the four parameters are significant based on the LR test ($\chi^2 = 440.03$, $df = 4$, $p < 0.01$). This suggests that the items are less discriminating for choosing against 'disagree' (Node 2) and for choosing in favor of 'strongly agree' (Node 3) than choosing against 'strongly disagree' (Node 1). The threshold is higher for choosing against 'strongly disagree' (Node 1) than for choosing against 'disagree' (Node 2). The model fit of the three-dimensional model is better than the unidimensional model (Model 5) ($\chi^2 = 23.42$, $df = 3$, $p < 0.01$) but it is similar to the bifactor tree model (Model 6) ($\chi^2 = 3.83$, $df = 8$, $p = 0.87$). Similar to the previous example, this analysis shows that (1) node-measurement invariance may not hold with the response scale and (2) the scale is multidimensional, while the bifactor structure can provide a summary measurement.

Similar to the previous example, we compare latent trait estimates from the graded response model and the general factor of the bifactor item response tree model with the sum scores. The sum scores are more highly correlated with the latent trait estimates from the graded response model (with the correlation of 0.96) than with the latent trait estimates from the bifactor tree model (with the correlation of 0.86). This result implies that this scale may not be perfectly ordinal. In addition, approximately 32 % of the people show more than 10 % difference in ranking when the bifactor item response tree model is applied than when the graded response model is applied. This result implies that inference based on subjects' latent trait scores can be biased when ordinality is forced to be assumed in data analysis.

Modeling missing responses

Missing data occur in most psychological and behavior assessments (Orme and Reis, 1991; Stevens, 1996; Allison, 2001; Pigott, 2001; Streiner, 2002; Acock, 2005). The APA Task Force on Statistical conference (Wilkinson & Statistical Inference, 1999) recommended that researchers report patterns of missing data and use statistical techniques to address the missing data problems. However, adequate reporting and handling of missing data is often ignored in practice (Peng et al., 2006; Saunders et al., 2006; Schlomer et al., 2010).

Common treatments for missing data are deletion (e.g., listwise, pairwise) and imputation methods (e.g., non-stochastic, stochastic). Deletion methods remove cases with missing values from data analysis whereas imputation methods substitute plausible values for missing values. Deletion of full cases can be used only when missing data are missing completely at random (MCAR); otherwise, the observed data are a biased subset of the complete data. Imputation methods are generally preferred to the deletion methods, but imputation methods still require at least missing at random (MAR). When missing is not at random (MNAR), any missing data technique can result in a biased inference. A better strategy is to incorporate the underlying missing data mechanism in modeling the data.

An item response tree approach can be a useful tool for modeling the missing data mechanism. For instance, the missing item response can be specified as another response category in addition to the existing response categories. Holman and Glas (2005) and Glas and Pimentel (2008) utilized tree-type IRT models to deal with omitted and non-reached item responses in ability testing, respectively. Recently, Debeer et al. (submitted) adopted an IRT tree modeling for analyzing both skipped and non-reached item responses in ability tests. Here we apply the proposed generalized item response tree approach to investigate omitted item responses in psychological assessments.

Modeling omitted item responses

The data come from the 38th round of the State Survey conducted by University's Institute for Public Policy and Social Research (2005). The survey was administered in 2005 to 949 Michigan citizens. Five questions measured the public's faith and trust in charity organizations, such as 'Charitable organizations are more effective now in providing services than they were five years ago' and 'I place a low degree of trust in charitable organizations'. A four-point Likert scale was used with four response categories 'strongly agree', 'somewhat agree', 'somewhat disagree', and 'strongly disagree'. In this dataset, the responses were coded from 0 to 3, with larger scores indicating less favorable views of charities. There were also missing values for each item.

Linear tree with ordinal responses Let us assume that we are only interested in separating out omitted responses from regular ordinal item responses. The following two stages can then be posited: (1) in Stage 1, the person decides on responding (i.e., omit or not omit) and (2) in Stage 2, the person chooses one category among the four ordinal response categories (i.e., strongly disagree to strongly agree). See Fig. 5 for the tree representation of this process.

The tree in Fig. 5 includes two internal nodes: Y_1^* represents response omission, and Y_2^* represents a relative degree of faith/trust in charity. The four regular response categories are commonly assumed to be ordinal. Although ordinality of the scale will be tested as in the previous applications, we focus first on modeling the missing responses.

We fit a two-dimensional tree model to the data where a regular two-parameter IRT model is used in Dimension 1 but a graded response model is applied in Dimension 2. The two-dimensional model fits better than the unidimensional

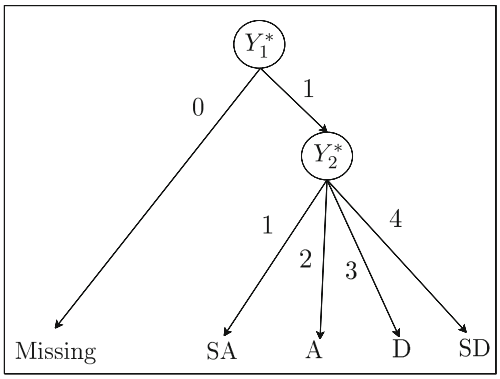


Fig. 5 Linear tree with two nodes for omitted (missing) responses as well as four response categories. SA is strongly agree, A is agree, D is disagree, and SD is strongly disagree

model based on the LR test ($\chi^2 = 131.53, df = 1, p < 0.01$). In the two-dimensional model, the estimated correlation is very low (-0.10) between Dimension 1 (involvement) and Dimension 2 (charity). This makes clear that skipping items differs from the construct that is intended to be measured with the scale (with the regular item responses). The full two-dimensional model (log-likelihood $= -5796.03$) does not fit better than the reduced version with a zero correlation between the two dimensions (log-likelihood $= -5796.80$). This implies that for this data the missing data mechanism may be considered ignorable (i.e., missing at random; MAR), but it will be shown that this conclusion needs to be modified based on further analysis.

Elaborated nested tree We expand the linear tree specified above by elaborating the regular response process. The following four-stage process can then be postulated: (1) in Stage 1, the person decides on responding (i.e., omit or not omit) (2) in Stage 2, the person decides on the direction (i.e., agree or disagree), (3) in Stage 3, the person decides on the intensity of admission (i.e., strongly agree or agree), and (4) in Stage 4, the person decides on the intensity of negation (i.e., strongly disagree or disagree). See Fig. 6 for the tree representation of this process.

Note that Stages 3 and 4 are not sequential because either Stage 3 or Stage 4 can follow after Stage 2 but not both. It means that the specified tree is nested instead of linear.

This tree includes four internal nodes: Y_1^* represents involvement, Y_2^* represents direction (i.e., admission vs. negation), Y_3^* represents strong admission, and Y_4^* represents strong negation. The terminal nodes (i.e., omitted response and four observed response categories) represent the result of this nested four-stage decision process.

We fit a four-dimensional tree model to the data. The estimated correlation is -0.04 between Dimension 1 (responding) and Dimension 2 (direction), 0.20 between Dimension 1 (responding) and Dimension 3 (strong admission), and -0.35 between Dimension 1 (responding) and Dimension 4 (strong negation). This confirms again that the decision for skipping items is distinct from the regular response processes. Omission (Dimension 1), the opposite of response, is positively related to strong negation (Dimension 4), which suggests that omission is related to negative attitudes, and the MAR condition is in fact not fulfilled. In addition, the estimated correlation is 0.34 between Dimension 2 (direction) and Dimension 3 (strong admission), and 0.49 between Dimension 2 (direction) and Dimension 4 (strong negation). The estimated correlation between Dimension 3 (strong admission) and Dimension 4 (strong negation) is 0.63. This latter conclusion suggests an underlying extreme response style.

We also fit the following three models to evaluate the dimensionality and node-measurement-invariance of the tree:

(1) Model 7: an unidimensional model that assumes a single latent variable across all nodes (with node-specific item slopes and intercepts),
(2) Model 8: a three-dimensional model with a latent trait for Node 1 (responding), a latent trait for Node 2 (direction), and a latent trait for Nodes 3 and 4 (extremity) (with node-specific item slopes and intercepts), and
(3) Model 9: a reduced three-dimensional model that includes the main-node effects slope and intercept parameters for Node 4 (compared to Node 3) (with node-specific item slopes and intercepts for Nodes 1 and 2).

We did not fit the bifactor version of the model here because a general trait would not be helpful in explaining both the omission and regular item processes.

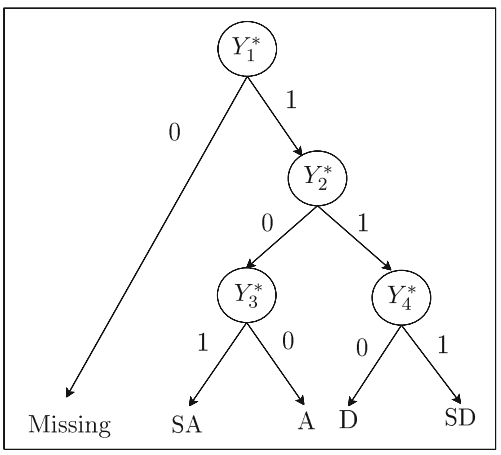


Fig. 6 Nested tree with four nodes for omitted (missing) responses as well as four response categories. SA is strongly agree, A is agree, D is disagree, and SD is strongly disagree

The four-dimensional model fits better than the unidimensional model (Model 7) ($\chi^2 = 255.13, df = 6, p < 0.01$) or the three-dimensional model (Model 8) ($\chi^2 = 85.86$, $df = 3$, $p < 0.01$). The three-dimensional model fits better than its reduced, node-main effects model (Model 9) based on the LR test ($\chi^2 = 19.32$, $df = 8$, $p = 0.01$). Compared with Node 3, the estimated main effects for the discrimination parameter is 0.29 (1.34 in the odds) and the estimated main effects for the intercept parameter is −0.33 (0.72 in the odds) for Node 4. The two parameters are jointly significant based on the LR test ($\chi^2 = 20.35$, $df = 2$, $p < 0.01$). This suggests that overall the items are more discriminating and the threshold is higher for choosing in favor of strongly disagree than choosing in favor of strongly agree. Because the four-dimensional model seems to have a better fit, it can be concluded that although being extreme on one end of the response scale is highly positively related to being extreme on the other hand (correlation of 0.63), being extreme is not just one latent variable.

Consequences of ignoring missing not at random (MNAR) We conducted an experiment to illustrate the consequence of ignoring missing not at random (MNAR). We utilized the verbal aggression data used in the "Three-point Likert scale" section. The data consist of complete responses to 24 items for 316 students. We generated MNAR in the data by applying the following procedure: First, we computed the probability (p_{ij}) of a positive reaction (Yes and Perhaps) for all item responses. Second, if $p_{ij} < 0.3$ (low probability), we replaced at random 50 % of such cases with missing (*NA*). For one item (item 21), which produced no missing with $p_{ij} < 0.3$, a higher threshold $p_{ij} < 0.5$ was applied. As a result, the simulated data contain about 8 % to 49 % (with average 36 %) of MNAR cases across items. We then specified the item response tree model by treating missing values as an additional response category. The tree structure used for the model formulation is shown in Fig. 7.

We then fit the following three models to the simulated data:

(1) Model 10: a two-parameter graded responses model (ignoring missing data),
(2) Model 11: a two-parameter graded responses model (treating missing data as incorrect ('No') responses), and
(3) Model 12: a two-parameter item response tree model with the tree structure in Fig. 7.

Note that the regular graded response model (Model 10) treats missing data as ignorable (missing at random), whereas the item response tree model (Model 12) takes into account the possibility that missing data follow a different response mechanism than non-missing responses. Treating missing data as incorrect responses is a common practice in applied research. In a formally equivalent way, we treated missing responses as 'No' (Model 11).

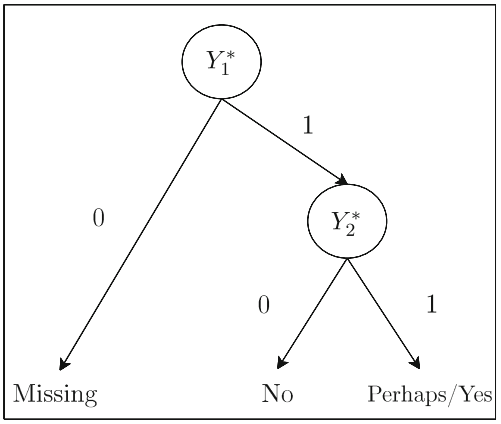


Fig. 7 Tree with two nodes for missing responses as well as two response categories ('No' vs. 'Perhaps'/'Yes')

Finally, we obtained and compared the latent trait estimates from the three models. For the item response tree model (Model 12), the latent trait estimates for Node 2 were considered for comparison because Node 2 refers to the psychological mechanism of main interest, while Node 1 refers to a response omission mechanism. Figure 8 displays the estimated latent traits from the three models.

The figure shows that the latent trait estimates from Model 12 are different (tend to be smaller) than those from Models 10 and 11, while the latent trait estimates from Models 10 and 11 are similar to each other. Specifically, the correlation between the estimated latent trait scores from the graded response model (Model 10) and from the model treating missing as 'No' responses (Model 11) is 0.97. However, the correlations between the Node 2 latent score from the tree model (Model 12) and the latent score from the graded response model (Model 10) and the model treating missing as 'No' (Model 11) are much lower: 0.50, 0.38, respectively. In terms of subject ranking, approximately 91 % of the subjects show more than 10 % difference in ranking when Model 12 is applied than when Models 10 and 11 are applied. These results suggest that when MNAR is treated as ignorable (Model 10) and treated as 'No' responses (Model 11), the inference on the persons' latent traits can be biased to some degree.

Discussion

Traditional IRT models focus on studying observed categorical item responses. The "black-box" decision process that leads to the observed item responses is, however, typically not of much interest in common IRT analysis. In this article, we presented generalized item response tree models that can be useful for investigating underlying decision processes. The item response tree models can also be useful for describing distinctive features of item response categories and for investigating multiple sources of heterogeneity in a response scale. These purposes are often neglected in traditional IRT analysis.

The proposed generalized item response tree modeling framework is flexible in its parametric form, dimensionality, and choice of explanatory variables. In particular, one can test whether a general latent trait can be used to summarize potentially multidimensional response categories with a bifactor structure. It is noteworthy that the bifactor version of the tree model has a computational advantage for maximum likelihood estimation, because the total dimensionality involved in calculating the likelihood can be remarkably reduced based on the conditional independence assumption between latent variables when a bifactor structure is exploited (Gibbons and Hedeker, 1992; Gibbons et al., 2007; Rijmen, 2009; Jeon et al., 2013).

We provided two kinds of applications to demonstrate the utilities of the proposed approach for investigating Likert-scale item responses and missing responses for psychological assessments. We also showed that exploiting a bifactor structure in a item response tree model can be a measurement solution when the response categories are multidimensional and may not be perfectly ordinal. In addition, we showed that a proposed tree modeling approach can be useful for testing the MAR assumption on missing item responses.

We introduced the new IRT package, *flirt* for maximum likelihood estimation of the proposed framework. The package *flirt* can be a useful tool for efficiently estimating high-dimensional tree models with explanatory variables. With *flirt*, different types of IRT models can be specified per node (dimension) in terms of parametric forms, dimensionality, link functions, covariates, and so on.

One limitation of the proposed model is that the tree allows only one unique path to an observed response category. This restriction may be relaxed by allowing for multiple paths to an observation as in Boeckenholt (2013). However, further research is needed to embrace multiple-path tree models in the current framework because including multiple paths may not only produce a model identification problem but also make it impossible to convert the tree

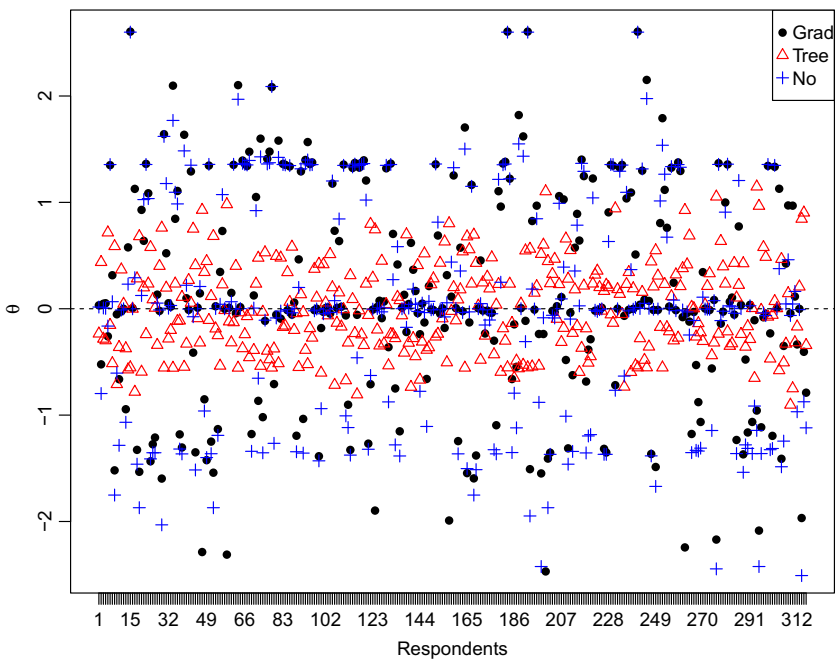


Fig. 8 Latent trait estimates from the regular graded response model (Graded), the item response tree model (Tree), and the graded response model with missing treated as incorrect responses (No)

model to a standard IRT model that can be estimated with a standard IRT software package.

Acknowledgments The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305D110027 to Educational Testing Service. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Appendix

We carried out a simulation study to show parameter recovery of the generalized item response tree models. We utilized the verbal aggression data used in the "Three-point Likert scale" section. We specified a two-dimensional tree model with two nodes (based on the tree structure shown in Fig. 4). The parameter estimates from the model were used as the data generating values. The two-dimensional latent scores were generated from a multivariate normal distribution with the estimated means and covariance matrix. Based on a sample size of $N = 316$ (same as the empirical data), 100 datasets with a total of 97 parameters (48 item discrimination parameters + 48 intercept parameters + 1 covariance parameter) were generated and evaluated using the package `flirt`.

Figure 9 displays boxplots of the estimated error ($\hat{\psi}_m - \psi_m$, where ψ_m and $\hat{\psi}_m$ are the true and estimated values for the mth parameter, $m = 1, ..., 97$).

In Fig. 9, each boxplot shows the distribution of the estimated error for the corresponding parameter across 100 simulated datasets. The bold horizontal line (inside the inner box) denotes the median of the data, the lower and upper ends of the box indicate the first and third quartiles, respectively, and the lower and upper ends of whiskers denote the minimum and maximum values, respectively. For most parameters, the estimated errors ranged from -1.0 to 1.0 (only a few parameters showed a maximum/minimum error greater than ± 1.0 but smaller than ± 2). The bias estimates (the means of the estimated errors across datasets) were small and ranged from -0.16 to 0.035 with a mean of -0.03 (only seven cases were greater than ± 0.10). The mean squared error (MSE) estimates were also quite small and ranged from 0.007 to 0.40 with the mean of 0.04 (only two cases were greater than 0.20).

This result suggests that true parameter values of the specified item response tree model were recovered quite well using the package `flirt`.

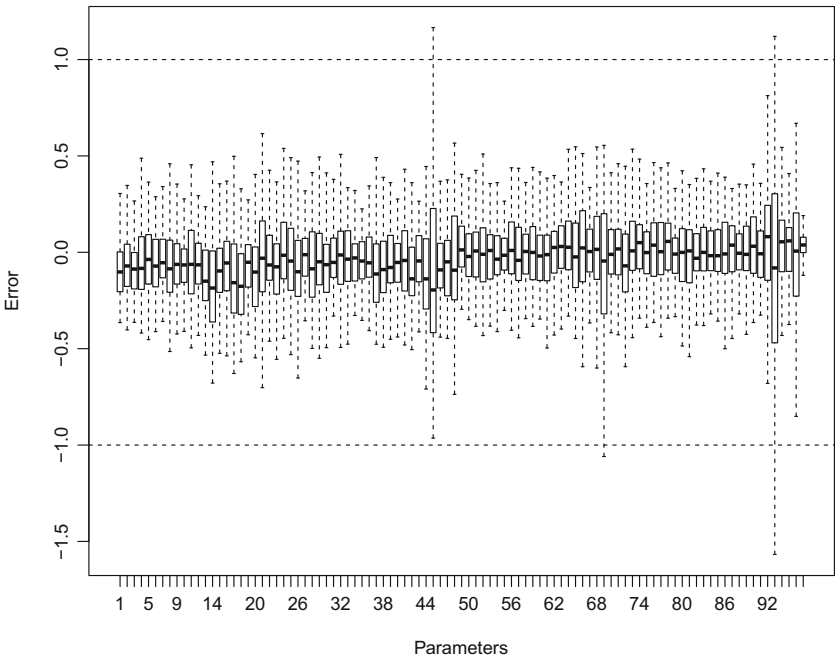


Fig. 9 Estimated error ($\hat{\psi}_m - \psi_m$) for 97 parameters ($m = 1, ..., 97$). The parameters are aligned in the order of 48 item discrimination parameters (item 1 to item 24 for Node 1 and item 25 to item 48 for Node 2), 48 intercept parameters (item 1 to item 24 for Node 1 and item 25 to item 28 for Node 2), and 1 covariance parameter

References

Acock, A.C. (2005). Working with missing values. *Journal of Marriage and Family*, *67*, 1012–1028.

Allison, P.D. (2001). *Missing data*. CA: Sage, Thousand Oaks.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modelling with crossed random effects for subjects and items. *Journal of Language and Memory*, *59*, 390–412.

Bartholomew, D.J., Steele, F., Moustaki, I., & Galbraith, J.I. (2008). *Analysis of Multivariate Social Science Data*, Second Edition. Boca Raton, FL: Chapman & Hall/CRC.

Batchelder, W.H. (2009). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In Embretson, S. (Ed.) *Measuring psychological constructs: Advances in model-based measurement* (pp. 71–94). Washington, DC: American Psychological Association.

Batchelder, W.H., Hu, X., & Smith, J.B. (2009). Multinomial processing tree models for discrete choice. *Journal of Journal of Psychology*, *217*, 149–158.

Bates, D., & Maechler, M. (2009). *lme4: Linear mixed-effects models using S4 classes R package version 0.999375-31*. Downloadable from http://CRAN.Rproject.org/package=lme4

Boeckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*, 665–678.

Boeckenholt, U. (2013). Modeling motivated misreports to sensitive survey questions. *Psychometrika*, *78*, 188–201.

Bolt, D., Wollack, J., & Suh, Y. (2012). Application of a multidimensional nested logit model to multiple-choice test items. *Psychometrika*, *77*, 339–357.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5–32.

Culpepper, S. (2014). If at first you don't succeed, try, try again: Applications of sequential IRT models to cognitive assessments. *Applied Psychological Measurement*, *38*, 632–644.

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*, 1–28.

De Boeck, P., & Wilson, M. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer.

Debeer, D., Janssen, R., & De Boeck, P. (submitted). Modeling responses and omissions within a tree-based IRT framework.

Gibbons, R.D., Bock, D., Hedeker, D., Weiss, D.J., Segawa, E., Bhaumik, D.K., Kupfer, D.J., Frank, E., Grochocinski, V.J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*, 4–19.

Gibbons, R.D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Glas, C.A., & Pimentel, J.L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*, 907–922.

Holman, R., & Glas, C.A. (2005). Modelling non-ignorable missing-data mechanism with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*, 1–17.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*, *38*, 32–60.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2014a). Flexible item response theory modeling with the R package flirt. *Applied Psychological Measurement*, *38*, 404.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2014b). `flirt`: Flexible item response theory modeling. R package version 1.15. downloadable from http://faculty.psy.ohio-state.edu/jeon/lab/flirt.php

Johnson, T.R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*, *68*, 563–583.

Johnson, T.R. (2007). Discrete choice models for ordinal response variables: A generalization of the stereotype model. *Psychometrika*, *72*, 489–504.

Klauer, K. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*, 70–98.

Lauritzen, S.L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, *19*, 191–201.

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

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

Orme, J.G., & Reis, J. (1991). Multiple regression with missing data. *Journal of Social Science Research*, *15*, 61–91.

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*, 23–32.

Peng, C.-Y.J., Harwell, M., Liou, S.-M., & Ehman, L.H. (2006). Advances in missing data methods and implications for educational research. In Sawilowsky, S. (Ed.) *Real data analysis* (pp. 31–78). Greenwich, CT: Information Age.

Pigott, T.D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, *7*, 353–383.

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*.

Reif, K., & Melich, A. (1992). Euro-barometer 38.1: Consumer protection and perceptions of science and technology. Technical report, INRA (Europe), Brussels.

Revuelta, J. (2010). Estimating difficulty from polytomous categorical data. *Psychometrika*, *331-350*, 331–350.

Rhemtulla, M., Brosseau-Liard, P.E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Psychological Methods*, *17*, 354–373.

Riefer, D.M., & Batchelder, W.H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318–339.

Rijmen, F. (2009). An efficient EM algorithm for multidimensional IRT models: Full information maximum likelihood estimation in limited time. ETS Research Report (RR0903).

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika*, *73*, 167–182.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *34*, 100–114.

Saunders, J.A., Morrow-Howell, N., Spitznagel, E., Dore, P., Proctor, E.K., & Pescarino, R. (2006). Imputing missing data: A comparison of methods for social work research. *Social Work Research*, *30*, 19–30.

Schlomer, G.L., Bauman, S., & Card, N.A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*, *57*, 1–10.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences*, 3rd ed. NJ: Erlbaum, Mahwah.

Streiner, D.L. (2002). The case of the missing data: Methods of dealing with dropouts and other research vagaries. *Canadian Journal of Psychiatry*, *47*, 68–74.

Suh, Y., & Bolt, D.M. (2010). Nested logit models for multiple-choice item response data. *Psychometrika*, *75*, 454–473.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*, 39–55.

University's Institute for Public Policy and Social Research (2005). Institute for public policy and Michigan State University social research. State of the state survey-38. Spring 2005. http://www.ippsr.msu.edu/SOSS

Vansteelandt, K. (2000). Formal models for contextualized personality psychology. Unpublished doctoral dissertation, K.U. Leuven, Belgium.

Wilkinson, L., & Statistical Inference, T.F. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.