JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
Journal of Educational Measurement
Fall 2017, Vol. 54, No. 3, pp. 333–363

Modeling Skipped and Not-Reached Items Using IRTrees

Dries Debeer
KU Leuven and University of Zurich
Rianne Janssen
KU Leuven
Paul De Boeck
KU Leuven and Ohio State University

When dealing with missing responses, two types of omissions can be discerned: items can be skipped or not reached by the test taker. When the occurrence of these omissions is related to the proficiency process the missingness is nonignorable. The purpose of this article is to present a tree-based IRT framework for modeling responses and omissions jointly, taking into account that test takers as well as items can contribute to the two types of omissions. The proposed framework covers several existing models for missing responses, and many IRTree models can be estimated using standard statistical software. Further, simulated data is used to show that ignoring missing responses is less robust than often considered. Finally, as an illustration of its applicability, the IRTree approach is applied to data from the 2009 PISA reading assessment.

In large-scale educational assessments, such as the Program for International Student Assessment (PISA), test takers do not always respond to all the items in their test booklet. In fact, two types of omitted responses can be discerned. First, there can be omissions near the end of the test so that the last attempted item is not the final item in the test booklet. In this case, the test taker *drops out* at a certain point and the items beyond that point can be considered as *not reached*. Second, there can also be omissions well before the last attempted item. In that case, the test taker *skipped* certain items. In PISA 2006, for instance, across all countries 10% of the items were skipped and 4% were not reached (OECD, 2009). In PISA 2009, the percentages are comparable: 9% skipped items and 4% not reached (OECD, 2012). In the remainder of this article, not-reached items and skipped items will be coded as r and d, respectively.

Following Rubin's (1976) categorization, three missing data mechanisms can be discerned: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Both MCAR and MAR do not pose a problem for likelihood-based inferences (Little & Rubin, 1987; Mislevy & Wu, 1996; Rubin, 1976). In these situations, the missingness is called ignorable. However, whenever the process underlying the missing responses is related to the measured attribute, the missingness is MNAR, and therefore not ignorable. Moreover, ignoring nonresponses when the missingness mechanism is MNAR can result in biased inferences.

It has been suggested that different processes underlie skipped and not-reached items. First, skipping is associated with the proficiency of the test taker and features

Copyright © 2017 by the National Council on Measurement in Education

of the item such as the perceived difficulty (e.g., Stocking, Eignor, & Cook, 1988), or the item format (i.e., open-ended vs. multiple-choice; Okumura, 2014). Second, not-reached items are associated with lack of motivation, test-taking strategy, and so on (e.g., Mislevy & Wu, 1988). Therefore, the two types of omissions should be treated differently. In addition, because both types may be related to the measured proficiency, the nonresponses can be informative and should not just be ignored.

The purpose of this article is to present an IRTree-based approach for modeling responses and omissions jointly, taking into account that test takers as well as items can contribute to the two types of omissions. In the following, we first give a brief overview of the current treatment of missing responses in educational assessments. Then, we introduce the IRTree framework of De Boeck and Partchev (2012) and apply it to missing responses. We will present two distinct IRTrees to simultaneously model skipped and not-reached items, and we will show that several existing models for missing responses can be formulated as IRTrees. We will argue that the IRTree framework is flexible, and that the models can be estimated using standard statistical software. Subsequently, using simulated data, we demonstrate the feasibility of the IRTree approach and compare it with ignoring missing responses. Finally, we apply the IRTree approach to empirical data form the PISA 2009 reading assessment in Argentina.

Treating Missing Responses

Although Huisman (2000) argued that the treatment of missing responses should depend on the mechanism that is causing the omissions, often the operational strategy to treat nonresponses is chosen in a somewhat *ad hoc* manner, ignoring possible bias. Common strategies to deal with missing responses can be divided in roughly three categories: response imputation, ignoring nonresponses, and modeling the missingness. The three strategies will be briefly discussed here; a more extensive overview can be found in Rose (2013).

Imputation Strategies

Although it has been criticized for more than 30 years (e.g., Lord, 1974), incorrect answer substitution (IAS) is still the most used form of response imputation in large-scale international assessments. Because IAS replaces a missing response by an incorrect response, it assumes that, regardless of the response pattern on the other items, the test taker could not have been able to solve the omitted item. In other words, IAS assumes that the missingness is MNAR in a very specific way.

Another commonly used imputation strategy is person mean imputation (PMI). PMI imputes the proportion of correct responses (only with respect to the items with a response). PMI implicitly assumes a similar proportion of correct responses (or a similar person mean) for the omitted items. Apart from IAS and PMI, there are other ways to impute item responses (e.g., random score imputation), some of which are very ingenious (Huisman, 2000).

Ignoring Missing Responses

Ignoring an item nonresponse is traditionally used in likelihood-based item response theory (IRT) approaches, where it is rather straightforward. One simply

treats the item with the missing response as if it was not administered (i.e., as if it was missing by design). Hence, ignoring strategies assume that the missingness is MAR, and therefore ignorable. When the skipped items and/or not-reached items are MNAR, however, the missing-data pattern is nonignorable, and ignoring the missing responses can cause bias on likelihood-based estimates (Finch, 2008; Mislevy & Wu, 1996).

When the two types of missing responses (i.e., skipped items and not-reached items) are both present, often a pragmatic approach is taken by a combination of an imputation strategy and ignoring nonresponses (e.g., treating skipped items as wrong and not-reached items as ignorable).

Modeling Strategies

In contrast with IAS and ignoring nonresponses, the treatment of missing responses in modeling strategies is not based on an *ad hoc* decision. Moreover, by including the missingness pattern in the measurement model, the modeling strategies are designed to handle cases of MNAR. Two modeling strategies have been repeatedly proposed, namely, a latent regression IRT approach, and a multidimensional IRT (MIRT) approach.

Latent regression modeling. A latent regression IRT model includes the missingness pattern in the measurement model by making the model for the observed data conditional upon an aggregate of the missing responses, such as the number of missing responses. Using this strategy, Rose, von Davier, and Xu (2010) proposed a multiple-group IRT approach, where the grouping variable represented the stratification of the test takers by the amount of omitted responses. A disadvantage of the latent regression approach is that the uncertainty of the estimates in the first step is not taken into account in the second step. Furthermore, the model does not supply a framework for interpreting the processes underlying the missing responses.

MIRT models. Another way to deal with missing responses is to identify and model the missing-data mechanism jointly with the relevant model for the observed data. In this way, Holman and Glas (2005) modeled skipped items, and Glas and Pimentel (2008) modeled not-reached items. Both these models combine an IRT model for the observed data (i.e., the proficiency process) with an IRT model for the missing data. By assuming a joint distribution for the two latent traits (proficiency and tendency to omit items) the models are two-dimensional IRT models, and the correlation between the latent traits can be estimated. Because both nonresponses can occur in the same test, combinations of both models have been proposed (Glas & Pimentel, 2008; Rose, 2013) and used (Pohl, Gräfe, & Rose, 2014). Recently, the MIRT approach has been extended to polytomous items (Glas, Pimentel, & Lamers, 2015; Jeon & De Boeck, 2016), and covariates were introduced in the model for the missing data (Glas et al., 2015; Okumura, 2014).

The benefit of these MIRT approaches is that they model the missing data and the responses with separate latent variables. If the model is correctly specified, it can deal with cases of MNAR. Moreover, the estimates of the correlation between the proficiency process and the missingness process are indicative of which missingness

mechanism is present in the data; that is, when the missing data pattern is MAR, the propensity to omit and the proficiency will be uncorrelated.

IRTrees for Missing Responses

In this section, we present the IRTree framework (De Boeck & Partchev, 2012) as a new approach to model missing responses within the MIRT tradition. In contrast with current MIRT approaches, and in line with the ideas of cognitive psychometrics (Batchelder, 2010), IRTrees are process models, with a behavioral process interpretation. We will propose four IRTrees for modeling missing responses. In addition, we will show that the framework is flexible, that it includes several previously proposed MIRT models, and that it can be used both descriptively and explanatorily. Finally, many IRTrees can be estimated using standard statistical packages.

The IRTree Approach

The IRTree framework (De Boeck & Partchev, 2012) applies the logic of a tree-based model to the process of responding to a test item. The categorical outcomes are modeled as a tree structure of sequentially interconnected subprocesses. Each subprocess is represented as an internal node that branches out in other nodes with a certain probability. The observed response categories are represented as the endnodes of the IRtree. Every endnode is considered to be the result of following a specific path through the tree. Hence, the probability of an observed response corresponds to the probability of passing through the branches that lead to the endnode representing that response. Within an IRTree, the internal nodes are modeled as IRT models. Generally, and using a common notation, the probability that test taker p passes through a specific branch leaving internal node n of item i is modeled as a combination of both the test taker's location $\theta_p^{(n)}$ and the item's location (or threshold) $\beta_i^{(n)}$ on the latent propensity scale that is relevant for node n. Interpretative examples are given below, and an overview of the IRTree framework can be found in De Boeck and Partchev (2012).

Four IRTrees for Missing Responses

Figure 1 shows graphical representations of four IRTrees that we present to model nonresponses in binary test items. Yet, applying the IRTree framework to missing responses is not limited to these four models. In Figure 1, the internal nodes or subprocesses are represented by circles. The endnodes, which refer to the observed response categories, are represented by squares. Because the number of branches leaving an internal node is always two, the presented IRTrees are binary. For simplicity, only the binary case will be discussed; an extension to polytomous items (or polytomous subnodes) is straightforward (Jeon & De Boeck, 2016).

An IRTree for skipped items. When there are only skipped items in an administration, the IRTree in Figure 1(a) can be used. In this IRTree, there are three possible endnodes: a correct response ($x = 1$), an incorrect response ($x = 0$), and a skipped item ($x = d$); and there are two internal nodes $y^{(1)}$ and $y^{(2)}$. Internal node $y^{(1)}$ represents the skipping process, with two possible outcomes (i.e., two branches leaving

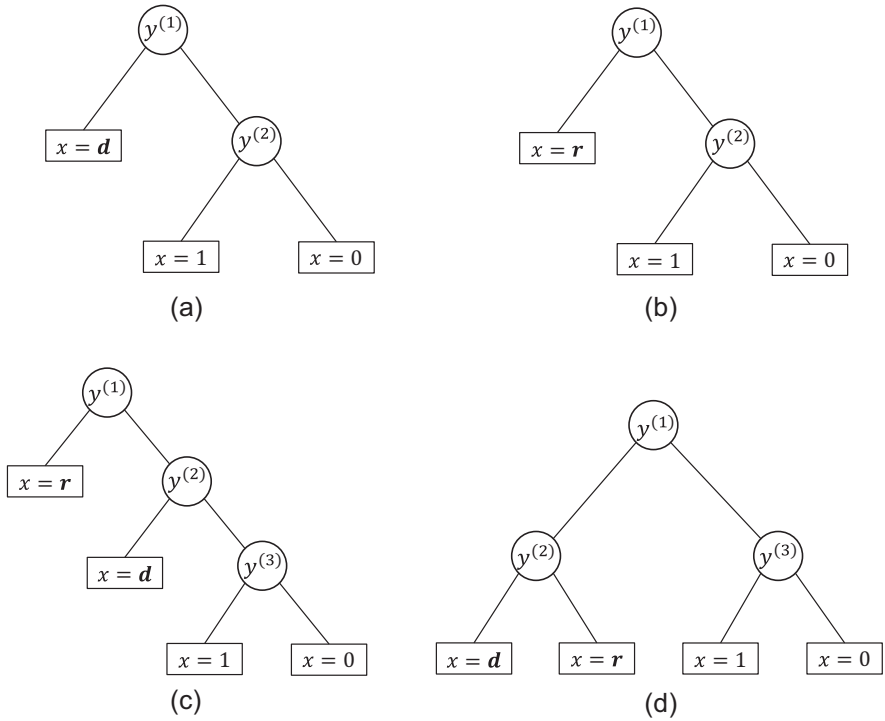


Figure 1. Graphical representations of four IRTrees to model nonresponses in binary test items. The internal nodes, or latent subprocesses, are represented as circles. The endnodes, or observed responses categories, are represented as squares.

the internal node). The right branch leads to internal node $y^{(2)}$, which represents the proficiency process.

Within our notation, $\theta_p^{(1)}$ and $\theta_p^{(2)}$ refer to the test taker's propensity to skip and the proficiency, respectively. Likewise, $\beta_i^{(1)}$ and $\beta_i^{(2)}$ are the item's skipping threshold and the item's difficulty. The joint model produces the probabilities for the observed response categories (i.e., $x_{pi} = 0$, 1 or d) as combinations of the probabilities of the subprocesses (see below).

An IRTree for not-reached items. When there are only not-reached items in an administration, the IRTree in Figure 1(b) can be used. The left endnode is now ($x = r$), $\theta_p^{(1)}$ is interpreted as the test taker's propensity to drop out, and $\beta_i^{(1)}$ is interpreted as the item's threshold for dropping out.

Because the items in the first position(s) are usually attempted by all test takers, it is impossible to estimate the dropping-out thresholds separately for these items. Instead, a restriction can be imposed on the dropping-out thresholds. For instance, Glas and Pimentel (2008) proposed a linear function of item position,

$$\beta_{ik}^{(1)} = \delta_a + (k - K)\delta_b, \tag{1}$$

where $\beta_{ik}^{(1)}$ is the threshold for dropping out of item i in position $k = 1, 2, ..., K$, and K is the number of items in the test booklet. This restriction supports a monotonously increasing probability of dropping out, and makes it possible to estimate a dropping-out threshold for every item.

Two IRTrees for skipped and not-reached items. To deal with both skipped and not-reached items simultaneously, a third internal node can be added to the IRTrees in Figures 1(a) and 1(b) in different ways. Two possible IRTrees are represented in Figures 1(c) and 1(d), which we will refer to as the *item selection model* (ISM) and the *continuing effort model* (CEM), respectively.

The ISM. In the ISM (cf. Figure 1(c)), the three sequential internal nodes refer to the process of dropping out $y^{(1)}$, the process of skipping an item $y^{(2)}$, and the proficiency process $y^{(3)}$. Hence, $\theta_p^{(1)}$, $\theta_p^{(2)}$, and $\theta_p^{(3)}$ are, respectively, the propensity to drop out, the propensity to skip, and the proficiency. $\beta_i^{(1)}$, $\beta_i^{(2)}$ and $\beta_i^{(3)}$ are the corresponding item thresholds.

The CEM. The CEM (cf. Figure 1(d)) has the same endnodes as the ISM, but its internal nodes are sequenced differently. The subnodes in the CEM refer to deciding whether or not to give a response $y^{(1)}$, deciding whether to continue or drop out $y^{(2)}$, and the proficiency process $y^{(3)}$. Correspondingly, $\theta_p^{(1)}$, $\theta_p^{(2)}$, and $\theta_p^{(3)}$ are the propensity to omit, the propensity to continue (after a nonresponse) and the proficiency. $\beta_i^{(1)}$, $\beta_i^{(2)}$, and $\beta_i^{(3)}$ are the corresponding item thresholds. Note that the third subprocess is the same in the CEM as in the ISM.

Modeling the Response Probabilities Using Subitems

Passing through an IRTree can be seen as a sequence of intermediate outcomes on (unobserved) subitems. Each subitem corresponds with an internal node, and its possible outcomes correspond with the branches leaving the node. For instance, in the binary IRTrees in Figure 1, left and right branches can be coded 1 and 0, respectively. As a result, each observed response category, can be mapped onto a specific response pattern on the subitems (see De Boeck & Partchev, 2012 for an elaborate discussion). Each IRTree corresponds to a specific mapping matrix. Appendix A gives the mapping matrices of the four IRTrees presented above, as well as sample data on which the mapping is applied.

Using the mapping matrix, the probabilities of the response categories can be formulated as

$$\pi(x_{pi} = m|\boldsymbol{\theta}_p, \boldsymbol{\beta}_i) = \prod_{n=1}^{N} \left[\frac{exp(\theta_p^{(n)} + \beta_i^{(n)})^{t_{mn}}}{1 + exp(\theta_p^{(n)} + \beta_i^{(n)})} \right]^{q_{mn}}, \tag{2}$$

with x_{pi} the response of person p on item i, $m = 0, 1, d$, or r, as the response categories, and $n = 1, 2$, or 3, as the internal nodes. t_{mn} represents the mapping of response category m on internal node n (see Appendix A). Furthermore, $q_{mn} = 0$ when $t_{mn} = NA$, and $q_{mn} = 1$ otherwise. For simplicity, the nodes in Equation 2 are one-parameter logistic models (1PLM; Rasch, 1966). However, other IRT models can be

used to model the subnodes. As an illustration, Appendix B applies Equation 2 to the response categories in the ISM and the CEM.

Multidimensional structure. In an IRTree, each subnode follows an IRT model, with an item location, and a person latent variable. Although a latent variable can return in multiple subnodes (cf. De Boeck & Partchev, 2012), or, oppositely, a MIRT model can be applied to an individual node (cf. Rose, von Davier, & Nagengast, 2015), we only consider the case where each subnode is modeled as a unidimensional IRT model. Then, the total IRTree is a MIRT model, with the number of latent variables equal to the number of subnodes. In contrast with traditional MIRT models, there is no rotational freedom, because the interpretation of the latent variables is determined by the structure of the IRTree. Furthermore, items have multiple item locations, rather than one; that is, one per subprocess.

The distributions of these latent traits and item locations are especially of interest, because they are indicative of the missingness mechanism. A correlation between (1) the measured proficiency and one of the missingness propensities, or (2) item difficulty and one of the thresholds corresponding to a missingness propensity, indicates that the missingness pattern is MNAR.

Explanatory IRTrees. IRT models can be descriptive or explanatory (De Boeck & Wilson, 2004). The same accounts for the internal nodes in an IRTree. Both item covariates (i.e., linear logistic testing model) and person covariates (i.e., latent regression IRT model) can be introduced into the subnodes. When such covariates are included linearly in an IRTree with 1PLM subprocesses (cf. Equation 2), the model can be formulated as

$$\pi(x_{pi} = m|\boldsymbol{\theta}_p, \boldsymbol{\beta}_i) = \prod_{n=1}^{N} \left[\frac{exp(\theta_p^{(n)} + \beta_i^{(n)} + \sum_g^G \gamma_g^{(n)} Z_{pig})^{t_{mn}}}{1 + exp(\theta_p^{(n)} + \beta_i^{(n)} + \sum_g^G \gamma_g^{(n)} Z_{pig})} \right]^{q_{mn}} . \quad (3)$$

Here, Z_{pig} is the value for person p and item i on the gth covariate, with $g = (1, 2, 3, ..., G)$, and $\gamma_g^{(n)}$ is the coefficient for covariate g within node n. Hence, Z_{pig} can be both an item covariate, a person covariate, or a covariate varying across both persons and items. The included covariates can depend on the assessment or the test population, and they can be included in one or more subnodes. For instance, in the ISM, the item format (e.g., open-ended vs. multiple-choice items) may be a relevant item covariate for the skipping process, but not for the dropping out process. In contrast, item position may play a role in the dropping-out process, but not in the proficiency process. When there are multiple item orders in a test administration, $(k - K)$ in Equation 1 is an example of a covariate Z_{pig} that varies across persons and items.

Process Interpretation

An IRTree is more than a purely mathematical model that formulates probabilities for the response categories. It also represents a belief about the processes underlying an item response. To illustrate the importance of the behavioral process interpretation, consider the IRTrees in Figure 2. In Figure 2(a), the first encountered subnode is the proficiency process. The second subnode, the skipping process, is only passed

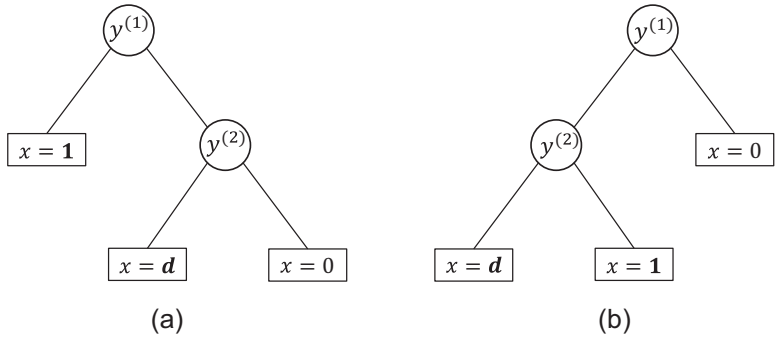


Figure 2. Graphical representations of two IRTrees to model skipped responses jointly, with counterintuitive behavioral process interpretations.

through when the correct response could not have been given. Thus, this IRTree assumes that an item can be skipped, only when the correct response is not known. This assumption corresponds with IAS. Although this IRTree also includes a skipping process and a proficiency process, its behavioral interpretation is different than the IRTree proposed above (cf. Figure 1(a)) and rather counterintuitive. The IRTree in Figure 2(b) assumes that items are only skipped when the correct response is known, which is even more unlikely. Although these IRTrees can model the item responses, their behavioral process interpretation makes them of limited use, if not useless.

The ISM and the CEM model the same response categories. However, due to their process interpretation, they may not suit every situation equally well. The behavioral process of the ISM can be described as follows. When confronted with an item, the test taker first (implicitly) decides whether to continue or to drop out. Next, if not dropped out, the test taker decides whether or not to skip the item. Finally, if not skipped, the test taker tries to solve the item correctly. Given this process interpretation, the ISM seems especially suitable for a speeded test situation, in which test takers are instructed to respond correctly to as many items as possible within a given time limit. The decision to skip an item in these situations may be driven by the test taker's impression of the item's difficulty and the required effort. This is in line with the finding that test takers estimate the amount of effort that is required to respond to a test item (Wolf, Smith, & Birnbaum, 1995).

In contrast, the CEM seems more applicable in low-stakes power tests. According to the CEM, the test taker decides after each nonresponse whether or not to continue and to devote effort to the next item. This process may be driven by the test taker's motivation. Because in low-stakes assessments, test takers' motivation is generally lower, and decreasing during the administration (Debeer, Buchholz, Hartig, & Janssen, 2014; Weirich, Penk, Hecht, Roppelt, & Böhme, 2015), the CEM seems especially suited for these situations.

Estimation and Model Fit

Model estimation. As noted above, the internal nodes in the IRTrees in Figure 1 are binary IRT models, and the total IRTree is a MIRT model. Hence, estimation techniques such as maximum-likelihood estimation (MML) using Gaussian quadrature

(Bock & Aitkin, 1981) or adaptive Gaussian quadrature (Pinheiro & Bates, 1995), and so-called limited information techniques (Jöreskog, 1994) can be applied. When 1PLMs are used for all the subprocesses (cf. Equations 2 and 3), the IRTree is a generalized linear mixed model (GLMM) (De Boeck & Partchev, 2012), and standard statistical packages for GLMM such as the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) in R (R Development Core Team, 2015), and SAS *PROC GLIMMIX* or *PROC NLMIXED* (SAS Institute Inc., 2014) can be used to estimate the parameters in the IRTree. Recently, Jeon and De Boeck (2016) showed that the *flirt* package for R (Jeon, Rijmen, & Rabe-Hesketh, 2014) can be used to fit IRTrees with other IRT models such as the 2PLM or the partial credit model (PCM; Masters, 1982). However, *flirt* does not yet allow covariates that vary over items and persons such as item position.

In addition, the *mdltm* software (von Davier, 2005) uses discrete latent variable distributions (i.e., latent classes) that can be structured using log-linear smoothing (Xu & von Davier, 2008). Analyzing the missingness in a German educational assessment, Köhler, Pohl, and Carstensen (2015) showed that the common distributional assumptions about the latent variables (i.e., multivariate normal) can be relaxed and tested using *mdltm*. A disadvantage, however, is that *mdltm* does not allow adding covariates to the model.

Model comparison and model fit. In order to decide which process interpretation best describes the data, different IRTrees that have the same endnodes can be compared using goodness-of-fit measures such as Akaike's information criterion (AIC) or the Bayesian information criterion (BIC). IRTrees with different endnodes cannot be compared, because their modeled outcome variable differs. Nested IRTrees can be compared using likelihood ratio tests (LRT). For instance, one can test whether the proficiency process is correlated to the missingness process(es) (i.e., MNAR) by comparing an IRTree with freely estimated correlations to the IRTree with the correlations constrained to zero.

Relation to Existing Models

The IRTree approach provides a flexible framework for modeling missing responses. By adding subnodes, using other models for the subnodes, setting latent variables equal across subnodes, and so on, new tailor-made IRTrees can be formulated. In addition, many previously proposed models can be formulated as IRTrees. For instance, (without being exhaustive) the model proposed by Holman and Glas (2005) is equivalent to the IRTree for skipped items (cf. Figure 1(a)) when the skipping process $y^{(1)}$ is modeled as a one-parameter logistic model and the proficiency process $y^{(2)}$ as a two-parameter logistic model (2PLM; Birnbaum, 1968). The model of Glas and Pimentel (2008) is equivalent to the IRTree for not-reached items (cf. Figure 1(b)) with a 1PLM for the dropping out process $y^{(1)}$ and a 2PLM for the proficiency process $y^{(2)}$. The model used by Pohl et al. (2014) is equivalent to an ISM with 1PLM subnodes. To our knowledge, a model similar to the CEM has not yet been proposed.

In the remainder of this article, we demonstrate the applicability of the IRTree framework for modeling nonresponses, using the glmer function (Bates et al., 2014)

function in R (R Development Core Team, 2015). First, using simulated data, we show that ignoring missing responses can lead to biased estimates, and that an IRTree can reduce this bias. Second, as an empirical illustration, we apply the IRTree framework to data form the 2009 PISA reading assessment in Argentina. The results indicate that missingness pattern is MNAR; that the ISM best fits the data; and that the ISM estimates differ from estimates obtained when ignoring the missing responses.

Simulation Study

Both analytic derivations (e.g., Lord, 1974) and simulation studies (e.g., Finch, 2008; Rose et al., 2010) have indicated that scoring missing responses as incorrect results in biased parameter estimates. The impact of ignoring missing responses, however, is less clear. On the one hand, the simulation studies of Holman and Glas (2005) and Glas and Pimentel (2008) found bias on item parameters when, respectively, skipped and not-reached items were ignored under MNAR conditions. Applying the MIRT approach reduced the bias in both studies. On the other hand, Rose et al. (2010) and Pohl et al. (2014) did not find differences in parameter estimations between ignoring missing responses and using a MIRT model in case-based simulation studies. The findings in these studies, however, have some limitations. First, the conclusions of Rose et al. (2010) and Pohl et al. (2014) were based on single and not on replicated simulations. Second, the studies of Holman and Glas (2005) and Glas and Pimentel (2008) considered only one type of response omission, and the proportions of missing responses were simulated to be .25 or .50, which are unlikely to be encountered in low-stakes educational assessments.

In this simulation study, we aimed to investigate the bias on item difficulty and person proficiency estimates (1) when missing responses were ignored, and (2) when missing responses were modeled using an IRTree. We focused on situations where both skipped and not-reached items were present, and where the proportions of missing responses were more realistic: 20% or 10% missing responses.

Data were simulated under MAR or MNAR conditions. When the missingness pattern was MAR, we expected no bias and similar results for ignoring versus IRTree modeling. When the missingness pattern was MNAR, we expected that ignoring would produce bias on item difficulty and that IRTree modeling could reduce the bias. In addition, we investigated the impact of the treatment of missing responses on the proficiency estimates. Although not reported by previous studies, we expected systematic differences between the two treatments when the missingness was MNAR. We found similar results using the ISM and the CEM. However, for brevity, only the results for the ISM are reported, because this model corresponds with the best fitting model in the empirical illustration (see below), and with the model used by Pohl et al. (2014).

Method

Preliminary results using only one item order indicated that the bias due to not-reached responses increased with item position (cf. Glas & Pimentel, 2008). Therefore, like many large-scale assessments, we chose to use multiple item orders across test takers, to counterbalance for these item-position effects. In each simulated test, the items were split into four equally sized blocks. The order of these item blocks

Table 1
Rotated Block Design Used in the Four Orders Condition

Item Order	Order of Blocks				*N*
A	BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4	250
B	BLOCK 2	BLOCK 4	BLOCK 1	BLOCK 3	250
C	BLOCK 3	BLOCK 1	BLOCK 4	BLOCK 2	250
D	BLOCK 4	BLOCK 3	BLOCK 2	BLOCK 1	250

Table 2
Design of the Simulation Study

	20% Missing		10% Missing	
	MAR	MNAR	MAR	MNAR
$I = 20$	1PLM	1PLM	1PLM	1PLM
	ISM	ISM	ISM	ISM
$I = 40$	1PLM	1PLM	1PLM	1PLM
	ISM	ISM	ISM	ISM

was rotated according to the rotated block design in Table 1, resulting in four item orders (A–D). Within each item block the item order was fixed.

Data were simulated in $2 \times 2 \times 2 = 8$ conditions, using $I = 20$ or $I = 40$ items; with 20% or 10% missing responses; under a MAR or a MNAR pattern. Table 2 gives the design of the simulation study. In each condition, 100 data sets were simulated according to the ISM depicted in Figure 2(c) with 1PLM internal nodes. When $I = 20$, the item difficulties ($\beta_i^{(3)}$, with $i = 0, 1, ..., I$) were set to two times 10 equidistant values between -1.5 and 1. The items were randomly assigned to four item blocks, so that the average difficulty in each block was similar, and the correlations between item position and difficulty in each item order were less than .15.[1] When $I = 40$, the difficulty values were repeated twice.

The thresholds for dropping out ($\beta_{ik}^{(1)}$) decreased linearly with item position (cf. Equation 1) and the skipping thresholds were equal for all items (and item positions). The parameters were chosen to produce 20% or 10% missing responses, with equal proportions of skipped and not-reached items.[2]

For each test, 1,000 persons were sampled from a multivariate normal distribution with mean vector $\boldsymbol{\mu}_\theta' = (0, 0, 0)$. The variance-covariance matrix $\boldsymbol{\Sigma}_\theta$ of the multivariate distribution reflected a MAR or a MNAR situation. In the MAR condition, $\boldsymbol{\Sigma}_\theta$ was the identity matrix (i.e., zero correlations between the propensity to drop out, the propensity to skip, and the proficiency). In the MNAR condition, the variances were also set to 1, but the propensity to drop out and the propensity to skip were both negatively related to proficiency ($\rho_{13} = -.50$ and $\rho_{23} = -.50$), and negatively related to each other ($\rho_{12} = -.50$). This covariance structure corresponds to the situation where lower ability test takers are more likely to skip items and drop out before the end of the assessment. In addition, test takers who are more likely to skip items are less likely to drop out.

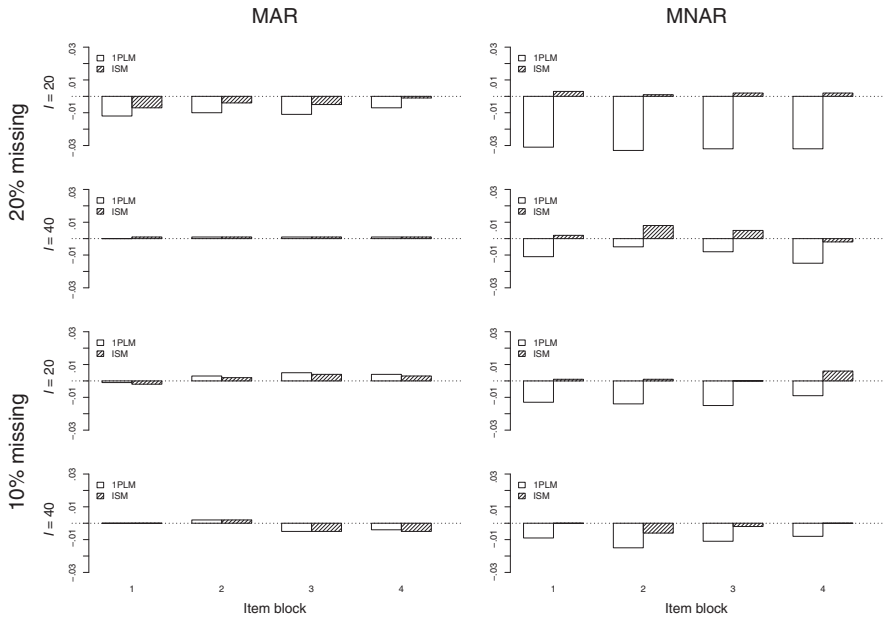


Figure 3. Bias on item difficulty estimates by the 1PLM and the ISM, averaged per item block in the $2 \times 2 \times 2$ conditions.

The sampled test takers were split into four equally sized subsamples, which were assigned to one of the four item orders (cf. Table 1). Each data set was analyzed twice, once using the ISM (i.e., the data generating model) and once using a 1PLM that ignored missing responses. For identifiability, the means for the latent variables were set to zero. Proficiency estimates were obtained using empirical Bayes estimates.

To compare the recovery of the item difficulties, bias and root mean squared error (RMSE) were averaged per item block. To investigate the impact of the missing data treatment on the proficiency estimates, the test takers were divided in seven groups based on the number of missing responses, so that the proportion of test takers in each group is about the same (between 10% and 18%).[3] The average bias and RMSE for the ability estimates were computed within each group. Finally, for the ISM, bias and RMSE were also computed for the other model parameters (i.e., the standard deviations and correlations of the latent traits, the skipping threshold, and the parameters in the linear function of the dropping-out threshold).

Results

The numerical values of the results are found in Appendix C. Figure 3 presents the bias on item difficulty per item block. Under MAR, the 1PLM and the ISM performed equally. Under MNAR, the 1PLM produced negatively biased item difficulties, while the ISM produced unbiased estimates. The bias produced by the 1PLM, however, decreased in conditions with smaller proportions of missing responses and with longer tests. Moreover, in the condition with 10% missing responses and $I = 40$

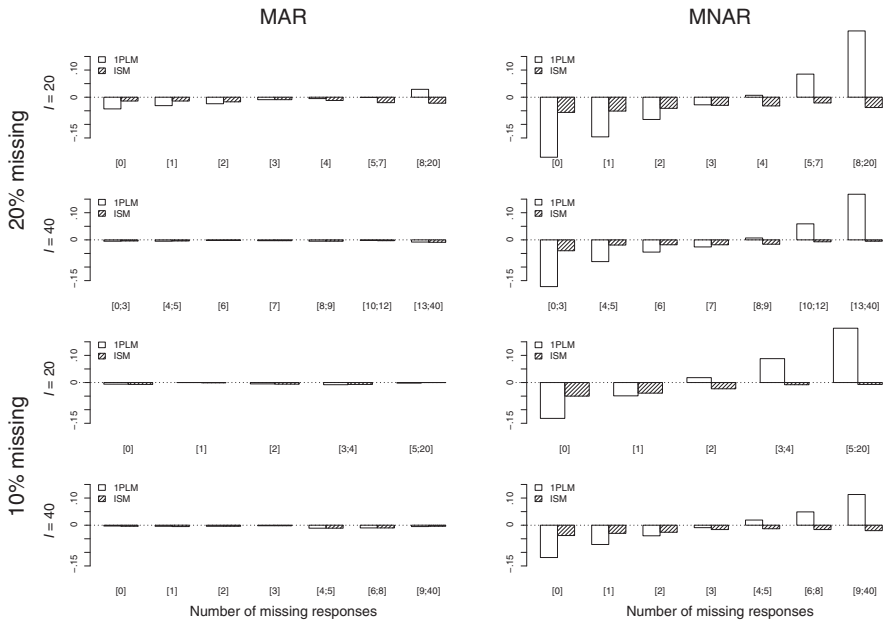


Figure 4. Average bias on the ability estimates of the 1PLM and the ISM, grouped according to number of missing responses in the $2 \times 2 \times 2$ conditions.

the bias on item difficulty became very small. The RMSE decreased with lower proportions of missing responses and longer test forms, both when the missingness was MAR or MNAR.

Regarding the proficiency estimates, no overall bias was found for both the 1PLM and the ISM, even in the MNAR conditions. However, grouping the test takers according to their number of missing responses revealed that the 1PLM proficiency estimates in the MNAR condition were negatively biased for test takers with no (or only a few) missing responses, and positively biased for test takers with high numbers of missing responses (see Figure 4). The ISM, in contrast, produced a small negative bias in the MNAR conditions, and this bias decreased with the number of missing responses per test taker. This pattern was present in all MNAR conditions, but less pronounced when the test was longer ($I = 40$) or when the proportion of missing responses was lower.

Similar simulations with different MNAR patterns for $\mathbf{\Sigma}_\theta$ (not included in the article) indicated that the pattern of bias depends on the MNAR pattern. For instance, when using positive instead of negative covariances in $\mathbf{\Sigma}_\theta$ the pattern in Figure 4 was reversed: ignoring resulted in positively biased estimates for test takers with few missing responses and negatively biased estimates for test takers with high numbers of missing responses. This MNAR pattern, however, seems less likely in real-life data.

Regarding the other model parameters of the ISM, the recovery of the parameters related to the dropping out process was problematic ($\sigma_{\theta^{(1)}}$, ρ_{12}, ρ_{13} and δ_b) when $I = 20$ or when there are only 10% missing responses (see Table C3, Appendix C). In the

condition with $I = 40\%$ and 20% missing responses, the recovery for the dropping-out process was better, but still worse than for the other model parameters. Note that, even in the 20% missing condition, only about 10% of the responses were not reached. Therefore, and because each test taker can only drop out once, the available information about this process was limited, which could explain the problematic recovery.

Conclusions

The results of the simulation study showed that ignoring missing responses can lead to biased item difficulty and person proficiency estimates when the missingness is MNAR. A negative relation between the propensities to skip and to drop out resulted in negatively biased item difficulties. This bias decreased (in absolute value) with longer tests and smaller proportions of missingness. Nevertheless, ignoring missing responses still resulted in biased proficiency estimates. Although there was no bias on the proficiency estimates overall (e.g., Pohl et al., 2014; Rose et al., 2010), grouping the test takers revealed that for some test takers the proficiency estimate was positively biased, while for others it was negatively biased. Modeling the missing responses using the ISM considerably reduced this bias in all conditions.

In data sets with a limited proportion of not-reached items, the dropping-out process in the ISM was not recovered at a satisfactory level. In these situations, the parameters related to the dropping-out process should be interpreted very carefully. However, even in these cases, applying the IRTree framework reduced the bias caused by ignoring missing responses.

Finally, the results suggest that problems with equity can arise when missing responses are ignored in situations where one of the missingness propensities is related to proficiency. Because the bias on proficiency estimates differs across test takers, not all test takers benefit equally from ignoring the missing responses. Moreover, in situations where the correlation structure between proficiency and the missing response propensities is similar to the one used in the simulation study, low-ability test takers that strategically omit are favored by ignoring the missing responses.

Empirical Illustration

The goal of the empirical illustration is to show that the IRTree framework can be used to investigate and model missing data processes in real-life educational assessments; and, further, that the choice for ignoring versus modeling missing responses can have an impact on measurement in practice. The 2009 PISA data in Argentina is a good candidate for an IRTree analysis, because in total about 25% of the responses are missing, and both skipped and not reached items are present. We analyzed data using different IRTrees, and compared the best fitting IRTree (which was an ISM) with a 1PLM that ignored the missing responses.

PISA is a system of large-scale international assessments that focus on reading, mathematics, and science literacy competencies of 15-year olds, with 65 participating countries in 2009. In Argentina, students were drawn through a two-tiered stratified sampling procedure. A total of 176 individual schools were systematically sampled and within each school 35 students were randomly selected, resulting in data for 4,736 students. Ten students did not respond to any of the reading items and

were dropped from the analysis. The final sample consisted of the 4,726 students that responded to at least some of the 93 reading items that were used in Argentina. Although PISA assessments are designed not to be speeded and students should have ample testing time, about 14% of the items was skipped and about 11% not reached in Argentina. More information on the sampling procedure, the items, the response formats, and the scoring rules can be found in the PISA 2009 technical report (OECD, 2012).

The main research questions were the following: (1) Which IRTree best describes the processes underlying the response behavior in the assessment? (2) Is there evidence that the pattern of missing responses is MNAR? (3) Are there differences between the IRTree results and the results of a model that ignored the missing responses?

Method

To fit a binary response model, 13 polytomous items (full credit – partial credit – no credit) were dichotomized by only considering a full credit response as correct, resulting in four possible response categories (i.e., incorrect, correct, skipped, and not reached). Data were analyzed using the ISM and the CEM. Technical details about the parameterization and the item-position covariates are found in Appendix D. To test whether the missingness was MAR, two versions of each IRTree were compared. In the first version, the latent traits were assumed to follow a multivariate normal distribution with the covariances constrained to zero (in accordance with MAR). In the second version, the covariances were unconstrained, allowing MNAR. The two versions are referred to as ISM_{MAR} and ISM_{MNAR}, and CEM_{MAR} and CEM_{MNAR}, respectively. In addition, we analyzed the data using a 1PLM that ignored the missing responses. In line with the 2009 PISA analyses (OECD, 2012), no item discriminations nor guessing parameters were included in the five models. For identification, the mean of the latent variable (vector) was set to zero.

Within each IRTree, the MAR and MNAR versions were nested models, and their fit was compared using an LRT. Across the ISM and the CEM, the model fit was compared using the AIC and the BIC.[4]

Only the best fitting IRTree was compared to the 1PLM. Because the standard deviation of the reading proficiency was a free parameter in both models, the item difficulty and empirical Bayes proficiency estimates of the two models were not necessarily on the same scale. Therefore, in each model, the estimates were standardized using the standard deviation of proficiency.

Results

IRTree selection. Table 3 gives the negative log-likelihood, the AIC, and the BIC of the four IRTrees. Both the AIC and the BIC marked the ISM as the better fitting model. An LRT showed clear evidence that the ISM_{MNAR} better fitted the data than the ISM_{MAR} ($\chi^2(3) = 582.38$, $p < .00001$).

The estimated correlations for the latent traits in the ISM_{MNAR} showed that reading proficiency was negatively related to the propensity to skip items ($r = -.63$), but not—or only slightly negatively—related to the propensity to drop out ($r =$

Table 3
Goodness-of-Fit Measures for the Four IRTrees Applied to the 2009 PISA Reading Assessment Data in Argentina

Model	No. Parameters	$-LL$	AIC	BIC
ISM_{MAR}	191	94,248.82	188,879.65	190,734.9
ISM_{MNAR}	194	93,666.44	187,720.89	189,605.3
CEM_{MAR}	192	95,522.85	191,429.70	193,294.7
CEM_{MNAR}	195	94,817.29	190,024.57	191,918.7

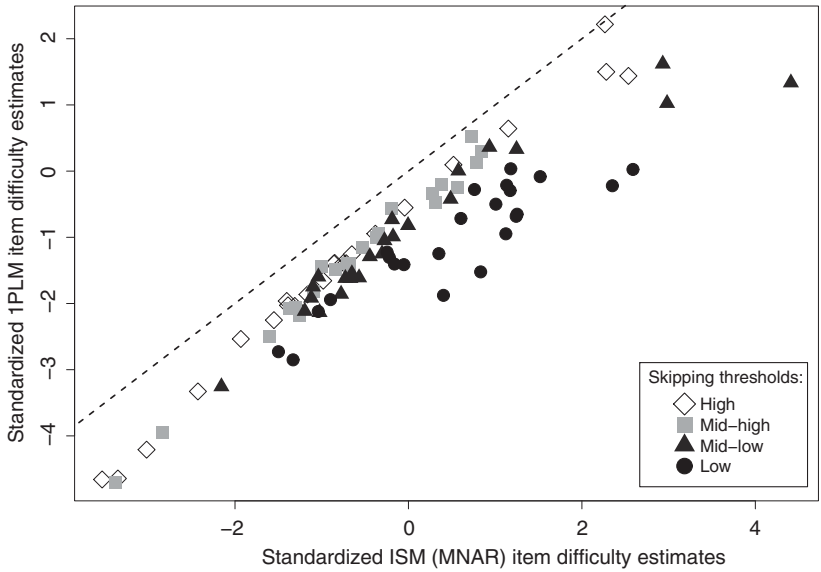


Figure 5. Scatter plot of ISM_{MNAR} versus 1PLM standardized item difficulty estimates for the PISA 2009 reading assessment in Argentina. The items were grouped according to their estimated skipping thresholds. Each of the four groups contains about 25% of the items. The dashed line is the bisect: for points below this line, the 1PLM gave lower item difficulty estimates than the ISM_{MNAR}.

$-.12$). There was a weak positive relation between the propensity to drop out and the propensity to skip ($r = .20$). With respect to the items, a linear regression showed evidence that more difficult items were more likely to be skipped ($F(1, 91) = 15.98$, $p = .00013$). However, only 14% percent of the differences in skipping thresholds were explained by differences in item difficulty $\hat{\omega}^2 = .138$, which suggests that the skipping thresholds may also be related to other features of the items.

Comparison of the ISM_{MNAR} with the 1PLM. The difficulty estimations of the ISM_{MNAR} and the 1PLM were highly correlated ($r = .93$), but, as can be seen in Figure 5, in general the 1PLM had lower item difficulty estimates than the ISM_{MNAR}. Furthermore, by grouping the items according to their estimated skipping thresholds,

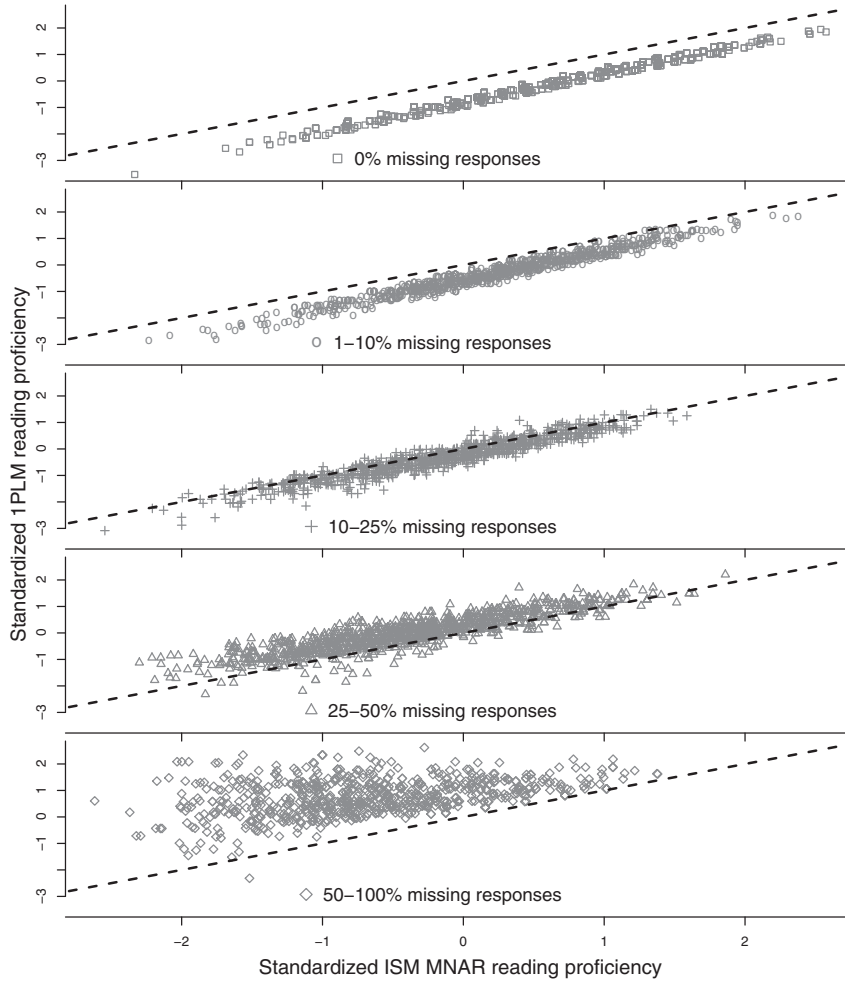


Figure 6. Scatter plots of standardized ISM_{MNAR} versus 1PLM reading ability estimates for the PISA 2009 reading assessment in Argentina. Students were grouped in five groups according to their proportion of missing responses. The dashed line in each plot is the bisect: for points above this line, the 1PLM gave higher ability estimates than the IRTree; for points below this line, the 1PLM gave lower ability estimates.

Figure 5 shows that especially the items with a lower skipping threshold (i.e., items that were more likely to be skipped) had a lower item difficulty estimate by the 1PLM.

Figure 6 gives five scatter plots of the ISM_{MNAR} versus the 1PLM reading proficiency estimates. The students were grouped according to their percentage of missing responses and each plot corresponds with one of the five groups. Figure 6 shows that, for students without missing responses, the ISM consistently produced higher proficiency estimates and the correlation between the estimates was very high. In contrast, for students with a high percentage of missing responses, the ISM gave, on

average, lower proficiency estimates and the correlation between the estimates was low. A table with additional statistics across the five groups is found in Appendix E. Overall, systematic differences in the proficiency estimates were found, according to a pattern similar to the one observed in the simulation study.

Conclusions

With regard to the research questions, the following conclusions can be made. (1) The ISM was the best representation of the response process in the 2009 PISA reading assessment in Argentina. (2) Evidence indicated that the pattern of missing responses was MNAR: reading proficiency was negatively related to the propensity to skip and slightly negatively related to the propensity to drop out. In addition, more difficult reading items were more likely to be skipped. (3) There were systematic differences between the item difficulty and the proficiency estimates of the IRTree and the 1PLM.

The results suggest there was information about the proficiency of the students in the pattern of missing responses. While the 1PLM ignored this information, the IRTree took it into account. For items and persons without missing responses, the available information about the proficiency process was the same in both models, resulting in highly correlated estimates. For items and persons with many missing responses, there was less direct information about the proficiency process. However, the IRTree incorporated additional information from the missingness pattern. This explains the lower correlation between the estimates of both models with respect to these items, and especially these persons.

Note that these findings do not necessarily generalize to PISA reading assessments in other countries, or to other large-scale assessments in Argentina. Different patterns may be found and analyses will have to show whether or not the missingness processes should be included in the model. Finally, the data from Argentina were specifically chosen because they had a large proportion of missing responses. In situations with fewer missing responses, the differences between the estimates can be expected to be smaller, even under MNAR.

General Discussion

An IRTree-based framework to model response omissions was presented and four IRTrees were proposed, two of which jointly model skipped and not reached items. Modeling aspects and the process interpretation were discussed. Subsequently, a simulation study indicated that ignoring missing responses under MNAR can result in biased proficiency and item difficulty estimates, in situations with reasonable proportions of missing responses. Finally, an empirical illustration demonstrated that IRTrees can be used to investigate and model missing responses in real-life assessments.

The remainder of the general discussion is organized as follows. First, the limitations of this study and the presented approach are discussed, with potential solutions and extensions for future research. Afterward, the applicability of the IRTree framework in educational measurement is addressed. we will argue that an IRTree model is suited for proficiency measurements in low-stakes assessments, but not necessarily

in high-stakes assessments. Finally, implications with respect to fairness and strategically omitting responses are discussed.

Limitations and Potential Extensions

The IRTree analysis of the Argentinian PISA data served merely as an illustration; an elaborate discussion and substantial interpretation of the findings were beyond the scope of this article. Nevertheless, future research could further investigate the missingness in PISA across different countries, while focusing on the implications and explanations of the IRTree results. Explanatory IRTrees using background variables and knowledge about the cultural and administrative differences across countries could provide additional insights.

The simulation study and the empirical illustration only considered the 1PLM as model for the IRTree subnodes. The IRTree approach is, however, not limited to 1PLMs subnodes (Jeon & De Boeck, 2016). Future research should investigate the performance of IRTrees for modeling missing responses with other IRT models (e.g., the 2PLM, polytomous IRT models). Given the work of Holman and Glas (2005), Glas and Pimentel (2008), and Glas et al. (2015), results similar to what was found in our simulation study can be expected.

In the empirical illustration, we assumed a multivariate normal distribution for the missingness propensities. A violation of this assumption could pose a threat to the validity of the IRTree analysis. One way nonnormality could arise is when the population consists of two (or more) groups of persons with a different mean missingness propensity. In these cases, the relevant characteristics (i.e., group membership) can be included in the model (see Equation 3). Similarly, item characteristics can be included in the model. Differential item functioning (or differential facet functioning) at the level of the missing data processes can also be modeled. In other cases of nonnormality, the distributional assumptions can be relaxed and tested by using a discrete latent variable distribution that can be structured using log-linear smoothing (Köhler et al., 2015). In essence, methods that can deal with violations of the distributional assumptions in IRT models can also be applied to IRTree models.

Applicability

The simulation study showed that, when one is interested in reliably modeling the missingness processes using the IRTree framework, the percentage of skipped or not reached responses should be high enough. It was found that at least 10% of the items should be not reached to include the dropping-out process in the ISM, while 5% skipped items was sufficient to include the skipping process. Note, however, that even in the conditions with 5% skipped and 5% not-reached items, the ISM reduced the bias in the proficiency estimates produced by ignoring the missing responses.

The IRTree approach is especially suited for low-stakes assessments. In these assessments, no personal consequences are attached to the performance of the test takers, and therefore, the treatment of missing responses does not affect the test takers individually. For test administrators and policy makers, however, these assessments have high stakes, and fair and unbiased measurement is of great

importance. Hence, in situations with enough missing responses, the IRTree framework can have an added value.

In contrast, the IRTree framework may not be suited to obtain performance measurements in some high-stakes assessments. In administrations where the treatment of missing responses has direct consequences for the test takers, this treatment, usually a scoring rule, should be secured well before the assessment. Test takers should be properly informed about the treatment or scoring rule, so that they can adjust their response behavior accordingly. Therefore, IRTree proficiency estimates, for which the treatment of missing responses is based on the data, seem unethical in these situations.

However, even if one does not want to use an IRTree for the proficiency estimates, an IRTree analysis can help in understanding the missingness pattern in a specific assessment, whether it is low-stakes or high-stakes. In low-stakes assessments, the results of IRTree analyses can be used to make a reasoned decision on how to treat the missing responses. In high-stakes assessments, the IRTree approach can be used to investigate whether the chosen scoring rule is eliciting the desired response behavior. For instance, when it is communicated that omissions will be scored as incorrect (i.e., IAS or number-correct scoring), IRTree analyses can indicate whether the propensity to skip or drop out is indeed related to the proficiency. If this is not the case, one can decide to alter the instructions and put more emphasis on the scoring rule in future administrations.

Strategically Omitting and Fairness

In educational measurement, great importance is attached to fair and impartial testing. Because an IRTree takes into account the missingness pattern of test takers in the proficiency estimation (if proficiency is related to at least one of the missingness propensities), the question arises whether test takers can "game the system" and benefit from strategically skipping items or dropping out. Only in the cases where proficiency is positively related to one or more missingness propensities (i.e., higher-ability test takers are more likely to skip or to not reach items), a low-ability test taker could benefit from strategically omitting responses to items for which he/she does not know the answer. Note, however, that a positive correlation between proficiency and a missingness propensity is highly unlikely. Furthermore, to be able to strategically omit, the test taker should somehow know in advance that in the current administration and sample a positive correlation will be present. Moreover, if more low-ability test takers start omitting responses strategically, this would automatically result in a smaller, or even a negative correlation between proficiency and the missingness propensities.

Interestingly, gaming the system by strategically omitting responses is more likely to occur when missing responses are ignored. For instance, if during a certain administration many low-ability test takers would strategically skip items or drop out, a correlation pattern similar to the one used in the simulation study would arise (i.e., a negative correlation between proficiency and the propensities to skip and drop out). In these cases, ignoring missing responses results in overestimated proficiency estimates for test takers with many missing responses. Hence, low-ability test takers that

strategically omit responses would be favored by ignoring the response omissions. Using an IRTree would reduce this bias.

Deciding on a strategy to handle missing responses is not an arbitrary choice. Different strategies may favor some students while disadvantaging others. Therefore, the treatment of missing responses should depend on the mechanism that is causing the omissions (Huisman, 2000). Because IRTree models can model patterns of MNAR in a natural and interpretable way, they can help in pursuing fair and impartial measurement.

Acknowledgments

The research leading to the results reported in this paper was supported in part by the Research Fund of KU Leuven (GOA/15/003) and by the Interuniversity Attraction Poles programme financed by the Belgian government (IAP/P7/06).

Appendix A

This appendix discusses the mapping of the response categories onto the subitems according to the four IRTrees for nonresponses that were presented in the main article. The graphical representations of the four IRTrees are given in Figure A1.

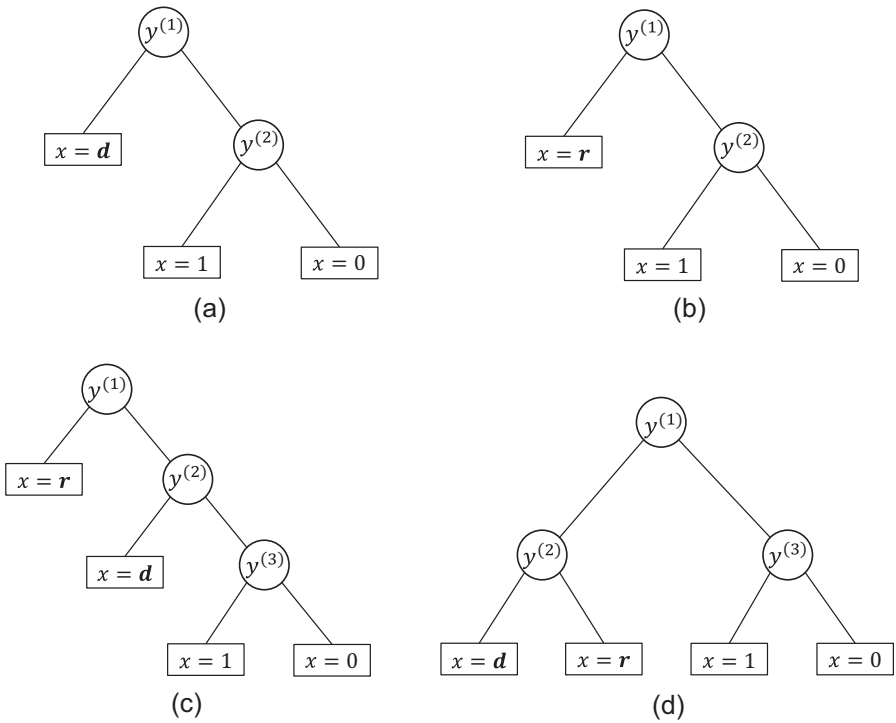


Figure A1. Graphical representations of (a) an IRTree for skipped items; (b) an IRTree for not-reached items; (c) the item selection model (ISM); and (d) the continuing effort model (CEM).

Table A1
Mapping Matrix the IRTree for Skipped Items

	Subitem Scores	
Response Category	Subitem $y^{(1)}$	Subitem $y^{(2)}$
Skipped ($x = d$)	1	9
Correct ($x = 1$)	0	1
Incorrect ($x = 0$)	0	0

Table A2
Mapping Matrix of the IRTree for Not-Reached Items

	Subitem Scores	
Response Category	Subitem $y^{(1)}$	Subitem $y^{(2)}$
Not Reached ($x = r$)	1	9
Correct ($x = 1$)	0	1
Incorrect ($x = 0$)	0	0

Table A3
Mapping Matrix of the Item Selection Model

	Subitem Scores		
Response Category	Subitem $y^{(1)}$	Subitem $y^{(2)}$	Subitem $y^{(3)}$
Not Reached ($x = r$)	1	9	9
Skipped ($x = d$)	0	1	9
Correct ($x = 0$)	0	0	1
Incorrect ($x = 1$)	0	0	0

Because all the presented IRTrees are binary, there are three possible *scores* for the subitems: 0 when the left branch is taken to leave the subnode; 1 when the right branch is taken; or 9 when the path through the tree does not pass through the subnode. The resulting mapping matrices are presented in Tables A1–A4, for (A1) the IRTree for skipped items; (A2) an IRTree for not-reached items; (A3) the item selection model (ISM); and (A4) the continuing effort model (CEM), respectively.

For each of the four IRTrees, Table A5 provides sample data with the corresponding subitem scores, for a test with six items and four test takers. Note that, after the first not-reached item (r), the test taker has dropped out. Therefore, the remaining items are not attempted, and none of the subitems is passed through, resulting in all nine scores. Although the sample data for the ISM (c) and the CEM (d) in Table A5 is the same, the corresponding subitem scores differ between the two models.

Table A4
Mapping Matrix of the Continuing Effort Model

Response Category	Subitem Scores		
	Subitem $y^{(1)}$	Subitem $y^{(2)}$	Subitem $y^{(3)}$
Not Reached ($x = r$)	1	0	9
Skipped ($x = d$)	1	1	9
Correct ($x = 0$)	0	9	1
Incorrect ($x = 1$)	0	9	0

Table A5
Sample Data With Their Mapping Onto the Subitems for the Four Presented IRTrees

		Response to Item i						Subitems $y_i^{(1)}$						Subitems $y_i^{(2)}$						Subitems $y_i^{(3)}$					
	Persons	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
(a)	1	0	1	1	*d*	1	0	0	0	0	1	0	0	0	1	1	9	1	0						
	2	1	0	*d*	1	*d*	0	0	0	1	0	1	0	1	0	9	1	9	0						
	3	1	1	0	1	0	*d*	0	0	0	0	0	1	1	1	0	1	0	9						
	4	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1	0	1	1						
(b)	1	0	1	1	*r*	*r*	*r*	0	0	0	1	9	9	0	1	1	9	9	9						
	2	1	0	*r*	*r*	*r*	*r*	0	0	1	9	9	9	1	0	9	9	9	9						
	3	1	1	0	1	*r*	*r*	0	0	0	0	1	9	1	1	0	1	9	9						
	4	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1	0	1	1						
(c)	1	0	*d*	1	0	*d*	1	0	0	0	0	0	0	0	1	0	0	1	0	0	9	1	0	9	1
	2	1	0	*r*	*r*	*r*	*r*	0	0	1	9	9	9	0	0	9	9	9	9	1	0	9	9	9	9
	3	1	*d*	*d*	1	*r*	*r*	0	0	0	0	1	9	1	0	0	1	9	9	1	9	9	1	9	9
	4	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1
(d)	1	0	*d*	1	0	*d*	1	0	1	0	0	1	0	9	1	9	9	1	9	0	9	1	0	9	1
	2	1	0	*r*	*r*	*r*	*r*	0	0	1	9	9	9	9	9	0	9	9	9	1	0	9	9	9	9
	3	1	*d*	*d*	1	*r*	*r*	0	1	1	0	1	9	9	1	1	9	0	9	1	9	9	1	9	9
	4	0	1	0	0	1	1	0	0	0	0	0	0	9	9	9	9	9	9	0	1	0	0	1	1

Notes. The four IRTrees are (a) the IRTree for skipped items; (b) the IRTree for not-reached items; (c) the ISM; and (d) the CEM. "9" marks "did not pass through this subnode."

Appendix B

This appendix discusses the probability functions of the ISM and the CEM, for dichotomous items. For simplicity, the subprocesses are assumed to follow the one-parameter logistic model (1PLM). Then, the probabilities of the response categories can be formulated as

$$\pi(x_{pi} = m|\boldsymbol{\theta}_p, \boldsymbol{\beta}_i) = \prod_{n=1}^{N} \left[\frac{exp(\theta_p^{(n)} + \beta_i^{(n)})^{t_{mn}}}{1 + exp(\theta_p^{(n)} + \beta_i^{(n)})} \right]^{q_{mn}}, \tag{B1}$$

Table B1
Probabilities of the Response Categories in Function of the Subitem Scores According to the Item Selection Model and the Continuing Effort Model

IRTree	Response	Probability
ISM	$x_{pi} = r$	$\pi(y_{pi}^{(1)} = 1 \mid \theta_p^{(1)}, \beta_i^{(1)})$
	$x_{pi} = d$	$\pi(y_{pi}^{(1)} = 0 \mid \theta_p^{(1)}, \beta_i^{(1)}) \times \pi(y_{pi}^{(2)} = 1 \mid \theta_p^{(2)}, \beta_i^{(2)})$
	$x_{pi} = 0$	$\pi(y_{pi}^{(1)} = 0 \mid \theta_p^{(1)}, \beta_i^{(1)}) \times \pi(y_{pi}^{(2)} = 0 \mid \theta_p^{(2)}, \beta_i^{(2)}) \times \pi(y_{pi}^{(3)} = 0 \mid \theta_p^{(3)}, \beta_i^{(3)})$
	$x_{pi} = 1$	$\pi(y_{pi}^{(1)} = 0 \mid \theta_p^{(1)}, \beta_i^{(1)}) \times \pi(y_{pi}^{(2)} = 0 \mid \theta_p^{(2)}, \beta_i^{(2)}) \times \pi(y_{pi}^{(3)} = 1 \mid \theta_p^{(3)}, \beta_i^{(3)})$
CEM	$x_{pi} = r$	$\pi(y_{pi}^{(1)} = 1 \mid \theta_p^{(1)}, \beta_i^{(1)}) \times \pi(y_{pi}^{(2)} = 0 \mid \theta_p^{(2)}, \beta_i^{(2)})$
	$x_{pi} = d$	$\pi(y_{pi}^{(1)} = 1 \mid \theta_p^{(1)}, \beta_i^{(1)}) \times \pi(y_{pi}^{(2)} = 1 \mid \theta_p^{(2)}, \beta_i^{(2)})$
	$x_{pi} = 0$	$\pi(y_{pi}^{(1)} = 0 \mid \theta_p^{(1)}, \beta_i^{(1)}) \times \pi(y_{pi}^{(3)} = 0 \mid \theta_p^{(3)}, \beta_i^{(3)})$
	$x_{pi} = 1$	$\pi(y_{pi}^{(1)} = 0 \mid \theta_p^{(1)}, \beta_i^{(1)}) \times \pi(y_{pi}^{(3)} = 1 \mid \theta_p^{(3)}, \beta_i^{(3)})$

Notes. $\theta_p^{(n)}$ and $\beta_i^{(n)}$ with $n = 1, 2$ or 3, are, respectively, the person location (i.e., the propensity) and the item location (i.e., the threshold) on the latent variable scale that is relevant for node n.

with x_{pi} the response of person p on item i, $m = 0, 1, d$, or r, as the response categories, and $n = 1, 2$, or 3, as the internal nodes. t_{mn} represents the mapping of response category m on internal node n (for the mapping matrices of the ISM and the CEM, see Appendix A). Furthermore, $q_{mn} = 0$ when $t_{mn} = NA$, and $q_{mn} = 1$ otherwise. Table B1 presents the probabilities of the response categories based on Equation B1, according to the ISM and the CEM, respectively.

Appendix C

This appendix presents the results of the simulation study. In each condition of the $2 \times 2 \times 2$ conditions 100 data sets were simulated with 1, 000 test takers. Table C1 gives the bias and RMSE of the estimated item difficulty parameters according to both the 1PLM (which ignored the missing responses) and the ISM. Table C2 lists the bias and RMSE of the empirical Bayes proficiency estimates of both models. The statistics are grouped according to the number of missing responses of the

Table C1
Bias and RMSE of the Item Difficulty Estimates in the Simulation Study

Condition			MAR				MNAR			
		Item	1PLM		ISM		1PLM		ISM	
Items	Missing	Block	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
$I = 20$	20%	$\beta_{1-5}^{(3)}$	−.012	.076	−.007	.074	−.031	.075	.003	.071
		$\beta_{6-10}^{(3)}$	−.010	.078	−.004	.077	−.033	.074	.001	.071
		$\beta_{11-15}^{(3)}$	−.011	.073	−.005	.072	−.032	.076	.002	.072
		$\beta_{16-20}^{(3)}$	−.007	.071	−.001	.070	−.032	.076	.002	.071

(*Continued*)

Table C1
Continued

Condition		Item	MAR				MNAR			
			1PLM		ISM		1PLM		ISM	
Items	Missing	Block	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
	10%	$\beta_{1-5}^{(3)}$	−.001	.068	−.002	.068	−.013	.070	.001	.071
		$\beta_{6-10}^{(3)}$	.003	.068	.002	.069	−.014	.068	.001	.069
		$\beta_{11-15}^{(3)}$	.005	.065	.004	.065	−.015	.066	.000	.066
		$\beta_{16-20}^{(3)}$	.004	.064	.003	.064	−.009	.067	.006	.068
$I = 40$	20%	$\beta_{1-10}^{(3)}$	.000	.072	.001	.072	−.011	.074	.002	.074
		$\beta_{11-20}^{(3)}$	.001	.071	.001	.071	−.005	.070	.008	.070
		$\beta_{21-30}^{(3)}$	.001	.071	.001	.071	−.008	.075	.005	.075
		$\beta_{31-40}^{(3)}$	.001	.070	.001	.070	−.015	.068	−.002	.067
	10%	$\beta_{1-5}^{(3)}$	.000	.068	.000	.068	−.009	.068	.000	.067
		$\beta_{6-10}^{(3)}$	.002	.068	.002	.068	−.015	.067	−.006	.065
		$\beta_{11-15}^{(3)}$	−.005	.067	−.005	.067	−.011	.070	−.002	.069
		$\beta_{16-20}^{(3)}$	−.004	.067	−.005	.067	−.008	.065	.000	.064

Notes. The bias and RMSE were computed over 100 replications and averaged per item block.

Table C2
Bias and RMSE of the Proficiency Estimates in the Simulation Study

Condition		Item	MAR				MNAR			
			1PLM		ISM		1PLM		ISM	
Items	Missing	Block	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
$I = 20$	20%	[0]	−.043	.475	−.014	.467	−.221	.522	−.056	.477
		[1]	−.031	.476	−.014	.473	−.146	.485	−.051	.466
		[2]	−.024	.485	−.017	.484	−.082	.471	−.041	.465
		[3]	−.009	.488	−.009	.487	−.028	.469	−.030	.467
		[4]	−.005	.496	−.012	.495	.007	.481	−.032	.478
		[5;7]	−.001	.526	−.020	.525	.085	.513	−.021	.500
		[8;20]	.029	.627	−.022	.616	.244	.675	−.038	.605
	10%	[0]	−.006	.467	−.007	.467	−.132	.482	−.050	.467
		[1]	.000	.472	−.001	.472	−.049	.462	−.039	.461
		[2]	−.005	.483	−.006	.483	.018	.470	−.023	.470
		[3;4]	−.008	.501	−.007	.501	.088	.488	−.008	.478
		[5;20]	−.002	.548	.000	.549	.200	.580	−.007	.533
$I = 40$	20%	[0;3]	−.005	.360	−.004	.360	−.172	.416	−.040	.378
		[4;5]	−.005	.369	−.004	.369	−.080	.369	−.019	.358
		[6]	−.002	.375	−.002	.375	−.045	.361	−.018	.355
		[7]	−.003	.377	−.003	.377	−.026	.362	−.018	.358

(*Continued*)

Table C2
Continued

Condition		Item	MAR				MNAR			
			1PLM		ISM		1PLM		ISM	
Items	Missing	Block	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
		[8;10]	−.005	.388	−.005	.388	.007	.369	−.016	.365
		[11;12]	−.002	.405	−.003	.405	.059	.384	−.007	.375
		[13;40]	−.008	.441	−.009	.441	.168	.465	−.005	.419
	10%	[0]	−.003	.353	−.004	.353	−.119	.386	−.038	.369
		[1]	−.004	.354	−.005	.354	−.071	.358	−.030	.352
		[2]	−.004	.362	−.004	.362	−.039	.348	−.026	.346
		[3]	−.002	.361	−.002	.361	−.009	.348	−.016	.347
		[4;5]	−.011	.366	−.011	.366	.019	.353	−.013	.351
		[6;8]	−.010	.377	−.010	.378	.049	.371	−.016	.366
		[9;40]	−.005	.409	−.004	.409	.113	.429	−.020	.405

Notes. The test takers were grouped according to their number of missing responses, the bias and RMSE were computed within groups.

Table C3
Bias and RMSE for the Additional Parameters in the ISM and the 1PLM

		$I = 20$				$I = 40$			
		MAR		MNAR		MAR		MNAR	
	Parameter	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
20% *missing*	$\sigma_{\theta(3)}$	−.011	.037	−.007	.033	−.003	.029	−.006	.032
	$\sigma_{\theta(2)}$	−.009	.044	.001	.043	−.011	.036	−.007	.033
	$\sigma_{\theta(1)}$	.338	.476	.660	.681	−.172	.193	.087	.114
	ρ_{12}	−.217	.350	−.202	.239	−.018	.069	−.045	.086
	ρ_{13}	−.122	.288	−.138	.167	−.004	.055	−.010	.065
	ρ_{23}	−.085	.222	−.000	.055	−.005	.043	−.001	.046
	δ_b	.020	.103	.002	.106	.175	.209	−.084	.135
	δ_a	.034	.044	.064	.068	−.032	.038	.017	.027
	β_{2i}	−.008	.053	−.009	.052	.005	.039	.006	.038
10% *missing*	$\sigma_{\theta(3)}$	−.006	.033	−.002	.032	−.006	.027	−.001	.029
	$\sigma_{\theta(2)}$	.005	.057	.029	.057	−.017	.041	.003	.042
	$\sigma_{\theta(1)}$	.652	1.780	2.212	3.320	.258	.380	.576	.642
	ρ_{12}	−.158	.229	−.411	.503	−.074	.128	−.179	.219
	ρ_{13}	.012	.098	−.274	.362	.005	.083	−.118	.171
	ρ_{23}	.002	.055	−.003	.056	.004	.050	−.001	.049
	δ_b	.261	.707	.779	1.360	.040	.094	.008	.106
	δ_a	.124	.451	.408	.702	.021	.030	.054	.061
	β_{2i}	−.015	.063	−.026	.061	.001	.053	−.001	.051

Notes. In the 1PLM, the standard deviation of proficiency was the only extra model estimated parameter.

corresponding (simulated) test takers. Finally, Table C3 gives the bias and RMSE of the other model parameters in the ISM, and the standard deviation of proficiency in the 1PLM.

Appendix D

This appendix discusses additional details about the four IRTrees that were used to analyze the PISA 2009 reading assessment data from Argentina. The following aspects are discussed: the item-position covariate and the parameterization of the item thresholds in the (best fitting) ISM an CEM.

The item-position covariate $\boldsymbol{q}$. In the PISA 2009 assessment the items were partitioned in 13 item blocks: seven for reading (R1–R7), three for mathematics (M1–M3), and three for science (S1–S3). Each item block represented 30 minutes of test time. Item blocks were presented in 13 test booklets (Booklets 1–13) with each booklet being composed of four blocks (Table D1). Using a balanced incomplete block design, each item block appeared once in each of the four possible block positions within a test booklet. Therefore, items were administered in different positions across the test takers. Within each item block items always appeared in the same order.

Because the number of items is different across item blocks (and across test booklets), the position of an item within a test booklet was not comparable across test booklets. Therefore, instead of using the position of the items itself, we expressed item position as the proportion of items (q) that was already administered in the booklet, including that item. As there were four item blocks in a test booklet, each

Table D1
Visual Representation of the PISA 2009 Rotated Block Design

Booklet	Block 1	Block 2	Block 3	Block 4
1	M1	R1	R3A	M3
2	R1	S1	R4A	R7
3	S1	R3	M2	S3
4	R3	R4	S2	R2
5	R4	M2	R5	M1
6	R5	R6	R7	R3
7	R6	M3	S3	R4
8	R2	M1	S1	R6
9	M2	S2	R6	R1
10	S2	R5	M3	S1
11	M3	R7	R2	M2
12	R7	S3	M1	S5
13	S3	R2	R1	R5

Notes. Each booklet consists of four of the 13 item blocks. There are four block positions, and each item cluster is administered at every block position once.

item block represented a proportion of $\frac{1}{4}$ of the complete booklet. The proportion that an item represents within an item block depends on the number of items in that item block. Following this rationale, an item appearing in the last position of the second item block in a test booklet has a $q = \frac{1}{4} + \frac{1}{4} = .50$ regardless of the number of items in the four item blocks. An item appearing in the first position of a third item block consisting of 20 items has $q = \frac{1}{4} + \frac{1}{4} + \frac{1}{20} = .5125$.

The parameterization of the thresholds in the ISM and CEM. Different ISMs and CEMs were fit to the data. The versions differed with respect to the parameterization of the item thresholds: item thresholds in each node could be modeled as an item-position effect, as an item content effect, or as a combination of both. Note that discerning an item-position effect from an item content effect is only possible when some items are presented in different item positions across the test booklets. Only the parameterization of the best fitting ISM and CEM are discussed.

The following item parameters were included in the two versions of the ISM (i.e., ISM_{MAR} and ISM_{MNAR}): an item difficulty ($\beta_i^{(3)}$) and a skipping threshold ($\beta_i^{(2)}$) for every item, and two parameters (δ_a) and (δ_b) for the dropping-out thresholds ($\beta_q^{(1)}$) that were modeled as a linear function of item position:

$$\beta_q^{(1)} = \delta_a + (q-1)\delta_b, \tag{D1}$$

where q is the proportion of items that is already administered in the booklet. Hence, δ_a is the dropping-out threshold for the last item in a booklet, and $\delta_a - \delta_b$ is the (theoretical) threshold for dropping out before addressing the first item in a booklet.

The parameters included in the CEM_{MAR} and CEM_{MNAR} were an item difficulty ($\beta_i^{(3)}$) and an omitting threshold ($\beta_i^{(1)}$) for every item, and two parameters for the continuing thresholds ($\beta_q^{(2)}$), so that an increasing probability to drop out when not responding was modeled without having to estimate a threshold for all items separately:

$$\beta_q^{(2)} = \delta_a + (q-1)\delta_b. \tag{D2}$$

In addition, on top of the omitting threshold for a specific item, a position effect was added to the omitting process to model an increasing probability to omit an item during the assessment. The resulting omitting threshold for an item i in position q, $\beta_{iq}^{(1)}$, was therefore modeled as the sum of an item-specific effect ($\beta_i^{(1)}$) and a linear function of item position ($(1-q)\delta^{(1)}$):

$$\beta_{iq}^{(1)} = \beta_i^{(1)} + (q-1)\delta^{(1)}. \tag{D3}$$

Appendix E

This appendix presents additional results from the empirical illustration. The 2009 PISA reading assessment data from Argentina were analyzed using the ISM and a 1PLM that ignored the missing responses. To compare the proficiency estimates across the two models, students were divided in five groups, according to their proportion of nonresponses. Table E1 lists the following five statistics per group as well as across groups: standard deviation of the reading proficiency estimates of the two models; average and standard deviation of the difference between the ISM_{MNAR} and

the 1PLM reading proficiency estimates; and the correlation between the ISM_{MNAR} and the 1PLM reading proficiency estimates.

Table E1
Grouped and Overall Statistics for the Standardized Proficiency Estimates by the ISM_{MNAR} and the 1PLM in the 2009 PISA Reading Assessment in Argentina

		SD Proficiency Estimates		ISM_{MNAR} versus 1PLM Estimates		
% Missing	% Students	ISM MAR	1PLM	Mean Difference[a]	*SD* Difference[a]	Correlation
0	24.35	.795	.836	.721	.132	.99
1–10	19.28	.717	.805	.524	.189	.98
10–25	19.83	.679	.722	.173	.224	.95
25–50	21.05	.741	.696	−.364	.339	.89
50–100	15.49	.750	.672	−1.434	.788	.39
Overall	100.00	.863	.838	.012	.820	.54

[a]Standardized ISM_{MNAR} minus standardized 1PLM reading proficiency estimates.

Notes

[1]The item difficulties per item block were: BLOCK 1: 1, −1.5, .722, −1.5, and −.389; BLOCK 2: −.944, .444, −.389, 1, and −1.222; BLOCK 3: −.111, .722, −1.222, −.111, and −.667; and BLOCK 4: .444, −.667, .167, −.994, and .167.

[2]The following parameters were used. For the 20% conditions: $\delta_a = 2$ and $\delta_b = -.3$, and $\beta_{2i} = 2.5$ for $i = 1, 2, ..., I$ when $I = 20$; and $\delta_a = -.5$ and $\delta_b = -.5$, and $\beta_{2i} = 2.5$ for $i = 1, 2, ..., I$ when $I = 40$. For the 10% conditions: $\delta_a = 2$ and $\delta_b = -.5$, and $\beta_{2i} = 3.3$ for $i = 1, 2, ..., I$ when $I = 20$; and $\delta_a = 2$ and $\delta_b = -.3$, and $\beta_{2i} = 3.3$ for $i = 1, 2, ..., I$ when $I = 40$.

[3]There were five instead of seven groups of test takers in the $I = 20$ condition with 10% missing responses, because it was impossible to have seven groups of about equal size.

[4]In the computation of the BIC, not-reached items after the first not-reached item of a test taker are not included in the total number of observations. The fit of the 1PLM could not be compared with the fit of the IRTrees because they did not share the same outcome variable.

References

Batchelder, W. H. (2010). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 71–93). Washington, DC: American Psychological Association.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Birnbaum, A. (1968). Some latent ability models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, *39*, 502–523.

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, NY: Springer.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*, 225–245.

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

Glas, C. A. W., Pimentel, J. L., & Lamers, S. M. A. (2015). Nonignorable missing data in IRT models: Polytomous responses and response propensity models with covariates. *Psychological Test and Assessment Modeling*, *57*, 523–541.

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

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality and Quantity*, *34*, 331–351.

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, *48*, 1070–1085.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2014). Flexible item response theory modeling with FLIRT. *Applied Psychological Methods*, *38*, 404–405.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*, 381–389.

Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement*, *75*, 850–874.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: John Wiley.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*, 247–264.

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

Mislevy, R. J., & Wu, P.-K. (1988). *Inferring examinee ability when some item responses are missing* (Research Report). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Research Report). Princeton, NJ: Educational Testing Service.

OECD. (2009). *PISA 2006 technical report*. Paris, France: OECD Publishing.

OECD. (2012). *PISA 2009 technical report*. Paris, France: OECD Publishing.

Okumura, T. (2014). Empirical differences in omission tendency and reading ability in PISA: An application of tree-based item response models. *Educational and Psychological Measurement*, *74*, 611–626.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*(1), 12–35.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, *74*, 423–452.

R Development Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: Author.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, *19*, 49–57.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* (Dissertation). Friedrich-Schiller-Universität Jena, Jena, Germany.

Rose, N., von Davier, M., & Nagengast, B. (2015). Commonalities and differences in IRT-based methods for nonignorable item nonresponses. *Psychological Test and Assessment Modeling*, *57*, 472-498.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (Research Report). Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

SAS Institute Inc. (2014). *SAS/STAT 13.2 Users Guide*. Cary, NC: Author.

Stocking, M. L., Eignor, D. R., & Cook, L. L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed- and true-score equating procedures* (Research Report). Princeton, NJ: Educational Testing Service.

von Davier, M. (2005). mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models [Computer software]. Princeton, NJ: Educational Testing Service.

Weirich, S., Penk, C., Hecht, M., Roppelt, A., & Böhme, K. (2015, April). *Item-position effects are moderated by changes in test-taking effort*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test, motivation, and mentally taxing items. *Applied Measurement in Education*, *8*, 341–351.

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model to NAEP data* (Research Report). Princeton, NJ: Educational Testing Service.

Authors

DRIES DEBEER is Postdoctoral Assistant at the University of Zurich, Department of Psychology, Research Unit Psychological Methods, Evaluation and Statistics, Binzmuehlestrasse 14 (PB 27), 8050 Zurich, Switzerland; dries.debeer@uzh.ch. His primary research interests include psychometrics, item response modeling, and statistical methods in psychology.

RIANNE JANSSEN is Associate Professor at the Faculty of Psychology and Educational Sciences, KU Leuven, Dekenstraat 2 (PB 3773), 3000 Leuven, Belgium; rianne.janssen@kuleuven.be. Her current research interests include psychometrics and educational measurement.

PAUL DE BOECK is Emeritus Professor at KU Leuven and Full Professor at the Ohio State University, Department of Psychology, 232 Lazenby Hall, 1827 Neil Avenue, 43210 Columbus, OH; deboeck.2@osu.edu. His primary research interests include item response models.