Educational and Psychological Measurement
Volume 68 Number 6
December 2008 907-922
© 2008 Sage Publications
10.1177/0013164408315262
http://epm.sagepub.com
hosted at
http://online.sagepub.com

Modeling Nonignorable Missing Data in Speeded Tests

Cees A. W. Glas
Jonald L. Pimentel
University of Twente

In tests with time limits, items at the end are often not reached. Usually, the pattern of missing responses depends on the ability level of the respondents; therefore, missing data are not ignorable in statistical inference. This study models data using a combination of two item response theory (IRT) models: one for the observed response data and one for the missing data indicator. The missing data indicator is modeled using a sequential model with linear restrictions on the item parameters. The models are connected by the assumption that the respondents' latent proficiency parameters have a joint multivariate normal distribution. Model parameters are estimated by maximum marginal likelihood. Simulations show that treating missing data as ignorable can lead to considerable bias in parameter estimates. Including an IRT model for the missing data indicator removes this bias. The method is illustrated with data from an intelligence test with a time limit.

Keywords: *ignorability; item response theory; maximum marginal likelihood; missing data; nonignorable missing data; sequential model; steps model*

Missing data usually create problems for statistical analyses, especially when ignorability of the missing data does not hold (Heitjan, 1994; Little & Rubin, 1987; Rubin, 1976). This article deals with responses x_{nk} of respondents labeled n $(n = 1, \ldots, N)$ to items labeled k $(k = 1, \ldots, K)$. A missing data indicator d_{nk} is defined as $d_{nk} = 1$ if a realization x_{nk} was observed and $d_{nk} = 0$ if x_{nk} was missing. If x_{nk} was missing, x_{nk} assumes an arbitrary constant that does not influence the statistical inferences. The observations and missing data indicators are collected in $N \times K$ matrices $\mathbf{X}$ and $\mathbf{D}$, respectively.

Rubin's ignorability principle entails that ignoring the missing data mechanism in frequentist inferences based on the likelihood function and Bayesian inferences based on the posterior distribution does not create bias if two conditions are met. The first condition entails that the missing data should be missing at random (MAR). MAR holds if the distribution of $\mathbf{D}$ given the observed data $\mathbf{X}_{obs}$, the

Authors' Note: Please address correspondence to Cees Glas, Department of Research Methodology, Measurement and Data Analysis, Faculty of Behavioral Sciences, University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands; e-mail: c.a.w.glas@gw.utwente.nl.

missing data $\mathbf{X}_{mis}$, the parameters of the distribution, say, ζ, and observed covariates $\mathbf{Y}$ do not depend on the missing data $\mathbf{X}_{mis}$, that is, if $p(\mathbf{D} \mid \mathbf{X}_{obs}, \mathbf{X}_{mis}, \zeta, \mathbf{Y}) = p(\mathbf{D} \mid \mathbf{X}_{obs}, \zeta, \mathbf{Y})$. Another way of expressing this is by the condition that all variables on which the distribution of $\mathbf{D}$ depends are observed. The second condition entails that the parameters of interest, say, θ, should be distinct from the ζ parameters. In a frequentist framework, distinctness means that there are no functional dependencies in the parameter space of ζ and θ; in a Bayesian framework, it means that the prior distributions of ζ and θ are independent. If the MAR and distinctness conditions are satisfied, the missing data process is ignorable in statistical inferences, which means that ignoring the distribution of $\mathbf{D}$ does not produce bias in the parameter estimates (Rubin, 1976).

When the missing data are nonignorable, the likelihood function and likelihood ratios that ignore the missing data process give rise to biased parameter estimates. An appropriate method to deal with these problems is to model the process that caused the missing data (Heckman, 1979). The idea is to identify and model the explanatory variables of the missing data process and to make inferences concurrently on this missing data model and the relevant model for the observed data. Several authors show that this approach can reduce bias caused by ignoring nonignorable missing data (Holman & Glas, 2005; Moustaki & Knott, 2000; Moustaki & O'Muircheartaigh, 2000; O'Muircheartaigh & Moustaki, 1999).

The present article focuses on item nonresponse in psychological and educational tests where responses are missing consecutively on items at the end of the test, for instance, because the respondent has not reached the end of the test due to a time limit. It must be expected that the number of items endorsed is correlated with the respondent's proficiency level; therefore, the missingness is nonignorable. This form of missingness is closely related to missingness caused by skipping of items by respondents with a low proficiency. Holman and Glas (2005) show that ignoring this missing data process can lead to bias in the estimates of the item parameters. Bradlow and Thomas (1998) also showed that ignoring this type of missing data process can produce bias in the parameter estimates.

This article shows that the missing data indicator of a test with a time limit can be modeled with the sequential model by Tutz (1990), a model also known as the "steps model" (Verhelst, Glas, & De Vries, 1997). The observed responses x_{nk} are discrete and will be modeled by the two-parameter logistic model (2PL; Lord, 1980) and the generalized partial credit model (GPCM; Muraki, 1992).

This article is made up of six sections and is organized as follows. After this section, a general notation is presented for item response theory (IRT) models for the missing data process, and the model for observed data will be discussed. Then a presentation about the estimation procedure using the maximum marginal likelihood (MML) method follows. In the next section, the bias introduced by ignoring the missing data process will be investigated in a number of simulation studies. An application of the method to an intelligence test will be shown in the fifth section.

Table 1
Example of Coding the Data

	Observed Data **X**						Missing Indicator **D**					
Respondent	1	2	3	k		K	1	2	3	k		K
1	0	1	9	9	9	9	1	1	0	9	9	9
2	1	1	9	9	9	9	1	1	0	9	9	9
3	1	0	1	9	9	9	1	1	1	0	9	9
4	1	1	0	0	9	9	1	1	1	1	0	9
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
n	1	1	0	1	9	9	1	1	1	1	0	9
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	0	0	0	1	0	9	1	1	1	1	1	0
.	1	0	1	0	1	9	1	1	1	1	1	0
N	1	1	0	1	0	0	1	1	1	1	1	1

Finally, the last section gives some conclusions and recommendations for further research.

A Multidimensional IRT Model

To model responses that are missing as a result of the speededness of a test, the study focused on the unobserved responses at the end of the response pattern. First, intermediate missing responses were considered ignorable missing data. However, some alternatives are discussed below. The missing data indicator d_{nk} was defined as

$$d_{nk} = \begin{cases} 1 & \text{if } x_{nk} \text{ was observed;} \\ 0 & \text{if } x_{nk} \text{ was the first response of a sequence } x_{nh} = 9, h = k, \ldots, K; \\ 9 & \text{otherwise .} \end{cases} \tag{1}$$

An example of the data considered here is shown in Table 1. Note that the responses in the observed data matrix **X** were dichotomously scored and the missing data were coded as *9*s. Note further that all the missing responses appeared at the end of the response patterns.

The data were modeled by an IRT model known as the sequential model (Tutz, 1990, 1997) or the steps model (Verhelst et al., 1997). It was assumed that the respondent kept giving responses (so $d_{nk} = 1$) until the first missing response appeared and then stopped responding. The distribution of d_{nk} depends on a latent person parameter θ_{n0}, and the probability of $d_{nk} = 1$ is given by $p_k(\theta_{n0})$. This entails that if $0 < k < K$, then

$$p(d_{n1} = d_{n2} = \ldots = d_{nk-1} = 1, d_{nk} = 0) = \left[\prod_{h=1}^{k} p_h(\theta_{n0})\right][1 - p_{k+1}(\theta_{n0})]; \tag{2}$$

if $k = K$, then

$$p(d_{n1} = d_{n2} = \ldots = d_{n(K-1)} = d_{nK} = 1) = \prod_{k=1}^{K} p_k(\theta_{n0});$$

and

$$p(d_{n1} = 0) = [1 - p_1(\theta_{n0})]$$

if the respondent fails to respond on the first item right away. It is assumed that

$$p_k(\theta_{n0}) = \frac{\exp(\alpha_{k0}\theta_{n0} - \beta_{k0})}{1 + \exp(\alpha_{k0}\theta_{n0} - \beta_{k0})}. \tag{3}$$

So $p_k(\theta_{n0})$ is the 2PL model for dichotomously scored items where α_{k0} is a so-called discrimination parameter and β_{k0} is a so-called item difficulty parameter. The model entails that the respondent makes so-called item steps until the first wrong response and then stops taking more item steps. Usually, the data **D** have too little variation to estimate the slope parameter α_{k0}, so in most cases it is convenient to assume that $\alpha_{k0} = 1$. In that case, the model given by Equation 3 specializes to the 1PL model (Rasch, 1960).

Furthermore, a restriction was imposed on the difficulty of the item steps given by

$$\beta_{k0} = \tau_0 + (k - K)\tau_1, \tag{4}$$

where τ_1 models a uniform change in the probability of an observation as a function of the position of the item in the test. The reason for this restriction on β_{k0} is that the first item steps are usually taken by all respondents, so the difficulty of these steps cannot be estimated separately. In addition, the restriction supports a monotonously decreasing probability of observing a response. Another interpretation of Equation 4 is that τ_0 is the difficulty of the last item step, and $(k - K)\tau_1$ is some measure of available time that decreases as the respondent proceeds through the test.

Combined IRT Models for Responses and the Missing Data Indicator

The combination of the model for the responses **X** and the missing data indicators **D** proceeds analogously to the approach by Holman and Glas (2005) adopted for modeling skipped items. They considered two classes of models: the MAR and NONMAR models.

Let $p(\mathbf{x}_n|\mathbf{d}_n, \theta_{n1}, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_1)$ be some model for the observed response pattern $\mathbf{x}_n = (x_{n1}, \ldots, x_{nk}, \ldots, x_{nK})$, where θ_{n1} is a latent proficiency parameter, and $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ are item parameters. Furthermore, let $p(\mathbf{d}_n|\theta_{n0}, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ be the model for the missing data indicator as defined above, where θ_{n0} is a latent person parameter, and $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ are parameters associated with items. Finally, $g_0(\theta_{n0})$ and $g_1(\theta_{n1})$ are the prior densities of the latent person parameters. The sequel assumed these densities were normal. Then the posterior of the person parameters of respondent n is proportional to

$$p(\mathbf{x}_n|\mathbf{d}_n, \theta_{n1}, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_1)p(\mathbf{d}_n|\theta_{n0}, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)g_0(\theta_{n0})g_1(\theta_{n1}) \; . \tag{5}$$

In Equation 5, the latent variables θ_{n1} for the observed data and θ_{n0} for the missing data process are independent, so the posterior distribution factors into two independent components: one for $\mathbf{x}_n$ and one for $\mathbf{d}_n$. Hence, the model for the likelihood of the missing data $p(\mathbf{d}_n|\theta_{n0}, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)g_0(\theta_{n0})$ can be ignored and estimates can be obtained using only

$$p(\mathbf{x}_n|\mathbf{d}_n, \theta_{n1}, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_1)g_1(\theta_{n1}) \; . \tag{6}$$

The model given by Equation 6 will be called the MAR model.

A violation of ignorability is created if it is assumed that the latent variables for the observed data and the missing data indicators, θ_{n1} and θ_{n0}, are dependent, hence, the model labeled NONMAR. In the sequel, it was assumed that θ_{n1} and θ_{n0} had a multivariate normal distribution with a covariance matrix $\boldsymbol{\Sigma}$. To identify the latent scale, the mean of this distribution was set equal to zero. The posterior was then given by

$$p(\mathbf{x}_n|\mathbf{d}_n, \theta_{n1}, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_1)p(\mathbf{d}_n|\theta_{n0}, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)g(\theta_{n0}, \theta_{n1}|\boldsymbol{\Sigma}), \tag{7}$$

where $g(\theta_{n0}, \theta_{n1}|\boldsymbol{\Sigma})$ is the prior distribution of θ_{n0} and θ_{n1}. If the off-diagonal elements of $\boldsymbol{\Sigma}$ are nonzero, the complete model for $\mathbf{x}_n$ and $\mathbf{d}_n$ has to be considered to obtain unbiased estimates of the parameters.

The GPCM

Depending on the response format, the observed responses can be modeled by various IRT models, such as the 2PL and 3PL models for dichotomous items and the graded response model (Samejima, 1969) and the GPCM (Muraki, 1992) for polytomous responses. Below, the GPCM, with the 2PL model as a special case for dichotomous responses, is used as an example. In the GPCM, the probability of responding in a category $g (g = 0, \ldots, m_k)$ is given by

$$p_{kg}(\theta_n) = p(X_{nkg} = 1|\theta_n, \alpha_k, \beta_k) = \frac{\exp(g\alpha_k\theta_n - \beta_{kg})}{1 + \sum\limits_{h=1}^{m_k} \exp(h\alpha_k\theta_n - \beta_{kh})}, \tag{8}$$

where α_k is a discrimination parameter and $\beta_{kg}(h=1,\ldots,m_k)$ are scalar location parameters. The item parameters are usually reparameterized as $\beta_{kg}=\sum_{h=1}^{g}\xi_{kh}$. The parameters $\xi_{kh}(h=1,\ldots,m_k)$ can be interpreted as so-called boundary parameters: ξ_{kh} is the position on the latent θ scale where $p_{k(h-1)}(\theta)=p_{kh}(\theta)$. The model becomes the 2PL model if $m_k=1$.

MML Estimation

MML estimation derives its name from maximizing a log likelihood that is marginalized with respect to the person parameters, rather than maximizing the joint log likelihood of all parameters simultaneously. If the likelihood given by Equation 7 is integrated over the person parameters, the marginal log likelihood is obtained:

$$\log L=\log\sum_{n}^{N}\iint p(\mathbf{x}_n|\mathbf{d}_n,\theta_{n1},\boldsymbol{\alpha}_1,\boldsymbol{\beta}_1)p(\mathbf{d}_n|\theta_{n0},\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)g(\theta_{n0},\theta_{n1}|\boldsymbol{\Sigma})\,d\theta_{n1}d\theta_{n0}. \quad (9)$$

In IRT models, the marginal likelihood rather than the joint likelihood is marginalized because Neyman and Scott (1948) showed that if the number of person parameters grows proportional with the number of observations, this can lead to inconsistent parameter estimates. Simulation studies by Wright and Panchapakesan (1969) showed that these inconsistencies can indeed occur in IRT models. Kiefer and Wolfowitz (1956) demonstrated that MML estimates of structural parameters, say, the item and population parameters of an IRT model, are consistent under fairly reasonable regularity conditions, which motivates the general use of MML for the estimation of IRT models, including multidimensional IRT models (Bock, Gibbons, & Muraki, 1988). The MML estimation equations for the present model can be found in Glas (2008), together with an expression for the standard errors of the parameter estimates.

Simulation Studies

The effect on the bias of the item parameter estimates of using either the MAR model given by Equation 6 or the NONMAR model given by Equation 7 was investigated using simulation studies. Three sets of simulation studies were conducted: one where the observed data followed the 1PL model, one where the observed data followed the 2PL model, and one where the observed data followed the GPCM. The missing data indicator always followed the steps model given by Equation 2. The first set of simulations was used to show how the bias in the item parameters developed as a function of their position in the test. The second simulation assessed the magnitude of the bias as a function of test length, sample size, and the correlation ρ between the two latent dimensions θ_0 and θ_1. The third set was done to assess whether the results obtained in the first two studies using dichotomously scored items generalized to the case of polytomously scored items.

Method

The test length was varied as $K = 10$ and $K = 40$, and the number of respondents was varied as $N = 500$ and $N = 1{,}000$. The person parameters θ_{n0} and θ_{n1} were drawn from a bivariate normal distribution with a mean equal to 0, a covariance matrix with diagonal elements equal to 1, and an off-diagonal element ρ that was varied as $\rho = 0$, $\rho = .2$, $\rho = .4$, $\rho = .6$, and $\rho = .8$. Note that $\rho = 0$ is the condition where ignorability holds, and the violation of ignorability increases with the magnitude of ρ. One hundred replications were made for every condition.

In the first set of simulations, the item parameters pertaining to the observed responses were all equal to $\alpha_k = 1.0$ and $\beta_k = 0$. This choice was made to evaluate the direction of the bias of the estimates with respect to some straightforward baseline values. In the second set of simulations, for every replication the discrimination parameters α_k were drawn from a distribution $.5 + Z$, where Z was drawn from a uniform distribution $U(0, 1.0)$. In addition, for every replication the item difficulties β_k were drawn from a standard normal distribution. The third set of simulations pertained to items with four response categories; that is, $m_k = 3$. In this case the discrimination parameters were again drawn as outlined above, but the category bounds parameters of the items were fixed. This was done to guarantee that all response categories attracted enough observations. For all items the values were equal to $\xi_{k1} = -1.0$, $\xi_{k2} = 0$, and $\xi_{k3} = 1.0$. Finally, for all three sets of simulations, the item parameters of the model for the missing data indicator were chosen in such a way that the amount of missing data was varied as 25% and 50%.

Results

The first set of simulations was used to show how the bias in the item parameters developed as a function of their position in the test. To get an impression of the trend of the bias and standard errors within a test, the item parameters were kept constant at $\alpha = 1$ and $\beta = 0$. The example shown here pertains to a test length of 10 items and a sample size of 1,000 respondents. The correlation between θ_{n1} and θ_{n0} was equal to $\rho = .8$. After some trial and error, the parameters of the missing data process were set equal to $\tau_0 = 1.00$ and $\tau_1 = .50$ to produce 50% missing data. One hundred replications were made. For every replication, a data set was generated using the steps model to select the items responded to and the 1PL model to generate the observations. For the MAR model the item parameters of the 2PL model and the variance of θ were estimated using MML. For the NONMAR model the

Table 2
Bias and Standard Error of Parameter Estimates Under the Not-Missing-at-Random (NONMAR) and the MAR Models, 50% Missing Data

	NONMAR Model				MAR Model			
Item	Bias(α)	Bias(β)	$SE(\alpha)$	$SE(\beta)$	Bias(α)	Bias(β)	$SE(\alpha)$	$SE(\beta)$
1	.000	.005	.000	.086	.000	.040	.000	.082
2	.033	.001	.158	.080	.032	.077	.193	.072
3	.022	.004	.154	.083	.004	.121	.160	.079
4	.026	.011	.157	.091	.016	.169	.179	.083
5	.042	.006	.193	.117	.021	.238	.197	.104
6	.054	.010	.196	.132	.038	.308	.197	.112
7	.088	.030	.300	.194	.022	.373	.287	.139
8	.068	.034	.347	.263	.023	.440	.327	.174
9	.054	.025	.423	.368	.056	.535	.391	.219
10	.034	.030	.857	.794	.121	.604	.800	.413

Note: $\tau_0 = 1.00$, $\tau_1 = .50$, $\rho = .8$.

item parameters of the two-dimensional 2PL model and the covariance matrix $\boldsymbol{\Sigma}$ were estimated using MML. The results are shown in Table 2. In the table, the items are numbered as they appeared in the test. The entries of Table 2 give the bias and the standard errors of the estimates of the item parameters of the model for the observed data, estimated by averaging over the 100 replications.

Notice that for both models the standard errors increased with the item number. This is because the number of observations on the items decreased with their position in the test: Everybody responded to the first item, but only the most able respondents reach the last item. Notice further that the bias in the estimates did not increase for the NONMAR model, which takes the missing data process into account. For the MAR model, which builds on the violated assumption of ignorability, the absolute bias increased with the item position. This especially held for the estimate of item difficulty; the estimates of the discrimination parameters were far less affected.

The second set of simulations was made to assess the magnitude of the bias as a function of test length, sample size, and the correlation ρ between the two latent dimensions θ_0 and θ_1. In practice, the item parameters are constant over the test but show variability. Therefore, in the second set of simulations, for every replication the discrimination parameters α_k were drawn from a distribution $.5 + Z$, where Z was drawn from a uniform distribution $U(0, 1.0)$, and the item difficulties β_k were drawn from a standard normal distribution. For every item, the bias and standard error were estimated using the estimates obtained in the 100 replications and then averaged over the items in the last half of the test. This was done because the effects in the first half of the test, where most respondents still responded to most items, were very small. The MML estimates for both the MAR and NONMAR models were computed using the steps model for the missing data indicators and the 2PL model for the observed responses.

As in the previous simulations, for the NONMAR model the bias in the estimates of the α and β parameters did not depend on the size of ρ, that is, on the correlation between θ_0 and θ_1. The results for the MAR model with 50% missing data are shown in Table 3. For $\rho = 0$, the bias was of the same magnitude as the bias under the NONMAR model. However, the absolute bias in the β parameters increased with the magnitude of ρ. So the larger the violation of ignorability, the larger the absolute bias. Note further that the bias was negative. Because the bias was computed as the difference between the estimate averaged over replications and the true parameter, the conclusion is that the item difficulty was underestimated. The explanation is that these items were endorsed by relatively high-proficiency respondents reaching the end of the test, and these high-proficiency students produced a higher proportion of correct responses than the proportion of correct responses that would be obtained if the complete sample of respondents would have endorsed these items. However, the MAR model has the implicit assumption that all items were endorsed by the same sample of respondents from the same proficiency distribution, so the items at the end of the test appear easy. Essentially, the argument is that only high-ability students reach the end of the test, so these items appear easier.

Note that there was no clear effect of ρ on the bias when using the MAR model in the α parameters. In the MAR model, the discrimination indices were slightly underestimated. Although the latter effect was small, it also has a straightforward explanation. The variance of the proficiency parameters of the respondents present in the first part of the test was larger that the variance of the respondents still present in the second part. However, the MAR model could not accommodate to this difference in variability through the model for the proficiency distribution and accommodated to the lesser response variability in the second part by lowering the estimates of the discrimination parameters. Or, to put it in another way, one variance of ability was assumed in the MAR model, but in reality this variance was smaller at the end of the test through the selection process: Only the bright students got that far. But less variance means less discrimination, so the discrimination parameters must go down.

Interestingly, for both the NONMAR and MAR models, the standard errors correlated positively with ρ. At first, this may seem a little surprising because the proportion of missing responses did not depend on ρ, so the numbers of observations determining the standard errors did not change. The explanation is in the loss of information due to the selection of respondents that occurred if ρ was increased. If ρ was high, only the more proficient respondents reached the second part of the test. However, all other things being equal, the amount of information in the responses

Table 3
Bias and Standard Error of Parameter Estimates Averaged Over the Last Half of the Items Under the Missing-at-Random Model, Two-Parameter Logistic Model, Case of 50% Missing Data

		$K = 10$				$K = 40$			
N	ρ	Bias(α)	Bias(β)	*SE*(α)	*SE*(β)	Bias(α)	Bias(β)	*SE*(α)	*SE*(β)
500	.0	.166	−.019	.441	.386	.038	−.027	.417	.355
	.2	.157	−.174	.458	.288	−.088	.011	.304	.302
	.4	.172	−.259	.586	.438	−.063	−.141	.370	.323
	.6	.054	−.335	.583	.282	−.049	−.185	.509	.400
	.8	.063	−.574	.474	.328	−.068	−.289	.404	.310
1,000	.0	.048	−.003	.393	.198	−.016	−.020	.255	.227
	.2	.031	−.081	.306	.188	−.170	−.061	.247	.204
	.4	.024	−.237	.336	.206	−.096	−.127	.239	.203
	.6	−.013	−.368	.317	.234	−.082	−.194	.277	.224
	.8	−.037	−.484	.501	.362	−.117	−.269	.326	.276

Note: For $K = 10$: $\tau_0 = 1.00$, $\tau_1 = .50$; for $K = 40$: $\tau_0 = -.20$, $\tau_1 = .12$.

depends on the distance between the proficiency parameters and the item difficulties (Lord, 1980). Nevertheless, the expectation of the item difficulties was zero and proficient respondents were farther removed from zero than the average respondent. As a result, the information decreased, and the standard errors, which are the reciprocal of information, increased.

Finally, from inspection of Table 3 it can be seen that there were main effects on the bias and standard errors of test length, sample size, and the proportion of missing data. All these effects were in the expected direction.

The simulations were also done with 25% missing data. For a test length of $K = 10$, this was achieved using parameter values $\tau_0 = 1.00$ and $\tau_1 = 1.00$, and for a test length $K = 40$ this was achieved using parameter values $\tau_0 = .75$, and $\tau_1 = .30$. The overall trends in the results were similar to the overall trends in the case of 50% missing data, but the magnitudes of the effects were lower. For the case of a test length of $K = 10$, the biases in the item difficulty parameters for a correlation of .8 rose to $-.223$ and $-.211$ for sample sizes of $N = 500$ and $N = 1{,}000$, respectively. For 40 items, these biases were $-.072$ and $-.067$, respectively. So the latter two biases were quite small.

The last set of simulations pertained to polytomously scored items with four response categories. The discrimination parameters were again drawn from a distribution $.5 + Z$, where Z had uniform distribution on (0, 1.0) and the category bounds parameters were equal to $\xi_{k1} = -1.0$, $\xi_{k2} = 0$, and $\xi_{k3} = 1.0$. Other details of the simulations were analogous to the two previously reported studies. Results are displayed in Table 4. Note that for all category bounds parameters, the bias increased

Table 4
Bias and Standard Error of Parameter Estimates Averaged Over the Last Half of the Items Under the Missing-at-Random Model, Generalized Partial Credit Model, Case of 25% Missing Data

		$K = 10$				$K = 40$			
N	ρ	Bias(α)	Bias(β)	SE(α)	SE(β)	Bias(α)	Bias(β)	SE(α)	SE(β)
500	.2	−.187	−.300	.319	.326	−.163	−.325	.398	.490
	.4	.180	−.281	.310	.351	−.164	−.341	.405	.492
	.6	.182	−.296	.328	.360	−.164	−.343	.399	.486
	.8	.186	−.314	.315	.363	−.164	−.354	.407	.533
1,000	.2	−.114	−.193	.217	.236	−.100	−.221	.266	.324
	.4	−.121	−.198	.213	.228	−.099	−.237	.279	.345
	.6	−.116	−.200	.216	.235	−.103	−.246	.278	.347
	.8	−.120	−.226	.241	.260	−.106	−.266	.300	.386

Note: For $K = 10$: $\tau_0 = 1.00$, $\tau_1 = 1.00$; for $K = 40$: $\tau_0 = .75$, $\tau_1 = .30$.

in absolute magnitude with ρ. As above, both the discrimination and category bounds parameters had a negative bias. The explanation remains the same: In the MAR model it was assumed that all items were endorsed by the same sample from the proficiency distribution; therefore, the items at the end of the test appeared easier and appeared to discriminate less. The bias virtually vanished under the NOMAR model.

An Application to the NIO Intelligence Test

In the framework of classical test theory, Attali (2005) showed that the reliability of number-correct scores on speeded multiple choice tests is decreased as a result of the lower consistency of the last responses of the test with the responses of the other items. In the application presented here, the MAR and NONMAR models were used to estimate the reliability of scores on a test administered with a time limit. The example concerned an analysis of data from a calibration sample of the Nederlandse Intelligentietest voor Onderwijsniveau (NIO), which is an intelligence test for students in primary education in the Netherlands (van Dijk & Tellegen, 2004). The test was developed to support the streaming of students into the various tracks in secondary education. The test consists of a number of subscales. In the present study, five subscales were analyzed: the Synonymies, Analogies, Categories, Numbers, and Arithmetic scales. The numbers of items and the numbers of

Table 5
Correlation Between Latent Variables and Reliability of Subtests Estimated Using the Missing-at-Random (MAR) Model and the NONMAR Model

Subtest	Number of Items	*N* First Item	*N* Last Item	ρ (θ_0,θ_1)	MAR Model	NONMAR Model	Optimal Weighted
Synonymies	30	3,145	508	.429	.727	.737	.743
Analogies	25	3,141	1,210	.097	.643	.699	.699
Categories	30	3,145	1,176	.348	.793	.803	.804
Numbers	25	3,145	837	.170	.757	.759	.759
Arithmetic	20	3,145	1,286	.035	.800	.803	.803

students are given in the second and third column of Table 5, respectively. The scales were administered under a time limit. Therefore, the percentage of missing data was equal to 27%. The fourth column of Table 5 gives the number of students who endorsed the last item of each subscale.

For every subscale, MML estimates of the parameters of the MAR and NONMAR models were obtained. The estimated correlations between the latent parameters of the observed data and missing data are given in the fifth column of Table 5. All correlations differed significantly from zero, which indicates that the missing data process could not be ignored. However, especially for the second and fifth subscale, the correlations were quite small.

The impact of the difference between the MAR and NONMAR models was studied further by computing a global reliability coefficient. Usually, this is done using classical test theory. However, if missing data are present, it is more convenient to compute a global reliability coefficient via IRT. In an IRT framework, such a coefficient can be based on the identity

$$\mathrm{var}(\theta) = E(\mathrm{var}(\theta|\mathbf{x})) + \mathrm{var}(E(\theta|\mathbf{x}))$$

(Scheerens, Glas, & Thomas, 2003). This identity entails that the total variance of the proficiency parameters is a sum of two components. The first component, $E(\mathrm{var}(\theta|\mathbf{x}))$, relates to the uncertainty about the proficiency parameter. The posterior variance of proficiency, $\mathrm{var}(\theta|\mathbf{x})$, gives an indication of the uncertainty with respect to the proficiency parameter, once the response pattern $\mathbf{x}$ is observed. By considering its expectation over the distribution of $\mathbf{x}$, an estimate of the average uncertainty is obtained for the respondents' proficiency parameters. The second term, $\mathrm{var}(E(\theta|\mathbf{x}))$, is related to the systematic measurement component. The expectation serves as an estimate of proficiency, and the variance of these expectations over the distribution of $\mathbf{x}$ indicates the extent to which the respondents can be distinguished on the basis of their observed responses. Therefore, the reliability coefficient can be computed as the ratio of the systematic variance and the total variance; that is,

$$\rho = \frac{\text{var}(E(\theta|\mathbf{x}))}{\text{var}(\theta)} \,. \tag{10}$$

In the present application, there are several ways to score the test and to compute a global reliability coefficient. One way is to focus on the proficiency parameter θ_1 related to the observed responses. Then var($E(\theta|\mathbf{x})$) can be computed in two ways: under the MAR assumption conditioned only on the observed responses, and under the NONMAR assumption conditioned on both the observed responses **x** and the missing data indicator **d**. The latter case was integrated over both latent variables involved, so the expectation in the numerator of Equation 10 is computed as

$$\begin{aligned} E(\theta_1|\mathbf{x},\mathbf{d}) &= \int\int \theta_1\, p(\theta_0,\theta_1|\mathbf{x},\mathbf{d})\, d\theta_1 d\theta_0 \\ &= \frac{\int\int \theta_1\, p(\mathbf{x}|\mathbf{d},\theta_1)p(\mathbf{d}|\theta_0)g(\theta_0,\theta_1|\mathbf{\Sigma})\, d\theta_1 d\theta_0}{p(\mathbf{x}|\mathbf{d})} \,. \end{aligned}$$

The results are given in the last three columns of Table 5. The columns labeled “MAR Model” and “NONMAR Model” give the global reliability computed under the two models. Note that the reliability under the NONMAR model is always higher.

Both indices are estimates of the global reliability of the estimates of the proficiency parameter θ_1. However, θ_0 is correlated with θ_1, and θ_0 also contains information with respect to the proficiency level of the students. Therefore, one can construct a linear compound $\theta_1 + \beta\theta_0$ that maximizes the reliability given by

$$\rho = \frac{\text{var}(E(\theta_1 + \beta\theta_0|\mathbf{x}))}{\text{var}(\theta_1 + \beta\theta_0)} \,.$$

These reliability estimates are given in the last column of Table 5. Note that the reliabilities under the NONMAR model and the reliabilities of the linear compound are clearly higher only for the first two subscales.

Discussion

Most approaches to modeling the responses to speeded tests focus on a situation where the responses to the last items in the test are actually recorded. In these approaches, a latent class model is used to distinguish between the responses produced according to an IRT model and responses that are guessed (Bolt, Cohen, & Wollack, 2002; Mislevy & Verhelst, 1990; Yamamoto & Everson, 1996). This study considered a different situation, where the omitted responses were actually recorded as missing. It was shown that these missing data may cause a violation of the ignorability principle by Rubin (1976), which may lead to biased parameter estimates. The simulation studies showed that this bias actually occurred. Furthermore, it was shown that modeling the process that caused the missing data can significantly reduce this bias. The choice of the missing data model must reflect the sampling process involved. Holman and Glas (2005), for instance, considered the problem of skipping difficult items by low-ability respondents. They assumed that the Rasch model holds for the missing data indicator. So implicitly they assumed local independence between the realizations of the missing data indicator; therefore, the missing data could, in principle, emerge everywhere.

In the present case, the last string of items in the test was not reached, and the realizations of the missing data indicators at the end of the test could not be viewed as independent observations. Therefore, the pattern of missingness had to be modeled as a whole, and this was done using the sequential model with linear restrictions described above. To keep the presentation relatively simple, it was assumed that the only missing data were the not-reached items at the end of the test. Intermediate missing responses were regarded as ignorable. However, this may not always be realistic. In these cases one may use the approach of Holman and Glas (2005) and apply the Rasch model as a model for the missing data indicator for intermediate missing responses. One can then either assume that these responses depend on a third latent proficiency variable, say, θ_3, that is correlated with the two other variables, or one may assume that these intermediate missing responses depend on the same latent proficiency variable that also describes the not-reached items through the steps model.

The final remark pertains to model fit. In itself, the method of adding a model with latent variables to describe the missing data process and checking whether the parameter estimates issued from the MAR and NONMAR models are comparable can be viewed as a test of model fit, that is, as a test of the assumption of ignorability. However, rejection of the MAR model does not imply that the NONMAR model holds. The NONMAR model can be viewed as a special case of a multidimensional GPCM model, and the methods for testing that model apply directly to the NONMAR model. One can, for instance, use the Lagrange multiplier tests for testing the GPCM model in an MML framework, which are discussed by te Marvelde, Glas, van Landeghem, and van Damme (2006). Especially, versions of these tests that focus on the assumption of local stochastic independence (e.g., see Glas, 1999) may be applied to evaluate the dependence between x_{nk} and d_{nk} for fixed k. The optimal choice of test statistics for the present specific application, however, is a topic of further study and, as such, beyond the scope of this article.

References

Attali, Y. (2005). Reliability of speeded number-right multiple-choice tests. *Applied Psychological Measurement*, *29*, 357-368.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement*, *12*, 261-280.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331-348.

Bradlow, E. T., & Thomas, N. (1998). Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics*, *23*, 236-243.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, *64*, 273-294.

Glas, C. A. W. (2008). *A marginal maximum likelihood procedure for an IRT model for speeded tests* (OMD Progress Report, 08-01). Enschede, Netherlands: Twente University. Available from http://www.utwente.nl

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrika*, *46*, 931-961.

Heitjan, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika*, *81*, 701-708.

Holman, R., & Glas, C. A. W. (2005). Modeling non-ignorable missing data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*, 1-18.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, *27*, 887-903.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J., & Verhelst, N. D. (1990). Modeling item responses with different subjects employing different solution strategies. *Psychometrika*, *55*, 195-216.

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A*, *163*, 445-459.

Moustaki, I., & O'Muircheartaigh, C. (2000). A one dimensional latent trait model to infer attitude from nonresponse for nominal data. *Statistica*, *10*, 259-276.

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

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, *16*, 1-32.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, Series A*, *162*, 177-194.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

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

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika Monograph Supplement*, *17*.

Scheerens, J., Glas, C. A. W., & Thomas, S. M. (2003). *Educational evaluation, assessment, and monitoring: A systemic approach*. Lisse, Netherlands: Swets & Zeitlinger.

te Marvelde, J. M., Glas, C. A. W., Van Landeghem, G., & Van Damme, J. (2006). Application of multidimensional IRT models to longitudinal data. *Educational and Psychological Measurement*, *66*, 5-34.

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

Tutz, G. (1997). Sequential models for ordered responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 139-152). New York: Springer.

van Dijk, H., & Tellegen, P. (2004). *NIO, Nederlandse Intelligentietest voor Onderwijsniveau.* [NIO, Dutch intelligence test for educational level]. Amsterdam: Boom Test Uitgevers.

Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp.123-138). New York: Springer.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*, 23-48.

Yamamoto, K., & Everson, H. (1996). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89-98). New York: Waxmann.