

# Simple imputation methods versus direct likelihood analysis for missing item scores in multilevel educational data

Damazo T. Kadengye · Wilfried Cools ·  
Eva Ceulemans · Wim Van den Noortgate

Published online: 15 October 2011  
© Psychonomic Society, Inc. 2011

**Abstract** Missing data, such as item responses in multilevel data, are ubiquitous in educational research settings. Researchers in the item response theory (IRT) context have shown that ignoring such missing data can create problems in the estimation of the IRT model parameters. Consequently, several imputation methods for dealing with missing item data have been proposed and shown to be effective when applied with traditional IRT models. Additionally, a nonimputation direct likelihood analysis has been shown to be an effective tool for handling missing observations in clustered data settings. This study investigates the performance of six simple imputation methods, which have been found to be useful in other IRT contexts, versus a direct likelihood analysis, in multilevel data from educational settings. Multilevel item response data were simulated on the basis of two empirical data sets, and some of the item scores were deleted, such that they were missing either completely at random or simply at random. An explanatory IRT model was used for modeling the complete, incomplete, and imputed data sets. We showed that direct likelihood analysis of the incomplete data sets

produced unbiased parameter estimates that were comparable to those from a complete data analysis. Multiple-imputation approaches of the two-way mean and corrected item mean substitution methods displayed varying degrees of effectiveness in imputing data that in turn could produce unbiased parameter estimates. The simple random imputation, adjusted random imputation, item means substitution, and regression imputation methods seemed to be less effective in imputing missing item scores in multilevel data settings.

**Keywords** Item response theory · Multilevel data · Missing data · Imputation methods

Multilevel data in education settings can contain complex patterns of nested sources of variability. For instance, suppose that exercise items nested in courses or chapters with varying difficulty levels are presented to students nested in schools or classes with varying ability levels. For each student, scores to the items, along with person and item properties, can be recorded. The collected data are clustered or multilevel in nature, consisting of the properties of schools, students, chapters, and items and the students' item scores (e.g., binary, pass/fail) on attempted items. These data can be modeled statistically—say, using item response theory (IRT; van der Linden & Hambleton, 1997)—to explain and understand student characteristics in relation to the item properties.

De Boeck and Wilson (2004) described IRT models within the framework of generalized linear mixed models (GLMMs) or nonlinear mixed models (NLMMs), also accounting for more complex multilevel structures than the structure of measurement occasions within subjects,

---

D. T. Kadengye (✉)  
Faculty of Psychology and Educational Sciences,  
Katholieke Universiteit Leuven,  
Etienne Sabbelaan 53,  
8500 Kortrijk, Belgium  
e-mail: Trevor.Kadengye@kuleuven-kortrijk.be

D. T. Kadengye · W. Cools · W. Van den Noortgate  
ITEC-IBBT-K.U.Leuven,  
8500 Kortrijk, Belgium

W. Cools · E. Ceulemans · W. Van den Noortgate  
Katholieke Universiteit Leuven,  
8500 Kortrijk, Belgium

which is usually accounted for in ordinary IRT models. When clusters are looked at as being randomly chosen from a population of clusters, the cluster effects can be treated as random effects in the GLMMs or NLMMs. In this way, the GLMMs or NLMMs not only account for within-person differences in the item response probabilities and between-person differences in the latent construct(s) (Briggs, 2008), but also for differences between groups of persons and groups of items. Moreover, the GLMM and NLMM frameworks suggest including predictors. This is appealing if one is interested in using student and item properties to explain group differences in item scores. In this regard, we follow the ideas of De Boeck and Wilson, who used the term *explanatory item response modeling* to refer to the use of IRT as a tool not only for measurement, but also for explanation.

### Explanatory item response modeling

Let  $Y_{pi}$  denote a binary score of person  $p$  ( $p = 1, \dots, P$ ) to item  $i$  ( $i = 1, \dots, I$ ). For a basic IRT model, commonly referred to as the *Rasch model*, the score of person  $p$  to item  $i$  is regarded as a function of person and item parameters,  $\theta_p$  and  $\beta_i$ , that can be interpreted as the person ability and item difficulty, respectively, such that

$$\begin{aligned} \text{logit}(\pi_{pi}|\theta_p) &= \ln(\pi_{pi}/1 - \pi_{pi}) \\ &= \theta_p - \beta_i \text{ and } Y_{pi} \sim \text{binomial}(1, \pi_{pi}), \end{aligned} \quad (1)$$

where  $\pi_{pi}$  is the probability of success for person  $p$  on item  $i$ .

In addition to Eq. 1, one can assume that persons are a random sample from a population in which people's abilities (the  $\theta_p$ s) are normally distributed, such that  $\theta_p \sim N(0, \sigma_\theta^2)$ . Whereas the Rasch model in Eq. 1 is a purely descriptive model, we can try to explain differences in item difficulty by using real item predictors, as in the linear logistic test model (LLTM; Fischer, 1973), instead of estimating the difficulty for each item separately. De Boeck and Wilson (2004) therefore called the LLTM an item explanatory IRT model. On top of the item predictors, random item effects can be included, where the *variance* of these random item effects refers to the variance in item difficulty that is not explained by the item predictors in the model.

In the same way, we can try to explain differences in person ability using a latent regression IRT model by including person covariates as fixed effects (Zwinderman, 1991) in addition to the random person effects, to get a person explanatory IRT model that does not assume that the person predictors explain all variance (De Boeck & Wilson, 2004). Kamata (2001) for example applied a latent regression model with person characteristic variables to analyze the effect of studying at home on science achieve-

ment. Thus, Eq. 1 can be further extended by including person and item predictors, to get a (doubly) explanatory IRT model (De Boeck & Wilson, 2004). Formulated as a GLMM, the probability of a correct response is given by

$$\text{logit}(\pi_{pi}) = \sum_{j=1}^J \varphi_j X_{pj} + \sum_{q=1}^Q \gamma_q G_{iq} + \omega_p + \epsilon_i, \quad (2)$$

where  $X_{pj}$  is the value of predictor  $j$  for person  $p$ ;  $G_{iq}$  is the value of predictor  $q$  for item  $i$ ;  $\varphi_j$  and  $\gamma_q$  are the unknown fixed regression coefficients for  $X_j$  and  $G_q$ ; and  $\omega_p$  and  $\epsilon_i$  are random effects for persons and items, which are assumed to have independent normal distributions with means of 0—that is,  $\omega_p \sim N(0, \sigma_\omega^2)$  and  $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ . This model includes several popular IRT models as special cases. For instance, when item predictors (the  $G_q$ s) are dummy indicators, such that  $G_{iq} = 1$  if  $i = q$  or 0 if  $i \neq q$ , and dropping the person predictors and the random item effects, Eq. 2 simplifies to Eq. 1, the basic Rasch model, with fixed item effects (where  $-\gamma_q$  can be interpreted as the difficulty of item  $i$ ) and random person effects (where  $\omega_p$  can be interpreted as the ability of person  $p$ ).

An advantage of the explanatory item response modeling framework is its flexibility. Depending on a researcher's interest, the model in Eq. 2 can be broken down or extended to different explanatory IRT models to explain latent properties of items, persons, groups of items, or groups of persons. For instance, when persons are grouped within schools, and if schools  $k = 1, 2, \dots, K$  are thought of as being a random sample from a population of schools, such that school effects ( $u_k$ s) are normally distributed with  $u_k \sim N(0, \sigma_u^2)$ , and letting  $\pi_{kpi}$  be the probability of success for person  $p$  from school  $k$  on item  $i$ , then Eq. 2 can be extended as

$$\text{logit}(\pi_{kpi}) = \sum_{j=1}^J \varphi_j X_{pj} + \sum_{q=1}^Q \gamma_q G_{iq} + \omega_p + u_k + \epsilon_i, \quad (3)$$

The explanatory IRT framework is also an important tool for the validation of research instruments. When an instrument or a test is interpreted as a measure of some postulated attribute of people that is not operationally defined but is assumed to be reflected in test performance, construct validation is said to be involved (Embretson, 1983). Modeling random item effects allows one to specify item features as predictors of item difficulty in order to assess internal evidence for validity or construct representation. In other words, the validity of the inferences made from a test is demonstrated through the underlying relations between item features and performance (Hoffman, Yang, Bovaird, & Embretson, 2006), which is not the case in the LLTM or other traditional IRT models that assume perfect prediction of item difficulty. Similarly, by including random

person effects, individual differences can be assessed in order to understand the degree to which expected relationships are observed with measures of theoretically related constructs—that is, assessing convergent validity or external evidence.

#### Missing item scores in multilevel data: An example from e-learning environments

In education settings, students can receive information to build and improve their knowledge in traditional classroom settings, via the Internet (and other electronic multimedia), or through a combination of both. A specific kind of Internet-based learning is an item-based e-learning environment in which persons are required to study and attempt exercises online. An example of such e-learning environments is the FraNel project (Desmet, Paulussen, & Wylin, 2006). In such environments, exercise items are structured within different groups such as chapters, and persons who are themselves structured in different social communities, such as schools, can learn by logging into the learning environment at any time to freely engage exercises.

The use of IRT for analyzing the tracking and logging data resulting from such learning environments can result in interesting applications, such as rendering items with difficulties that are matched with the proficiency of the persons (Wauters, Desmet, & Van den Noortgate, 2010). Yet the analyses may be hampered by missing item scores that might occur due to several mechanisms and factors. For instance, persons may leave items blank that proved to be too difficult for them, they may lose interest midway through the test, they may skip certain sections inadvertently, or perhaps they may refuse to respond to sensitive topics. Large numbers of missing values may also arise where persons navigate freely through the items and only engage in a relatively small number of independent exercises out of a vast number that are normally provided within learning environments (Wauters et al. 2010). A high number of missing item scores poses difficulties for using IRT to estimate item difficulties and person abilities—for instance, nonconvergence problems in the estimation of the IRT model parameters. Moreover, during statistical analysis of data with missing scores, mechanisms that lead to missing items need to be identified in order to avoid biased estimates (Little & Rubin, 2002).

#### Missing-data mechanisms

Item scores can be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Data missing completely at random occur when the probability of an item having a missing score does not depend on any of the observed or unobserved quantities

(item responses or properties of persons or items). Once this assumption holds, the process(es) generating the missing values can be ignored during the analysis, and estimators from a complete case analysis can be unbiased (Horton & Kleinman, 2007; Molenberghs & Verbeke, 2005), but there could be substantial efficiency losses (Little & Rubin, 2002). However, MCAR is a strong and unrealistic assumption in most research experiments, because very often some degree of relationship exists between the missing values and some of the covariate data. Thus, it might be more reasonable to state that data are missing at random (MAR), which occurs when a missing item depends only on some observed quantities, which may include some outcomes and/or covariates. If neither MCAR nor MAR holds, then missing items depend on the unobserved quantities and data are said to be missing not at random (MNAR). For instance, missing item scores could be due to those items being too difficult for certain students.

These missing-data mechanisms need to be taken into account during statistical modeling to be able to produce estimates that are unbiased and efficient. However, as noted by Molenberghs and Verbeke (2004), no modeling approach, whether for MAR or for MNAR, can fully compensate for the loss of information that occurs due to incompleteness of the data. In likelihood-based estimation, a missing item arising from MCAR or MAR is ignorable, and parameters defining the measurement process are independent of the parameters defining the missing item process (Beunckens, Molenberghs, & Kenward, 2005). In the case of a nonignorable missing item arising from an MNAR mechanism, a model for the missing data must be jointly considered with the model for the observed data, which can, for example, be done using selection models and/or pattern mixture models (Molenberghs & Verbeke, 2005). However, the causes of an MNAR mechanism are difficult to ascertain a priori (or are unknown), and the methods to implement the remediation of such a mechanism are complex and beyond the scope of this study. Thus, MNAR is not considered further in the present study.

#### Dealing with missing data in IRT

Most methods for dealing with missing data were developed outside the IRT context, but some of them have been applied and examined in traditional item response models. For instance, Finch (2008) compared the performance of several missing-data techniques, and noted that several of these methods exhibited varying degrees of effectiveness in terms of imputing data for a simulated data set from a three-parameter logistic model. Whereas Finch investigated missing scores for the same 4 items out of a set of 20 items, in Web-learning environments missing scores can occur on any of the items. Sheng and Carrière (2005)

examined the implications of applying common missing-data strategies to Rasch models for item response data under various missing-data mechanisms, but they limited the investigation to data with continuous measurements, which may not always be the case with item-based data. In addition, Sheng and Carrière proposed a bootstrap technique under imputation for missing item response data and noted that for the case of 20% missing data, this method appeared to be the best strategy for efficiently producing consistent estimators and their variances. The bootstrap technique method, however, did not work well when a large proportion of the data were missing, even with a large number of items being explored. In Web-learning environments in which persons are free to navigate and engage in a large number of available exercises, large amounts of missing item scores (greater than 20%) are common. Huisman (2000) investigated the effects of several naïve deterministic imputation methods on the estimation of the latent ability of respondents and Cronbach's alpha (indicating the reliability of a test). Deterministic methods, however, assume that all conditions leading to missing item scores are perfectly known, an assertion that may not be entirely valid. Sijtsma and van der Ark (2003) discussed some simple methods and proposed two nonparametric single-imputation methods, one of which seemed to be superior in recovering several statistical properties of the original complete data from an incomplete data set. In addition, van Ginkel, van der Ark, and Sijtsma (2007) showed that multiple-imputation versions of some methods discussed by Sijtsma and van der Ark produced small discrepancies when compared to the statistical properties of the completely observed data. However, the designs of the data structures used in both studies to test these methods did not contain a hierarchical structure, and the proportions of missing item scores studied were small—between 1% and 15%.

Another limitation of most of these studies is that they do not examine *explanatory* item response models. Moreover, most studies do not account for more complex multilevel structures, with persons and/or items nested in groups. This study therefore extends the fast growing research domain on missing data in three ways: (1) We focused on explanatory models, while previous studies focused mainly on the descriptive Rasch model; (2) we investigated the performance of some simple-imputation approaches in the case of multilevel data structures, with persons grouped in, for instance, schools; and (3) we focused on substantial amounts of missing item scores, such as are common in tracking and logging e-learning data. In the remainder of this article, we discuss some of the common methods for handling missing item responses, and in a simulation study, apply and compare their performance on multilevel data with substantial numbers of item scores missing.

## Missing-data methods

Several methods for dealing with missing data in IRT settings exist and can readily be implemented in such statistical software as R, SAS, or SPSS. For the present study, the approaches we discuss are by no means an exhaustive set, but were selected because previous research had demonstrated their potential effectiveness in estimating missing item scores for item response data; as such, they were deemed worthy candidates for use with multilevel item response data. For instance, some of the methods we describe have been shown to be effective in imputing missing item scores for test and questionnaire data (van Ginkel et al. 2007), to result in limited bias for item parameter estimates (Finch, 2008), or to eliminate the imputation variance of the estimator (Chen, Rao, & Sitter, 2000). Some methods are described or mentioned but not implemented in this analysis, because they were deemed superfluous in the presence of other imputation methods. For a comprehensive overview of missing-data methods, the interested reader is referred to Schafer and Graham (2002).

To set notation first, suppose a person  $p \in P$  responds to  $I_{rp}$  items but misses  $I_{mp}$  items, with  $I_{rp} + I_{mp} = I$ . Similarly, suppose an item  $i \in I$  is responded to by  $P_{ri}$  persons but is missed by  $P_{mi}$  persons, with  $P_{ri} + P_{mi} = P$ . Let  $Y_p^o = Y_{p1}^o, Y_{p2}^o, \dots, Y_{pI_{rp}}^o$  for all  $i \in I_{rp}$  be a vector of the observed scores for person  $p$ , such that  $Y_p^m = Y_{p1}^m, Y_{p2}^m, \dots, Y_{pI_{mp}}^m$ , for all  $i \in I_{mp}$  is a vector of his/her missing scores. From the item side, let  $Y_i^o = Y_{1i}^o, Y_{2i}^o, \dots, Y_{P_{ri}i}^o$  for all  $p \in P_{ri}$  denote a vector of the observed person scores on item  $i$ , such that  $Y_i^m = Y_{1i}^m, Y_{2i}^m, \dots, Y_{P_{mi}i}^m$  for all  $p \in P_{mi}$  is a vector of item  $i$ 's missing person scores.

## Complete case analysis (CCA)

A “case” in this situation refers to a person. A complete case analysis ignores persons who did not answer all items, thereby retaining only part of the observed data for analysis (see, e.g., Horton & Kleinman, 2007; Little & Rubin, 2002; Molenberghs & Verbeke, 2005). That is to say, person  $p$  is included within the analysis if he/she provided a fully observed response vector  $Y_p^o = Y_{p1}^o, Y_{p2}^o, \dots, Y_{pI}^o$  for all  $I = 1, \dots, I$ . This approach is problematic in some item-based research settings. For instance, in e-learning environments, the large number of items coupled with navigation freedom at a person's pace can make it likely that almost every person will have missing item scores, resulting in substantial efficiency losses of the complete case estimator. Moreover, it is unlikely that the persons that answered all items can be regarded as a random sample. Therefore, CCA will not be considered further in this study.

### Simple random imputation (SRI)

Simple random or *hot deck* imputation (Little & Rubin, 2002), which is specially designed for categorical data (Rubin & Schenker, 1986), fills the missing components of a variable by drawing with replacement and with equal probabilities from the observed values of that particular variable. In its simplest form, SRI is fit for random imputation of a single variable (Little & Rubin, 2002). For item response data, SRI can be applied person by person or item by item, but for simplicity, we will consider only one case: item by item. In this case, SRI selects a simple random sample of size  $P_{mi}$  with replacement and equal probabilities from  $P_{ri}$ , and then uses the associated  $Y_i^0$  values as donors for  $Y_i^m$  for all  $p \in P_{mi}$ ; that is, a missing item score is filled with a value  $Y_{pi}^m = Y_{qi}^0$  for some  $q \in P_{ri}$ . As such, the distribution of the item values is preserved.

### Adjusted random imputation (ARI)

The adjusted random imputation (Chen et al. 2000) method is an adaptation of SRI that replaces missing item scores with random estimates in such a way that the average of each item remains the same as in the raw data table. The ARI method uses  $\tilde{Y}_{pi}^m = \bar{Y}_{\bullet i} + (\tilde{Y}_{pi}^{m*} - \tilde{Y}_{\bullet i}^*)$  for all  $p \in P_{mi}$  as imputed values for the missing scores instead of  $Y_{pi}^{m*}$ , where  $\bar{Y}_{\bullet i}$  is the mean of an item's observed scores and  $\tilde{Y}_{\bullet i}^*$  is the mean of random values  $Y_{pi}^{m*}$  for all of the  $p \in P_{mi}$  obtained from SRI. For binary values, the computed values  $\tilde{Y}_{\bullet i}^m$  must be rounded to 0 or 1 or used as parameters to make a random draw from the Bernoulli distribution.

### Mean imputation (IM and PM)

Two variants of mean imputation can be considered—item mean substitution (IM) and person mean substitution (PM; Huisman, 2000). Under IM, the arithmetic mean of an item's observed scores, computed as  $IM_i = \bar{Y}_{\bullet i} = \frac{1}{P_{ri}} \sum_{p=1}^{P_{ri}} Y_{pi}^o$ , is used to replace every missing score of the said item. For PM, the arithmetic mean of a person's observed scores, computed as  $PM_p = \frac{1}{I_{rp}} \sum_{i=1}^{I_{rp}} Y_{pi}^o$ , is used to replace all of his/her missing scores. For either case, the means can be rounded to 0 or 1 binary values, but this creates rounding-off errors. Alternatively, an IM or PM can be used as a parameter to make a random draw from the Bernoulli distribution for the missing binary data (Sijtsma & van der Ark, 2003).

### Two-way mean substitution (TW)

IM substitution corrects for score differences between items, but not for differences between persons. Similarly, PM substitution corrects for score differences between persons,

but not for differences between items. Therefore, Bernaards and Sijtsma (2000) proposed *two-way mean substitution*. Let  $OM$  be the overall mean of all observed item scores. Then a deterministic TW mean for person  $p$  on item  $i$  is calculated as  $TW_{pi} = PM_p + IM_i - OM$ . For binary item scores, the computed  $TW_{pi}$  can be rounded to 0 or 1, or used as a parameter to sample from a Bernoulli distribution. The TW method assumes that no person  $\times$  item interaction is present.

### Corrected item mean substitution (CIM)

Corrected item mean (CIM) substitution (Bernaards & Sijtsma, 2000; Huisman, 2000; Huisman & Molenaar, 2001) improves on unconditional item mean imputation (IM) by taking into account the overall mean performance of a person on the items he/she answered, as well as the difficulty of those items. Here, a weight reflecting the relative performance of a person on his/her observed item scores is calculated and applied to the mean performance on the item across persons, thus increasing the likelihood for imputing a correct score for persons whose relative performance is higher (Finch, 2008). In other words, the item mean is multiplied by a factor reflecting the ratio between the student's scores on available items and the average of available scores on these items. Again, these means can be rounded to binary values or used as parameters to sample from a Bernoulli distribution. A CIM for person  $p$  on item  $i$ —that is, cell  $(p, i)$ —is calculated as  $CIM_{pi} = \left( \frac{PM_p}{\frac{1}{I_{rp}} \sum_{i \in I_{rp}} IM_i} \right) \times IM_i$

### Regression imputation (RI)

The regression imputation method (see, e.g., Little & Rubin, 2002) takes information from observed auxiliary variables into account. Once again, consider scores on an item  $i$ , with  $Y_i^0$  observed for all  $p \in P_{ri}$  and  $Y_i^m$  missing for all  $p \in P_{mi}$ . Regression imputation uses persons' data to regress  $Y_i^0$  on observed auxiliary variables and then computes the missing values as predictions from the regression. For a qualitative dependent variable  $Y$ , logistic models may be used, and the auxiliary variables can be quantitative or qualitative—the latter being incorporated by means of dummy variables (Kalton & Kasprzyk, 1982). It is also possible to include useful interaction terms and transformations of the variables. Specifically, for predictor variables  $X = X_1, \dots, X_J$ , a missing value  $Y_{pi}^m \in Y_i^m$  is imputed using the regression model given as

$$\begin{aligned} \text{logit}[\pi_i(X)] &= \alpha_i + \sum_{j \in J} \varphi_{ij} X_j, \quad \text{with} \\ \pi_i(X) &= P(Y_{pi}^0 = 1 | X = X_1, \dots, X_J). \end{aligned} \quad (4)$$

The imputed score can then be estimated by  $Y_{pi}^m \sim \text{binomial}[1, \pi_i(X)]$ . A special case of the RI method is the ratio imputation method, in which  $Y_i^0$  is regressed on a single auxiliary variable and an intercept of zero (see, e.g., Arnab & Singh, 2006; Kalton & Kasprzyk, 1982; Shao, 2000). Therefore, ratio imputation will not be considered separately in this study, but only the more general regression imputation method.

### Multiple imputation (MI)

In all methods discussed up to now, a missing score was replaced by one single estimated value. Single-value imputation methods can be easily implemented, but replacing missing values with a single value is likely to decrease the variance that would be present if data were fully observed (Schafer & Graham, 2002). For instance, replacing each missing score in  $Y_i^m$  for all  $p \in P_{mi}$  with the mean of the observed scores in  $Y_i^0$  will likely underestimate the true variance had all item  $i$  scores been observed. Through an example, Baraldi and Enders (2010) illustrated that several single-value imputation methods introduce bias in the estimates of means, variances, and correlations. The main problem is that inferences about parameters based on the filled-in values do not account for imputation uncertainty, since variability from not knowing the missing values is ignored (Rubin & Schenker, 1986). As a result, the standard errors and confidence intervals of estimates based on imputed data can be underestimated (Little & Rubin, 2002). One alternative to single-value imputation methods is multiple imputation.

Multiple imputation (MI; Rubin, 1987) is a technique in which each missing item score is replaced with a set of  $M > 1$  plausible values according to a statistical model. Different imputation models can be adopted—for instance, using the previously described methods—but the exact choice typically depends on the nature of the variable to be imputed in the data set. For categorical variables, the log-linear model (Schafer, 1997) would be the most appropriate imputation model. However, the log-linear model can be applied only when the number of variables used in the imputation model is small (Vermunt, van Ginkel, van der Ark, & Sijtsma, 2008), whereby it is feasible to set up and process the full multiway cross-tabulation required for the log-linear analysis. As an alternative, imputations can be carried out under a fully conditional specification approach (van Buuren, 2007), where a sequence of regression models for the univariate conditional distributions of the variables with missing values are specified. Under this approach, a variable with missing binary scores can be modeled as a logistic regression model of the observed scores  $Y_i^0$  on the observed auxiliary variables, as in Eq. 4, with possible higher-order interaction terms or transformed variables (van Buuren & Oudshoorn, 2000).

Such a model is used to generate  $M$  imputations (or synthetic values) for the missing observations in the data by making random draws from a distribution of plausible values. This step results in  $M$  “complete” data sets, each of which is analyzed using a statistical method that will estimate the quantities of scientific interest. The resulting  $M$  analyses (instead of one) will differ because the imputations differ (van Buuren, 2007). Following Rubin’s rules (e.g., Little & Rubin, 2002), the results of the  $M$  analyses are then combined into a single estimate of the statistics of interest by combining the variation within and across the  $M$  imputed data sets. As such, the uncertainty about the imputed values is taken into account, and under fairly liberal conditions, statistically valid estimates with unbiased standard errors can be obtained.

### Direct likelihood (DL) analysis

Instead of imputing missing observations, Mallinckrodt, Clark, Carroll, and Molenberghs (2003) proposed the use of the direct likelihood methodology to deal with incomplete correlated data for ignorable missing-data mechanisms—that is, for MCAR or MAR (Little & Rubin, 2002). This approach is referred to as *likelihood-based ignorable analysis*, or simply *direct likelihood analysis* (see, e.g., Molenberghs & Kenward, 2007). Under the DL approach, all of the available observed data are analyzed without deletion nor imputation using models that offer a framework from which to analyze clustered data by including both the fixed and random effects in the model—for example, GLMMs for non-Gaussian data. In so doing, appropriate adjustments valid under the ignorable missing-data mechanism are made to parameters, even when the data are incomplete, due to the within-person correlation (Beunckens et al. 2005).

As noted before, IRT models can be reformulated in the framework of GLMM (De Boeck & Wilson, 2004), with persons as clusters and items for the repeated observations, and parameter estimates can easily be obtained using tools like the lmer function of the lme4 package for the R software (De Boeck, Bakker, Zwitser, Nivard, Hofman, Tuerlinckx & Partchev 2011). The GLMM for the observed binary item scores  $Y_{pi}^0$  is given as  $\text{logit}(\pi_{pi}) = \eta_{pi}$  and  $Y_{pi}^0 \sim \text{binomial}(1, \pi_{pi})$ . For each pair of a person  $p$  and an item  $i$ ,  $(p, i)$ , the value of the component  $\eta_{pi}$  is determined by a linear combination of the person and item predictors (De Boeck et al. 2011; De Boeck & Wilson, 2004), as given in Eq. 3. In other words, the same model that would be used for a fully observed data set is now fitted with the same software tools to a reduced data set with persons having unequal sets of item scores. For an extensive description of the DL approach as a method for analyzing clustered data with missing observations, the interested reader is referred to Molenberghs and Kenward (2007).

## Simulation study

A simulation study was conducted to compare the performance of the aforementioned methods. The study was based on “population” values similar to the estimated values obtained from two different empirical data sets—namely, the CTB–McGraw-Hill data used in De Boeck and Wilson (2004), and the data from a test for assessing attainment targets of reading comprehension in Dutch for pupils leaving primary education (De Boeck, Daems, Meulders, & Rymenans 1997). The latter data were reanalyzed by Van den Noortgate, De Boeck, and Meulders (2003) as an application for cross-classification multilevel logistic models. Both data sets depict a multilevel item response structure, with students grouped in schools but differing in the variance values of school effects, person abilities, and item difficulties. The Dutch comprehension data have a large difference between the variance of person abilities and that of item difficulties, with about 15% of the differences in item scores situated between schools. The CTB data have a small difference between the variances of person abilities and item difficulties, with about 30% of the differences in item scores situated between schools. Analyses based on the CTB data are used for comparison purposes, to understand the performance of the discussed methods given different population values.

### Simulating data

Both empirical data sets have a number of predictors, but for simplicity, one fixed person property and one fixed item property were considered, and clustering was done only at the person level, whereby persons are grouped into schools with equal probabilities. In both situations, persons, items, and schools were treated as random effects, while covariates of gender (male = 0, female = 1) and type of text for item tasks were considered as fixed effects. Type of text is a dummy variable, with 0 assigned to evaluation or science items and 1 assigned to other item types with a probability of .5, simply to obtain a generally balanced design. Population values for these factors were obtained by fitting Eq. 5 to the two empirical data sets.

$$\text{logit}(\pi_{kpi}) = \gamma_0 + \gamma_1 \text{gender}_p + \gamma_2 \text{type}_i + w_p + u_k + \varepsilon_i, \quad (5)$$

with  $u_k \sim N(0, \sigma_u^2)$ ,  $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ ,  $w_p \sim N(0, \sigma_w^2)$ . The score on item  $i$  by person  $p$  from school  $k$  ( $\text{bin}_{kpi}$ ) was then generated from a binomial distribution as  $\text{bin}_{kpi} \sim \text{binomial}(1, \pi_{kpi})$ . For each of the two empirical studies, 1,000 fully observed sample data sets were simulated, and the following factors were considered for simulating the data.

### Clustering

For simulations based on the Dutch comprehension data, two cases were considered—namely, no variability between person groups (school effects,  $\sigma_u^2 = 0$ ) versus a case in which there was variability between schools ( $\sigma_u^2 > 0$ )—in order to examine the effect of school clusters. For the simulations based on the CTB data, only the latter case was considered. The number of schools was fixed at  $K = 20$  for both simulation cases, which is the average number of schools in the two empirical data sets.

### Sample sizes

Two item sample sizes were compared—that is,  $I = 25$  and 50—while the persons’ sample size was fixed at  $P = 120$ —that is, 6 persons per school.

### Missing proportions ( $\zeta$ )

In each of the samples, a random selection of 15% or 50% of the item scores were removed, yielding low and high proportions of missing scores. These values were based on our experience with e-learning environments, which tend to have high levels of item nonresponse.

### Missing mechanisms

For all simulated samples, an ignorable missing-data mechanism was induced in item scores—that is, MCAR or MAR. To induce MCAR, item scores were removed at random from the data with equal probabilities of responding  $\pi(\text{resp}_{kpi} = 1) = .85$  and  $\pi(\text{resp}_{kpi} = 1) = .50$  for the 15% and 50% proportions of missing scores, respectively. A new response variable,  $\text{mcar\_bin}$ , was deducted from the original scores as  $\text{mcar\_bin}_{kpi} = \begin{cases} \text{bin}_{kpi}, & \text{if } \text{resp}_{kpi} = 1 \\ NA, & \text{if } \text{resp}_{kpi} = 0 \end{cases}$ , with  $\text{resp}_{kpi} \sim \text{binomial}[1, \pi(\text{resp}_{kpi})]$  and “NA” meaning that the item score is missing. To induce MAR, the probability of responding  $\pi(\text{resp}_{kpi} = 1)$  was dependent on a person’s gender. Girls were assigned a high probability of responding to items relative to boys, as shown in Table 1. Then, similar procedures were followed for the MCAR condition, to create a variable  $\text{mar\_bin}_{kpi}$  with scores MAR.

In Table 2, an overview of all of the factors and design characteristics for both simulation cases is given. Also presented in Table 2 are the “population” values—that is, the regression coefficients for the fixed effects and random effects variance estimates obtained after reanalyzing the two empirical data sets using the GLMM in Eq. 5.

**Table 1** Probabilities of responding for creating data missing at random

	Level of Response		Condition
	$\zeta = 15\%$	$\zeta = 50\%$	
$\pi(\text{resp}_{kpi} = 1) = \begin{cases} 0.95, & 0.60, \\ 0.75, & 0.40, \end{cases}$			$\text{if }  \text{gender}_p = 1$ $\text{if }  \text{gender}_p = 0$

$\pi(\text{resp}_{kpi} = 1)$  refers to the probability that person  $p$  in school  $k$  responds to item  $i$

### Imputation methods

Missing data were then estimated according to six imputation methods: SRI, ARI, IM, TW, CIM, and RI. In addition, DL analysis was conducted while the completely observed data (here referred to as *CD*) were analyzed for reference purposes. For the ARI, IM, TW, CIM, and RI imputation methods, three variations were possible—that is, (a) single-value imputation by rounding each estimated mean to 0 or 1, (b) single-value imputation by using each computed mean as a parameter to sample 0 or 1 values from a Bernoulli distribution, and (c) multiple imputation by using each computed mean as a parameter to sample, for each missing value, several values from a Bernoulli distribution, and then analyzing and combining the estimates to make one inference using Rubin's rules. Some preliminary analyses (not discussed further in this text) showed that deterministic single-value imputation methods (possibilities a and b) did not perform well, and the results

were unreliable. This was as expected, because single-value imputation methods do not take uncertainty about the missing values into account, such that estimates based on imputed data are biased. Therefore, the ARI, IM, TW, CIM, and RI methods were adapted to a multiple-imputation approach (possibility c).

For methods adapted to a multiple-imputation approach, the basic rule of thumb is that the number of imputations  $M$  is set to  $3 \leq M \leq 5$  to get sufficient inferential accuracy. Even though Schafer (1997) noted that basing conclusions on this range might be risky, Molenberghs and Verbeke (2005) showed that efficiency gains rapidly diminish after the first  $M = 2$  imputations for small fractions of missing information, and after the first  $M = 5$  for large fractions of missing information. For both simulation studies,  $M = 5$  was used in all conditions.

The RI method was implemented using a logistic regression model as incorporated in multivariate imputation using chained equations (MICE; van Buuren & Oudshoorn, 2000), a missing-data package for the R software, but an extra choice to make was the variables to include in the imputation model. Even though the MICE software can automatically select variables to include in the model, a proper multiple imputation requires correct adjustments to the model to reflect the variables of interest and any possible interaction effects—a process that is not straightforward. For this analysis, a person's gender, item indicators, and school indicators, and an interaction between the person's gender and an item's text type were all included as fixed effects.

**Table 2** Design characteristics of the two simulation studies

Study 1: Based on Dutch Comprehension Data	
Clustering:	None and persons in schools
Number of clusters:	$K = 20$
Item sample size:	$I = 25$ and $I = 50$
Person sample size:	$P = 120$
Predictors:	Intercept ( $\gamma_0 = 0.24$ ), gender ( $\gamma_1 = -0.14$ ), type ( $\gamma_2 = 0.50$ )
Variance of random effects:	Persons ( $\sigma_\omega^2 = 0.45$ ), items ( $\sigma_\epsilon^2 = 1.24$ ), schools ( $\sigma_u^2 = 0$ ; $\sigma_u^2 = 0.08$ )
Percent of missing scores:	$\zeta = 15\%$ and $\zeta = 50\%$
Missing data mechanism:	MCAR and MAR
Methods	SRI, ARI, IM, TW, CIM, RI, DL and CD
Study 2: Based on CTB Data in De Boeck and Wilson (2004)	
Clustering:	Persons in schools
Number of clusters:	$K = 20$
Item sample size:	$I = 50$
Person sample size:	$P = 120$
Predictors:	Intercept ( $\gamma_0 = 0.82$ ), gender ( $\gamma_1 = -0.1$ ), type ( $\gamma_2 = -0.28$ )
Variance of random effects:	Persons ( $\sigma_\omega^2 = 0.65$ ), items ( $\sigma_\epsilon^2 = 0.75$ ), schools ( $\sigma_u^2 = 0.30$ )
Percent of missing scores:	$\zeta = 15\%$ and $\zeta = 50\%$
Missing data mechanism:	MAR
Methods	SRI, ARI, IM, TW, CIM, RI, DL and CD



## Analysis procedure

In total, there were 18 conditions—namely, 2 (clustering situations)  $\times$  2 (item sample sizes)  $\times$  2 (missing proportions)  $\times$  2 (missing mechanisms) for simulations based on the Dutch comprehension data, and 1 (clustering situation)  $\times$  1 (item sample size)  $\times$  2 (missing proportions)  $\times$  1 (missing mechanisms) for simulations based on the CTB data. For all conditions, the simulated 1,000 data sets were analyzed by fitting the explanatory IRT from Eq. 5 to the fully observed data (CD), to the reduced/incomplete data (DL analysis), and to data whose missing item scores had been substituted using each of the six imputation methods. The purpose of fitting a model similar to the one that was used for simulating the fully observed data was to assess whether “population” values can be recovered from the data whose originally observed item scores have been deleted and are analyzed, either without imputation (for the DL analysis) or after the deleted item scores have been estimated by one of the imputation methods, under the different conditions described in preceding sections.

The outcomes of interest were the accuracy of the fixed parameters of gender and item type, of the standard errors of these fixed effects, and of the estimates for random variables’ variance over persons ( $\sigma_{\omega}^2$ ), schools ( $\sigma_u^2$ ), and items ( $\sigma_{\epsilon}^2$ ). Therefore, we will look at the bias, precision, and mean squared error (*MSE*) of the estimates of the fixed and random parameters. Bias is the difference between an estimator’s expected value and the true value of the parameter being estimated. *MSE* is a measure to quantify the accuracy of prediction, and is given by the sum of the variance (or the squared standard error) and the squared bias of an estimator. Fixed effects’ expected values are estimated as the mean of the 1,000 estimates, while random effects’ expected variances were estimated as the median of the estimates, because variance estimates are bounded above zero and positively skewed. The standard error estimates were evaluated by comparing the mean standard error estimate of each parameter to the standard deviation of 1,000 estimates of the corresponding parameter, and by looking at the coverage proportions of the respective confidence intervals. Model fitting and programming were done in the R statistical software, version 2.12.1. For each simulated data set, model parameters were estimated using the `lmer()` function in the `lme4` package (Bates & Maechler, 2010) under REML with the Laplace approximation.

## Results

Analyses from the small item samples showed trends similar to those from the large item samples for all conditions of this study, although standard errors for the small item samples

seemed to be biased even for a complete data analysis. Furthermore, there were no differences in the conclusions for analyses of samples drawn from the Dutch comprehension data and from the CTB data. Here, we present and discuss only the results obtained for the large item sample sizes drawn from the Dutch comprehension data. Results from smaller samples and from samples drawn from the CTB data were very similar and can be obtained from the authors.

## Fixed-effect parameter estimates

### Bias

Tables 3 and 4 show the bias and mean squared errors of fixed-effect estimates by imputation method, clustering condition, and missing mechanism at the 15% and 50% proportions of missing item scores, respectively. For the two clustering conditions, the results indicated that bias was generally larger in the MAR condition than in the MCAR condition across both missing proportions. The DL and TW methods resulted in bias that was comparable to that of the CD analysis, in which there were no missing item scores. Furthermore, bias was largest in all conditions when missing scores were imputed using the SRI, ARI, and IM methods. When missing scores were imputed using the CIM and RI methods, bias was small but not similar to that in the CD analysis. There was generally an increase in bias with increases in the proportions of missing item scores, except with the DL method, as can be seen when comparing Tables 3 and 4.

### Mean squared error

For the 15% proportion of missing item scores and for both clustering conditions, the *MSE* estimates in the MAR condition were similar to those in the MCAR condition, as can be observed in Table 3. However, for the 50% proportion of missing item scores, in Table 4, *MSE* estimates in the MAR condition were larger than those in the MCAR condition. For the 15% proportion of missing scores, *MSE* values for the DL analysis and the analyses after missing scores have been imputed using the TW and CIM methods were comparable to those of the CD analysis. This was not the case for the 50% proportion of missing scores, where the *MSE* was larger for the TW and CIM methods than for CD. In addition, the results indicate that *MSE* estimates were smaller than those from the CD analysis when missing item scores were imputed using the SRI, ARI, IM, and RI methods in all conditions. Smaller *MSE* values for the analyses after missing item scores have been imputed by the SRI, ARI, IM, and RI methods are an indication that the standard errors of fixed-effect estimates were underestimated, due to reduced variability in the imputed item scores for

**Table 3** Bias and mean squared error in fixed-effect estimates by imputation method and missing mechanism for  $\zeta = 15\%$ 

Method	No School Clustering						School Clustering					
	Bias			MSE			Bias			MSE		
	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$
$\gamma_0 = 0.24, \gamma_1 = -0.14, \gamma_2 = 0.50$												
CD	.008	-.006	-.001	.059	.018	.101	.004	-.007	.010	.063	.020	.100
MCAR												
DL	.007	-.005	-.001	.060	.019	.102	.003	-.005	.010	.063	.020	.101
SRI	-.009	.020	-.011	.055	.014	.099	-.014	.021	-.002	.058	.015	.097
ARI	-.012	.019	-.018	.054	.014	.096	-.017	.021	-.009	.056	.015	.093
IM	-.009	.021	-.011	.056	.015	.100	-.014	.021	-.002	.058	.016	.097
TW	.007	-.004	-.004	.060	.020	.102	.004	-.005	.006	.064	.022	.101
CIM	-.001	-.003	-.007	.060	.020	.101	-.005	-.004	.004	.063	.021	.100
RI	.001	-.002	-.009	.057	.015	.100	-.003	-.002	.002	.061	.016	.099
MAR												
DL	.008	-.007	-.001	.060	.019	.102	.005	-.007	.012	.064	.021	.101
SRI	-.020	.024	-.010	.056	.014	.099	-.025	.024	.002	.058	.015	.097
ARI	-.026	.031	-.017	.054	.015	.096	-.032	.032	-.008	.057	.016	.093
IM	-.020	.025	-.011	.057	.015	.100	-.025	.026	.000	.059	.016	.097
TW	.008	-.005	-.004	.061	.020	.102	.004	-.005	.008	.064	.022	.101
CIM	-.005	.004	-.005	.060	.020	.102	-.011	.005	.007	.064	.021	.100
RI	-.001	.002	-.009	.058	.015	.100	-.005	.002	.003	.062	.016	.099

CD = complete data, DL = direct likelihood, SRI = simple random imputation, ARI = adjusted random imputation, IM = item means, TW = two-way means, CIM = corrected item means, RI = regression imputation

these methods. This was more pronounced for the analyses of data sets whose missing scores were imputed by the SRI, ARI, and IM methods, especially because greater biases were associated with these three methods.

### Clustering

With persons clustered in schools (the six columns on the right of Tables 3 and 4), bias for fixed effects generally seemed to be slightly larger than when persons were not clustered within schools for the SRI, ARI, IM, CIM, and RI imputation methods, in both the MCAR and MAR conditions. For all methods, *MSE* values seemed to be larger for the case of school clustering than when persons were not clustered within schools. This was not unexpected, due to the additional variability from school effects. Generally for both clustering conditions, the DL analysis and the analysis after imputation of missing scores by the TW method seemed to result in biases and *MSE* values comparable to those of the CD analysis.

### Standard errors

Tables 5 and 6 show the mean of the standard errors and standard deviations of the fixed-effect estimates by impu-

tation method, clustering condition, and missing mechanism for 15% and 50% proportions of missing item scores, respectively. Standard deviations of the 1,000 estimates were better approximated by the estimated standard errors for the condition of no school clustering, as compared to when persons were clustered within schools. This was probably due to the small group sample sizes per school. Standard errors were found to be smaller for the SRI, ARI, and IM approaches, but this was also true for the standard deviations. In general, for both the MCAR and MAR conditions and both high and low proportions of missing item scores, there was little variation across imputation methods in the difference between the standard errors and the corresponding standard deviations.

### Coverage probabilities

The 95% confidence interval coverage probabilities by imputation method and missing mechanism are shown in Table 7. Good confidence procedures should have coverage probabilities equal (or close to) the nominal level of 95%. For all methods, the coverage probabilities of the confidence intervals for all fixed effects were generally close to the nominal value for the 15% proportion of missing item scores. However, for the 50% proportion of missing item

**Table 4** Bias and mean squared error in fixed-effect estimates by imputation method and missing mechanism for  $\zeta = 50\%$

Method	No School Clustering						School Clustering					
	Bias			MSE			Bias			MSE		
	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$
$\gamma_0 = 0.24, \gamma_1 = -0.14, \gamma_2 = 0.50$												
CD	.008	-.006	-.001	.059	.018	.101	.004	-.007	.010	.063	.020	.100
MCAR												
DL	.010	-.010	-.002	.062	.022	.105	.004	-.006	.009	.065	.023	.103
SRI	-.037	.069	-.021	.053	.011	.100	-.046	.077	-.018	.053	.013	.095
ARI	-.046	.073	-.039	.050	.012	.093	-.052	.073	-.031	.051	.012	.090
IM	-.039	.071	-.021	.055	.014	.102	-.046	.075	-.018	.055	.015	.098
TW	.014	-.010	-.011	.065	.026	.107	.006	-.005	.000	.068	.028	.104
CIM	-.021	-.005	-.020	.063	.025	.104	-.030	.001	-.011	.066	.026	.100
RI	-.008	-.001	-.020	.061	.012	.107	-.014	.002	-.009	.065	.013	.105
MAR												
DL	.010	-.009	-.003	.063	.022	.105	.005	-.006	.010	.066	.024	.104
SRI	-.054	.074	-.027	.055	.012	.100	-.062	.078	-.016	.055	.013	.096
ARI	-.063	.081	-.035	.052	.013	.093	-.067	.080	-.033	.052	.014	.090
IM	-.055	.075	-.023	.057	.014	.102	-.061	.078	-.016	.057	.015	.098
TW	.013	-.007	-.011	.066	.027	.107	.006	-.004	.002	.069	.028	.105
CIM	-.031	.011	-.018	.065	.025	.105	-.014	.017	-.007	.068	.027	.102
RI	-.012	.007	-.023	.062	.013	.107	-.017	.009	-.011	.066	.013	.106

CD = complete data, DL = direct likelihood, SRI = simple random imputation, ARI = adjusted random imputation, IM = item means, TW = two-way means, CIM = corrected item means, RI = regression imputation

**Table 5** Standard errors and standard deviations of fixed-effect estimates by imputation method and missing mechanism for  $\zeta = 15\%$

Method	No School Clustering						School Clustering					
	SE			SD			SE			SD		
	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$
$\gamma_0 = 0.24, \gamma_1 = -0.14, \gamma_2 = 0.50$												
CD	.243	.135	.318	.245	.137	.310	.251	.141	.316	.261	.145	.330
MCAR												
DL	.244	.138	.320	.246	.140	.310	.252	.143	.317	.261	.148	.331
SRI	.235	.117	.314	.240	.118	.304	.239	.121	.311	.253	.123	.324
ARI	.231	.117	.309	.236	.116	.300	.236	.120	.305	.249	.122	.318
IM	.236	.120	.316	.239	.117	.305	.240	.124	.312	.252	.122	.324
TW	.245	.142	.320	.245	.140	.308	.252	.147	.317	.261	.148	.330
CIM	.244	.140	.318	.245	.138	.306	.251	.145	.315	.260	.146	.327
RI	.239	.123	.316	.242	.138	.305	.247	.127	.315	.256	.146	.326
MAR												
DL	.245	.138	.319	.247	.142	.310	.252	.143	.317	.264	.148	.331
SRI	.236	.118	.315	.237	.119	.305	.240	.122	.311	.252	.124	.324
ARI	.232	.118	.309	.233	.117	.301	.236	.122	.306	.247	.123	.318
IM	.237	.121	.316	.237	.116	.305	.241	.125	.312	.252	.121	.324
TW	.246	.142	.320	.245	.142	.309	.253	.147	.317	.263	.148	.329
CIM	.245	.141	.319	.247	.140	.308	.252	.146	.316	.263	.146	.327
RI	.241	.124	.317	.241	.141	.307	.249	.127	.314	.258	.147	.326

CD = complete data, DL = direct likelihood, SRI = simple random imputation, ARI = adjusted random imputation, IM = item means, TW = two-way means, CIM = corrected item means, RI = regression imputation

**Table 6** Standard errors and standard deviations of fixed-effect estimates by imputation method and missing mechanism for  $\zeta = 50\%$ 

Method	No School Clustering						School Clustering					
	SE			SD			SE			SD		
	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$
$\gamma_0 = 0.24, \gamma_1 = -0.14, \gamma_2 = 0.50$												
CD	.243	.135	.318	.245	.137	.310	.251	.141	.316	.261	.145	.330
MCAR												
DL	.248	.147	.324	.249	.148	.316	.255	.152	.321	.267	.162	.334
SRI	.228	.081	.315	.235	.082	.305	.226	.083	.308	.247	.089	.319
ARI	.219	.082	.302	.227	.079	.296	.219	.084	.298	.235	.086	.306
IM	.231	.093	.319	.233	.072	.306	.230	.095	.312	.244	.078	.319
TW	.255	.162	.326	.248	.149	.311	.261	.166	.322	.265	.162	.327
CIM	.250	.157	.321	.250	.144	.307	.255	.161	.317	.263	.157	.321
RI	.247	.111	.327	.239	.147	.305	.254	.112	.324	.258	.161	.323
MAR												
DL	.250	.148	.324	.252	.151	.316	.257	.153	.322	.268	.158	.331
SRI	.228	.082	.315	.233	.081	.307	.227	.084	.309	.245	.083	.318
ARI	.220	.083	.303	.227	.080	.296	.218	.085	.297	.238	.083	.309
IM	.232	.094	.318	.232	.071	.304	.231	.096	.313	.245	.074	.317
TW	.257	.163	.327	.252	.154	.311	.262	.168	.324	.267	.157	.325
CIM	.252	.157	.323	.252	.146	.308	.257	.162	.319	.267	.152	.321
RI	.249	.112	.327	.240	.152	.305	.257	.114	.320	.257	.160	.320

CD = complete data, DL = direct likelihood, SRI = simple random imputation, ARI = adjusted random imputation, IM = item means, TW = two-way means, CIM = corrected item means, RI = regression imputation

scores, the coverage probabilities of the confidence intervals of estimates obtained after imputations by the SRI, ARI, and RI methods could fall farther below the nominal value, while those obtained from the DL analyses and the analyses after imputations by the IM, TW, and CIM methods were generally comparable to those of the CD analyses for all conditions.

#### Random-effect variance estimates

For both clustering conditions, the CD analysis resulted in the median variance estimate that was about 95% of the true variability within person abilities and within item difficulties, and about 90% of the true variability within school effects for the case of school clustering. Variance estimates therefore were slightly negatively biased.

#### Bias

Tables 8 and 9 show the bias and *MSE* values of the random-effect variance estimates by imputation method, clustering condition, and missing mechanism at the 15% and 50% proportions of missing item scores, respectively. For all conditions, the obtained results showed greater bias of all random effects' variance estimates at the 50% than at the 15% proportion of missing item scores. The DL analysis resulted in bias that was comparable to that of

the CD analysis for all conditions, and more especially at the 15% proportion of missing item scores. Although the biases of the DL analyses at the 50% proportion of missing item scores were increased, they were the least relative to those obtained after estimating missing scores by any of the imputation methods. For all conditions, analyses after imputing missing scores by the SRI, ARI, IM, and RI methods resulted in larger negative bias, while analyses after imputations by the TW and CIM methods resulted in small absolute biases.

#### Mean squared error

For all conditions, the obtained *MSE* values based on the DL analysis for all random-effect variance estimates were comparable to those of the CD analysis. Analyses of data sets whose missing item scores were imputed by the TW and CIM methods showed *MSE* values that were comparable to those of the CD analysis for the variance estimates of person abilities and school effects, but not of item difficulties. This was more noticeable at the 15% than the 50% proportion of missing item scores. For all conditions shown in Tables 8 and 9, analyses after imputations by the SRI, ARI, IM, and RI methods resulted in greater *MSE* values. In general, there was an increased loss of prediction accuracy for analyses of data sets whose missing item scores were estimated using any of the imputation methods

**Table 7** 95% confidence interval coverage probabilities by imputation method and missing mechanism

Method	No School Clustering						School Clustering					
	$\zeta = 15\%$			$\zeta = 50\%$			$\zeta = 15\%$			$\zeta = 50\%$		
	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$	$\gamma_0$	$\gamma_1$	$\gamma_2$
CD	.95	.94	.95	.95	.94	.95	.94	.94	.94	.94	.94	.94
MCAR												
DL	.94	.94	.94	.95	.95	.95	.94	.94	.93	.94	.93	.94
SRI	.94	.93	.95	.93	.86	.95	.93	.95	.93	.91	.82	.92
ARI	.94	.94	.95	.94	.87	.95	.94	.94	.94	.92	.86	.94
IM	.94	.94	.95	.94	.94	.95	.94	.96	.94	.93	.91	.94
TW	.94	.94	.95	.95	.96	.95	.94	.95	.94	.95	.96	.94
CIM	.94	.95	.95	.94	.96	.95	.94	.95	.94	.94	.96	.94
RI	.94	.91	.95	.96	.86	.96	.94	.92	.94	.95	.82	.95
MAR												
DL	.95	.94	.95	.95	.94	.95	.94	.95	.94	.94	.94	.94
SRI	.95	.94	.95	.94	.85	.95	.94	.95	.94	.91	.85	.94
ARI	.95	.94	.95	.94	.84	.95	.94	.94	.94	.92	.84	.94
IM	.95	.95	.95	.95	.93	.96	.94	.96	.94	.93	.92	.94
TW	.95	.94	.96	.96	.96	.96	.94	.96	.94	.94	.96	.94
CIM	.95	.95	.95	.95	.96	.96	.94	.95	.94	.94	.96	.94
RI	.96	.90	.95	.96	.84	.96	.95	.92	.94	.95	.83	.94

CD = complete data, DL = direct likelihood, SRI = simple random imputation, ARI = adjusted random imputation, IM = item means, TW = two-way means, CIM = corrected item means, RI = regression imputation

**Table 8** Bias and mean square error in random-effect variance estimates by imputation method and missing mechanism for  $\zeta = 15\%$

Method	No School Clustering						School Clustering					
	Bias			MSE			Bias			MSE		
	$\sigma_\omega^2$	$\sigma_u^2$	$\sigma_\varepsilon^2$	$\sigma_\omega^2$	$\sigma_u^2$	$\sigma_\varepsilon^2$	$\sigma_\omega^2$	$\sigma_u^2$	$\sigma_\varepsilon^2$	$\sigma_\omega^2$	$\sigma_u^2$	$\sigma_\varepsilon^2$
$\sigma_\omega^2 = 0.45, \sigma_u^2 = 0.00, \sigma_\varepsilon^2 = 1.24$						$\sigma_\omega^2 = 0.45, \sigma_u^2 = 0.08, \sigma_\varepsilon^2 = 1.24$						
CD	-.025	.000	-.050	.006	.000	.079	-.019	-.009	-.078	.007	.003	.074
MCAR												
DL	-.027	.000	-.052	.007	.000	.081	-.020	-.009	-.076	.007	.004	.076
SRI	-.160	.000	-.091	.028	.000	.083	-.157	-.033	-.119	.028	.003	.081
ARI	-.166	.000	-.132	.031	.000	.087	-.162	-.033	-.163	.029	.003	.087
IM	-.159	.001	-.086	.028	.000	.082	-.157	-.032	-.114	.028	.003	.078
TW	.006	.001	-.052	.006	.000	.076	.012	-.010	-.080	.007	.003	.071
CIM	-.005	.001	-.067	.005	.000	.076	.003	-.011	-.091	.006	.003	.071
RI	-.152	.021	-.076	.026	.001	.081	-.152	.015	-.098	.027	.004	.076
MAR												
DL	-.026	.000	-.052	.006	.000	.081	-.021	-.008	-.074	.007	.004	.076
SRI	-.155	.000	-.083	.027	.000	.082	-.153	-.031	-.111	.027	.003	.080
ARI	-.160	.000	-.128	.029	.000	.085	-.153	-.034	-.156	.027	.003	.086
IM	-.155	.001	-.084	.027	.000	.081	-.152	-.032	-.117	.026	.003	.080
TW	.010	.001	-.053	.006	.000	.075	.015	-.008	-.071	.007	.004	.072
CIM	-.001	.001	-.061	.005	.000	.076	.004	-.011	-.083	.006	.003	.073
RI	-.148	.021	-.077	.025	.001	.079	-.149	.016	-.094	.026	.004	.076

CD = complete data, DL = direct likelihood, SRI = simple random imputation, ARI = adjusted random imputation, IM = item means, TW = two-way means, CIM = corrected item means, RI = regression imputation

**Table 9** Bias and mean square error in random-effect variance estimates by imputation method and missing mechanism for  $\zeta = 50\%$ 

Method	No School Clustering						School Clustering					
	Bias			MSE			Bias			MSE		
	$\sigma_{\omega}^2$	$\sigma_u^2$	$\sigma_{\varepsilon}^2$	$\sigma_{\omega}^2$	$\sigma_u^2$	$\sigma_{\varepsilon}^2$	$\sigma_{\omega}^2$	$\sigma_u^2$	$\sigma_{\varepsilon}^2$	$\sigma_{\omega}^2$	$\sigma_u^2$	$\sigma_{\varepsilon}^2$
	$\sigma_{\omega}^2 = 0.45, \sigma_u^2 = 0.00, \sigma_{\varepsilon}^2 = 1.24$						$\sigma_{\omega}^2 = 0.45, \sigma_u^2 = 0.08, \sigma_{\varepsilon}^2 = 1.24$					
CD	-.025	.000	-.050	.006	.000	.079	-.019	-.009	-.078	.007	.003	.074
MCAR												
DL	-.037	.000	-.066	.009	.001	.090	-.031	-.007	-.084	.010	.004	.082
SRI	-.364	.000	-.091	.133	.000	.095	-.363	-.066	-.131	.133	.005	.089
ARI	-.366	.000	-.186	.135	.000	.105	-.366	-.067	-.212	.135	.005	.108
IM	-.364	.002	-.089	.133	.000	.090	-.363	-.067	-.129	.132	.005	.086
TW	.133	.003	-.020	.025	.001	.069	.140	-.010	-.046	.028	.004	.062
CIM	.088	.003	-.068	.014	.000	.073	.089	-.014	-.092	.016	.003	.066
RI	-.357	.082	-.060	.128	.008	.085	-.359	.074	-.064	.129	.009	.075
MAR												
DL	-.044	.000	-.053	.011	.001	.085	-.025	-.011	-.082	.010	.004	.083
SRI	-.361	.000	-.079	.131	.000	.089	-.358	-.067	-.123	.129	.005	.088
ARI	-.364	.000	-.169	.133	.000	.098	-.361	-.069	-.216	.131	.005	.107
IM	-.360	.002	-.075	.130	.000	.084	-.358	-.068	-.123	.129	.005	.084
TW	.135	.004	-.006	.027	.001	.066	.151	-.010	-.044	.032	.004	.063
CIM	.090	.003	-.043	.015	.000	.069	.099	-.018	-.077	.018	.004	.067
RI	-.356	.085	-.047	.127	.009	.081	-.355	.078	-.061	.127	.010	.076

CD = complete data, DL = direct likelihood, SRI = simple random imputation, ARI = adjusted random imputation, IM = item means, TW = two-way means, CIM = corrected item means, RI = regression imputation

in almost all conditions, while the DL analysis results remained relatively stable.

## Discussion and conclusions

The aim of this study was to compare the performance of selected simple imputation methods adapted to a multiple-imputation approach versus a direct likelihood analysis approach when analyzing multilevel educational data having missing item scores with an explanatory IRT model. Therefore, the results obtained when analyzing a fully observed data set were compared to results from the same data when a specified number of item scores were considered missing either at random or completely at random and were ignored or estimated by an imputation method. Given that the fitted explanatory IRT model was specified correctly, a missing-data method was considered to perform effectively if the analysis results obtained from the reduced or imputed data were comparable to those obtained from analyses of the fully observed data sets.

For all conditions and compared to all imputation methods, the DL analysis approach produced unbiased fixed and slightly negatively biased random-effect variance parameter estimates, with *MSEs*, standard errors (and

standard deviations of the estimates), and confidence interval coverage probabilities comparable to those of a CD analysis. This might have been a result of the within-person correlation due to the repeated item scores, such that a moderate loss of (ignorable) scores does not necessarily lead to a substantial loss of predictive information. Additionally, and as noted by Beunckens et al. (2005), in DL analysis there is no distortion in the statistical information, since incomplete observations are neither removed (such as in the CCA analysis) nor added (such as in imputation methods).

Analyses of data sets whose missing values were imputed by the TW and CIM methods at a 15% proportion of missing item scores resulted in bias, standard errors, *MSE* values, and coverage probabilities of fixed-effect parameter estimates that were generally comparable to those of the CD analysis, though not as good as those of the DL analysis. The ability of the TW and CIM methods to perform better than other imputation methods could be attributed to the way in which they make imputations—that is, taking into account the relative performance of each person and correcting for item differences. These methods, however, tend to fail with large proportions of missing item scores, resulting in larger bias and *MSEs*. This outcome is in line with previous research. For instance, van Buuren

(2010) showed that the TW method tends to fail with large proportions of missing values set between 48% and 73%, and van Ginkel et al. (2007) showed that both TW and CIM methods performed quite well with 5% and 15% proportions of missing item scores. Analyses of data sets imputed by these two methods seemed to result in biased random-effect variance estimates (though they were smaller in magnitude when compared to those of other imputation methods). This highlights a difficulty faced by the considered imputation methods in accurately estimating missing observations in multilevel data sets. Indeed, the difficulties of multiple imputation with performing well in multilevel or clustered data settings have been noted with various existing multiple-imputation software packages (Yucel, 2008; Yucel, Schenker, & Raghunathan, 2006).

The SRI, ARI, and IM methods make imputations that result in biased fixed-parameter estimates with relatively small standard errors (and *MSE* values) in all conditions. This is a result of the way in which these methods substitute missing values. That is to say, due to missing scores, the set of observed scores from which to make random draws (for the SRI and ARI methods) or to compute item means (for the ARI and IM methods) becomes smaller, and tends to be uniform, thereby making the imputed item scores more similar (less variable) than they would have been, had item scores not been missing. The RI method does not perform well in most conditions (but not all), and it tends to be worse with higher proportions of missing items. One reason for this failure could be that a logistic regression model is not suitable for analyzing clustered data. Indeed, data structures in which observational units are clustered within groups (here, scores were clustered within persons) are in principle to be handled via multilevel analyses, as noted by Yucel (2008).

Standard errors for the CD and DL analyses, and those obtained from analyses of all imputed data sets, are larger when persons are clustered within schools, as compared to the case of no school clustering, which is an indication of increased uncertainty with clustering complexity. For instance, for all methods, standard errors in the no-school-clustering condition were comparable to their corresponding standard deviations of estimates, which was not the case in the school-clustering condition. This upward bias, however, might also be attributed to the small cluster sizes, as was highlighted for the mixed model by Cools, Van den Noortgate, and Onghena (2009). However, as the confidence interval coverage probabilities for all analyses with school or no school clustering were not very different, this did not seem to affect our conclusions.

The present article has verified that the DL analysis approach produces (almost) unbiased estimates of the fixed-effect parameters and random-effect variances of the explanatory IRT model, provided that the mechanism inducing the

missing item scores is ignorable (MCAR or MAR). This approach is further desirable because it can be easily implemented in most statistical standard software packages, with no additional programming involved (Beunckens et al. 2005), using the same model that would have been applied to a complete data set with no missing item scores. However, inference by multiple imputation may have some practical advantages over direct likelihood in some situations. For instance, Yucel (2008) noted that multiple imputation provides complete data sets for subsequent analyses, allowing analysts to use their favorite models and software. Molenberghs and Verbeke (2005) also discussed various situations in which multiple imputation might be preferred—say, in handling missing covariates when there is a combination of both missing covariates and missing outcomes—but these situations are not a focus of the present study.

We conclude that the direct likelihood analysis approach performs generally better than the considered imputation methods in the case of missing item scores in multilevel data sets. However, if there are reasons for using imputations, we recommend multiple imputation using the two-way mean and corrected-item-mean substitution methods, especially for low proportions of missing item scores. We advise against use of the simple random imputation, adjusted random imputation, and item or person mean imputation methods for multilevel data sets, based on our research.

**Author note** Kind acknowledgements to the Hercules Foundation and the Flemish Government—EWI department for funding the Flemish Supercomputer Centre (VSC), whose infrastructure we used to carry out simulations.

## References

- Arnab, R., & Singh, S. (2006). A new method for estimating variance from data imputed with ratio method of imputation. *Statistics & Probability Letters*, *76*, 513–519. doi:10.1016/j.spl.2005.08.019
- Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*, 5–37. doi:10.1016/j.jsp.2009.10.001
- Bates, D., & Maechler, M. (2010). lme4 0.999375-73: Linear mixed-effects models using Eigen and S4 classes [R package]. Available at <http://cran.r-project.org>
- Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, *35*, 321–364. doi:10.1207/S15327906MBR3503\_03
- Beunckens, C., Molenberghs, G., & Kenward, M. G. (2005). Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical Trials*, *2*, 379–386.
- Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*, 89–118. doi:10.1080/08957340801926086
- Chen, J., Rao, J. N. K., & Sitter, R. R. (2000). Efficient random imputation for missing data in complex surveys. *Statistica Sinica*, *10*, 1153–1170.

- Cools, W., Van den Noortgate, W., & Onghena, P. (2009). Design efficiency for imbalanced multilevel data. *Behavior Research Methods, 41*, 192–203. doi:10.3758/BRM.41.1.192
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., et al. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*, 1–28.
- De Boeck, P., Daems, F., Meulders, M., & Rymenans, R. (1997). *Ontwikkeling van een toets voor de eindtermen begrijpend lezen [Construction of a test for the educational targets of reading comprehension]*. Leuven/Antwerp, Belgium: Katholieke Universiteit Leuven/Universiteit Antwerpen.
- De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.
- Desmet, P., Paulussen, H., & Wylin, B. (2006). FRANEL: A public online language learning environment, based on broadcast material. In E. Pearson & P. Bohman (Eds.), *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 2307–2308). Chesapeake VA: AACE. Available at [www.edutlib.org/p/23329](http://www.edutlib.org/p/23329)
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.
- Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*, 225–245. doi:10.1111/j.1745-3984.2008.00062.x
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374. doi:10.1016/0001-6918(73)90003-6
- Hoffman, L., Yang, X., Bovaird, J. A., & Embretson, S. E. (2006). Measuring attentional ability in older adults: Development and psychometric evaluation of DriverScan. *Educational and Psychological Measurement, 66*, 984–1000. doi:10.1177/0013164406288170
- Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician, 61*, 79–90. doi:10.1198/000313007X172556
- Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality and Quantity, 34*, 331–351. doi:10.1023/A:1004782230065
- Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In A. Boomsma, M. A. J. V. Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 221–244). New York: Springer.
- Kalton, G., & Kasprzyk, D. (1982). Imputing for missing survey responses. *Proceedings of the Survey Research Methods Section*. American Statistical Association (pp. 22–31).
- Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79–93. doi:10.1111/j.1745-3984.2001.tb01117.x
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.
- Mallinckrodt, C. H., Clark, S. W. S., Carroll, R. J., & Molenberghs, G. (2003). Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *Journal of Biopharmaceutical Statistics, 13*, 179–190.
- Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies*. Hoboken, NJ: Wiley.
- Molenberghs, G., & Verbeke, G. (2004). An introduction to (generalized) (non)linear mixed models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 111–153). New York: Springer.
- Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. New York: Springer.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association, 81*, 366–374.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall/CRC.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177. doi:10.1037/1082-989X.7.2.147
- Shao, J. (2000). Cold deck and ratio imputation. *Survey Methodology, 26*, 79–86.
- Sheng, X., & Carrière, K. C. (2005). Strategies for analyzing missing item response data with an application to lung cancer. *Biometrical Journal, 47*, 605–615.
- Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research, 38*, 505–528. doi:10.1207/s15327906mbr3804\_4
- van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*, 219–242. doi:10.1177/0962280206074463
- van Buuren, S. (2010). Item imputation without specifying scale structure. *Methodology: A European Journal of Research Methods for the Behavioral and Social Sciences, 6*, 31–36. doi:10.1027/1614-2241/a000004
- van Buuren, S., & Oudshoorn, C. G. M. (2000). *Multivariate imputation by chained equations: MICE V1.0 user's manual*. TNO Prevention and Health, Public Health. Available at <http://web.inter.nl.net/users/S.van.Buuren/mi/docs/Manual.pdf>
- Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*, 369–386. doi:10.3102/10769986028004369
- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.
- van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research, 42*, 387–413. doi:10.1080/00273170701360803
- Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology, 38*, 369–397. doi:10.1111/j.1467-9531.2008.00202.x
- Wauters, K., Desmet, P., & Van den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning, 26*, 549–562. doi:10.1111/j.1365-2729.2010.00368.x
- Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A, 366*, 2389–2403. doi:10.1098/rsta.2008.0038
- Yucel, R. M., Schenker, N., & Raghunathan, T. E. (2006, October). *Multiple imputation for incomplete multilevel data with SHRIMP*. Paper presented at the annual conference on New Methods for the Analysis of Family and Dyadic Processes, Amherst, MA. Retrieved from [www.umass.edu/family/pdfs/talkyucel.pdf](http://www.umass.edu/family/pdfs/talkyucel.pdf)
- Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika, 56*, 589–600. doi:10.1007/BF02294492