Applied Psychological Measurement
http://apm.sagepub.com/

Direct Likelihood Analysis and Multiple Imputation for Missing Item Scores in Multilevel Cross-Classification Educational Data

Damazo T. Kadengye, Eva Ceulemans and Wim Van den Noortgate

Applied Psychological Measurement 2014 38: 61 originally published online 27 June 2013
DOI: 10.1177/0146621613491138

The online version of this article can be found at:
http://apm.sagepub.com/content/38/1/61

Published by:
SAGE
http://www.sagepublications.com

Additional services and information for *Applied Psychological Measurement* can be found at:

Email Alerts: http://apm.sagepub.com/cgi/alerts

Subscriptions: http://apm.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations: http://apm.sagepub.com/content/38/1/61.refs.html

>> Version of Record - Dec 8, 2013

OnlineFirst Version of Record - Jun 27, 2013

What is This?

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

Article

Applied Psychological Measurement
38(1) 61–80
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621613491138
apm.sagepub.com
SAGE

Direct Likelihood Analysis and Multiple Imputation for Missing Item Scores in Multilevel Cross-Classification Educational Data

Damazo T. Kadengye[1,2], Eva Ceulemans[2], and Wim Van den Noortgate[1,2]

Abstract

Multiple imputation (MI) has become a highly useful technique for handling missing values in many settings. In this article, the authors compare the performance of a MI model based on empirical Bayes techniques to a direct maximum likelihood analysis approach that is known to be robust in the presence of missing observations. Specifically, they focus on handling of missing item scores in multilevel cross-classification item response data structures that may require more complex imputation techniques, and for situations where an imputation model can be more general than the analysis model. Through a simulation study and an empirical example, the authors show that MI is more effective in estimating missing item scores and produces unbiased parameter estimates of explanatory item response theory models formulated as cross-classified mixed models.

Keywords

explanatory item response models, missing data, multiple imputation

Statistical models of data from educational research and other disciplines often have to deal with complex multilevel structures. In this article, the authors focus on item response theory (IRT) modeling of data sets, where the same set of exercise items are answered by several persons, resulting in a cross-classification of response scores at Level 1 within persons and items both at Level 2 (Beretvas, 2011; Fox, 2010; Van den Noortgate, De Boeck, & Meulders, 2003), and where items and/or persons belong to larger groups. A typical example of such a data structure is an educational testing and measurement study, where persons are repeatedly measured on the same sample set of items, with individual persons clustered in classes, schools, or a university's campus locations (Level 3). Such a structure is illustrated in Figure 1. The cross-classification is

[1]KU Leuven–Kulak, Kortrijk, Belgium
[2]KU Leuven, Belgium

Corresponding Author:
Damazo T. Kadengye, Faculty of Psychology and Educational Sciences, KU Leuven–Kulak, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium.
Email: trevor.kadengye@kulak.be

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

evident from the crossing of the lines connecting items to scores. Tabular illustrations of similar data structures are given, for instance, by Beretvas (2011) and Goldstein (2010). Although all Level 1 cells in the current example have an element within them, it is possible to have some empty cells in case of missing items scores by some persons. In such situations, caution must be taken when employing specific IRT models to avoid biased estimates.

IRT Model With Crossed Random Effects

In IRT modeling, we can consider the measurement occasions (Level 1) to be nested within persons (Level 2), and treat person parameters as random effects, and characteristics of the measurement occasions (i.e., what item was solved at that moment) as predictors in an IRT model. A model with dummy item indicators as predictors is equivalent to the marginal maximum likelihood (ML) procedure for the Rasch model (De Boeck & Wilson, 2004; Rasch, 1993). Similarly items can be considered as being repeatedly measured such that measurement occasions are nested within items and person dummy variables can be included in the IRT model as predictors. However, items and persons can be regarded as random effects and fit cross-classification multilevel logistic models (Van den Noortgate et al., 2003) for binary multilevel IRT data. A fully unconditional (or an empty) two-level cross-classification IRT model which takes into account the variation between items, between persons, and within persons and items is given as

$$\text{logit}\left(\pi_{pi}\right) = \alpha_0 + \omega_p + \varepsilon_i, \text{ and } Y_{pi} \sim \text{binomial}\left(1, \pi_{pi}\right), \tag{1}$$

where Y_{pi} denotes the binary score of person $p (p = 1, \ldots, P)$ to item $i (i = 1, \ldots, I)$ at measurement occasion m_{pi}, π_{pi} is the conditional success probability given the random effect for person p and item i, α_0 is the expected logit of the success probability for an average person on an average item, and ω_p and ε_i are random effects for persons with $\omega_p \sim N(0, \sigma^2_\omega)$ and for items with $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$ respectively. The person random effects from such a model are interpreted as ability levels of individuals and similarly, the item random effects can be interpreted as easiness levels of items.

The cross-classified IRT model in Equation 1 can be recognized as a generalized linear mixed model (GLMM) that can be extended to an explanatory IRT framework by including person- and item-specific predictors (De Boeck & Wilson, 2004). Level 1 characteristics can also be included in the GLMM, say, the time or place at which a measurement of a person on an item was taken. The explanatory IRT framework is very flexible as it can be used to build different IRT models to explain latent properties of items, persons, groups of items, or groups of persons. For instance, if schools, $k = 1, 2, \ldots, K$, in Figure 1 are thought of as being a random sample from a population of schools, such that school effects (u_ks) are normally distributed with $u_k \sim N(0, \sigma^2_u)$, then the resulting cross-classified multilevel explanatory IRT model including characteristics of items, persons, and measurement occasions can be formulated as follows:

$$\text{logit}\left(\pi_{kpi}\right) = \alpha_0 + \sum_{l=1}^{L} \gamma_l X_{il} + \sum_{q=1}^{Q} \lambda_q G_{pq} + \sum_{j=1}^{J} \beta_j m_{pij} + \omega_p + u_k + \varepsilon_i, \tag{2}$$

where π_{kpi} is the probability of success for person p from school k on item i, X_{il} is the value of predictor l for item i and G_{pq} is the value of predictor q for person p, m_{pij} is the value of predictor j for a characteristic of the measurement occasion on which item i was solved by person p, and β_j, γ_l, and λ_q are unknown fixed regression coefficients, respectively.

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

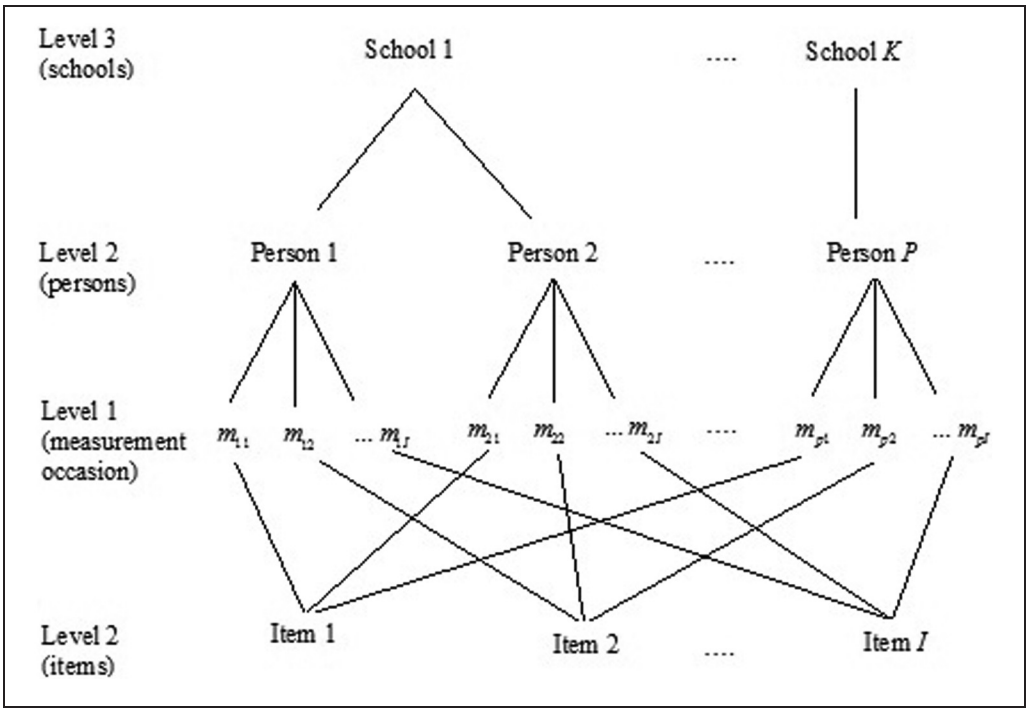

Figure 1. An illustration of a multilevel cross-classified data structure.

The model in Equation 2 includes several popular IRT models as special cases. For instance, the Linear Logistic Test Model (LLTM; Fischer, 1973) that is used to explain differences in item difficulty can be obtained, by dropping measurement predictors, person predictors and, the random item and school effects. De Boeck and Wilson (2004) showed that an interesting extension of the LLTM is an *item explanatory* IRT model that includes random item effects, referring to the part of the item difficulties that is not explained by the item predictors in the model. In the same way, differences in person ability can be explained by dropping measurement predictors, item predictors, and only including person predictors as fixed effects as in the case of a latent regression IRT model (Zwinderman, 1991) which De Boeck and Wilson call a *person explanatory* IRT model. By adding random person effects, the model does not assume that the person predictors explain all variance. For instance, Kamata (2001) has applied a latent regression model with person characteristic variables to analyze the effect of studying at home on science achievement. By dropping only the measurement predictors from Equation 2 and including person and item predictors, a *doubly explanatory* IRT model, in the terminology of De Boeck and Wilson, is obtained. Similarly, one can include only a Level 1 characteristic in the model or define several other IRT models from Equation 2. Moreover, Equation 2 can be extended to include meaningful group-by-item or person-by-measurement occasion random effects, though these are not considered further in the current study.

Presence of Missing Item Scores: Example of Electronic Learning Data

To build and improve their knowledge, persons can be taught in traditional classroom settings, via electronic learning (e-learning) environments, or through a combination of both. A specific

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

kind of an e-learning environment is an item-based web environment in which persons learn by freely engaging exercise items online, receiving automatic feedback after each response, and in which their answers are centrally logged. The Maths Garden (Klinkenberg, Straatemeier, & Van der Maas, 2011) and the FraNel project (Desmet, Paulussen, & Wylin, 2006) are good examples of such e-learning environments. Such web-based e-learning environments might contain more items than a typical user makes, resulting in relatively large numbers of missing item scores (Wauters, Desmet, & Van den Noortgate, 2010). In such environments, the place or time at which measurements are taken can be influencing factors: whether a person is motivated to engage items at school, at home, or other public place; at start or completion of the course; toward a high-stakes or a low-stakes test; and so on. For instance, it is likely that an average student will tend to be more active in engaging the exercise items toward a graded examination period than at the beginning of the course. High amounts of missing item scores can be a major challenge when using IRT models to obtain reliable parameter estimates, usually giving rise to nonconvergence problems or intractable likelihoods (Fox, 2010).

Therefore, during statistical analysis of data with missing scores, variables and mechanisms that lead to missing items need to be identified and corrected for, to avoid biased estimates (Little & Rubin, 2002). In particular, using the terminology of Little and Rubin (2002), item scores can be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). MCAR occurs 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). However, MCAR is a strong and unrealistic assumption in most studies because very often some degree of relationship exists between the missing values and some of the covariate data. It is often more reasonable to state that data are MAR, which occurs when a missing score depends only on some observed quantities that may include some outcomes and/or covariates. If neither MCAR nor MAR holds, then the missing scores that depend on the unobserved quantities and data are said to be MNAR. For instance, some students may not respond to certain items because they find them difficult.

These missing-data mechanisms need to be taken into account during statistical modeling to be able to produce estimates that are unbiased and efficient, although no modeling approach, whether for MCAR, 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 score 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 case of a nonignorable missing item response arising from the 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 (cf. Molenberghs & Kenward, 2007). However, the causes of an MNAR mechanism might be unknown and the methods to implement its remediation are complex and beyond the scope of this study. Thus, MNAR is not further considered in the current study.

Handling of Missing Item Scores

Direct Likelihood (DL) Analysis

Modern statistical software for fitting cross-classification multilevel IRT models (e.g., SAS procedure GLIMMIX, SAS Institute Inc, 2005, and the *lmer* function of R package *lme4*, D. Bates & Maechler, 2010) are generally equipped with a full information ML analysis approach. An interesting feature of this approach, also known as *DL analysis* (Beunckens et al., 2005; Mallinckrodt, Clark, Carroll, & Molenberghs, 2003; Molenberghs & Kenward, 2007), is that it

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

automatically handles missing responses under the assumption of MAR. Under the DL approach, all the available observed data are analyzed without deletion or 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 doing so, appropriate adjustments valid under the ignorable missing mechanism are made to parameters, even when data are incomplete, due to the within-person correlation (Beunckens et al., 2005). An extensive description of the DL approach as a method for analyzing clustered data with missing observations is given in Molenberghs and Kenward (2007). In the next paragraph, a brief description of the approach is given in the context of the IRT model described in Equation 2.

The explanatory IRT models belong to the framework of GLMMs (De Boeck & Wilson, 2004) with persons as clusters and items for the repeated observations. The GLMM for observed binary item scores Y^o_{kpi} is given as $\text{logit}(\pi_{kpi}) = \eta_{kpi}$ and $Y^o_{kpi} \sim \text{binomial}(1, \pi_{kpi})$. For each combination of a person p from school k and an item i, the value of the component η_{kpi} is determined by a linear combination of the person and item predictors as given in Equation 2. 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. The DL analysis is not a new approach for handling missing item response data. For instance, in a recent study by Kadengye, Cools, Ceulemans, and Van den Noortgate (2012), the DL analysis was shown to perform quite well for up to 50% proportions of missing scores in multilevel IRT data structures similar to those considered in the current study.

Despite the attractiveness of the DL analysis, van Buuren (2011) advised that this approach should be used with some care when data are incomplete as it is only preferred to handle missing data if certain conditions hold. For instance, the standard DL analysis approach assumes that missing item scores are MAR. As such, the analysis model must include both the predictors of interest and all other covariates that may be related to the missingness process. But this can be suspect in some explanatory IRT models. For example, it can easily be deduced that the described *person, item*, and *doubly* explanatory IRT models are in fact distinct GLMMs—all nested in the general IRT model of Equation 2. Let us, for instance, revisit the e-learning environment example where students learn by freely engaging exercises online anywhere at anyplace. Suppose that items scores are missing as a result of differences in a Level 1 property (e.g., some persons may be motivated to engage exercise items in the afternoon than in the morning). Then fitting the item, the person, or the doubly explanatory IRT model may result in biased parameter estimates if the Level 1 property that is related to the missingness process is not included in the model, because this violates the MAR assumption. Yet, the requirement of including all covariates related to the missingness process has at least two challenges for the analyst. First, an analyst is forced to use a model that is not necessarily of interest. Second, the so-called complete model may be more complex to interpret (van Buuren, 2011) given the research objective of the analyst, or worse still, impossible to estimate given the available tools.

Imputation

A different approach to dealing with missing item scores in IRT data structures is by replacing the missing scores with simulated values—a process called imputation. Several imputation methods have been developed to deal with missing item scores in the IRT context. Most of these methods, like simple forms of mean imputation for the missing item scores, have been widely applied and examined in traditional IRT models (cf. Finch, 2008; Sheng & Carrière, 2005; van Ginkel, van der Ark, Sijtsma, & Vermunt, 2007). However, not very much has been done about missing responses in the area of multilevel IRT models with crossed random effects.

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

Only recently, the performance of several imputation methods were compared in explanatory IRT models for multilevel educational data (Kadengye et al., 2012). The authors noted, however, that simple mean imputation approaches failed in clustered data structures and resulted in severely underestimated variability of the imputed data. This observation is in line with literature on missing data where simple forms of imputation are discouraged for multilevel data sets (cf. Little & Rubin, 2002; Schafer, 1997; van Buuren, 2011).

Imputation can take at least two different forms: single-value versus multiple-value. In single-value imputation, a missing score is replaced by one estimated value according to an imputation method of choice. However, Baraldi and Enders (2010) illustrated that several single-value imputation methods introduced bias in the estimates of means, variances, and correlations of a fictitious mathematics performance data set. The main problem is that inferences about parameters based on the filled-in single values do not account for imputation uncertainty, because variability induced by 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 and the Type I error rate of significance tests may be inflated (Baraldi & Enders, 2010; Little & Rubin, 2002; Schafer & Graham, 2002).

The multiple imputation (MI; Little & Rubin, 2002; Rubin, 1987) approach, however, involves replacing each missing item score by a set of more than one synthetic values generated according to an appropriate statistical model. However, some statistical programs that offer modules to implement MI for missing binary data, such as PROC MI (SAS Institute Inc, 2005) and R's package MICE (van Buuren & Oudshoorn, 2000), generally ignore the clustering structure in multilevel data. Indeed, difficulties of some MI procedures to perform well in clustered data settings have been noted for various existing MI software (Yucel, 2008). To the best of knowledge, appropriate MI models for clustered data structures like those considered in the current study and the IRT models that are focused on—the cross-classified mixed models—have not been explored in the IRT literature. For an educational researcher with little statistical knowledge, it is not very clear as to which imputation model one can employ for such problems. This leaves such researchers at the mercy of existing MI software that are not appropriate for the data structures focused on, or perhaps to blindly employ DL analyses that may not be appropriate in some situations.

In this article, therefore, the authors compare the performance of an appropriate MI model based on empirical Bayes techniques versus a DL analysis for cross-classified multilevel IRT data. Generally speaking, an imputation model does not have to be similar to that of the final analysis, as the former can be allowed to contain structures or covariates that the latter model does not (Schafer, 2003; Yucel, 2008). More specifically, the focus is on situations where the variables that are of importance to understand missingness may not be of scientific interest to the analyst such that the MI model is more general than is any possible subsequent IRT analysis model. In the remainder of the article, the authors discuss the MI approach and an appropriate imputation model for missing item scores in cross-classified multilevel IRT data. Then through a simulation study, they compare results of a DL analysis and those obtained from multiply imputed data, in the situation where imputation and analysis models are similar, and where they are different. Next, they apply both approaches to an empirical example using different values for the number of imputations. The article concludes with a brief discussion.

MI Procedure

Plausible values for the missing item scores can be generated from the observed data by using MI procedures (Little & Rubin, 2002; Rubin, 1987). A model is used to generate $M > 1$ imputations for the missing observations in the data by making random draws from a distribution of

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

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 results of the M analyses will probably differ slightly because the imputations differ. Following Rubin's rules (e.g., Little & Rubin, 2002), results of the M analyses are combined into a single estimate of the statistics of interest by combining the variation within and across the M imputed data sets. As such, uncertainty about the imputed values is taken into account and under fairly liberal conditions, statistically valid estimates with unbiased standard errors can be obtained.

Different models can be adopted in the MI approach, 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 would be an appropriate imputation model (Schafer, 1997). 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), making it feasible to set up and process a full multiway cross-tabulation as required for the log-linear analysis. Alternatively, 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 is specified. Missing binary scores can be predicted with a logistic regression model of the observed scores on observed auxiliary variables with possible higher order interaction terms or on transformed variables. Such a logistic model can easily be implemented in MI modules like MICE (van Buuren & Oudshoorn, 2000) of R software.

However, a simple logistic regression model may not be suitable for estimating missing item scores in multilevel data sets. For instance, implementation of this model under MICE by Kadengye et al. (2012) resulted in biased estimates of the fixed and variance parameters for IRT models similar to the ones considered in the current study. To create good imputed data sets, it is desirable to reflect the data design in the model employed for creating the MIs (Rubin, 1987), as well as all other covariates available in the data set. Schafer (2003), however, cautions against making the imputation model unnecessarily large and complicated. For instance, he warns that in case some covariates are unrelated to the outcome variable but are included in the imputation procedure, precision of the analysis estimates may be reduced. Rather, a good imputation model should contain variables and structures that are correlated either with both the outcome and the missingness process or one of the two. For multilevel data, van Buuren (2011) described a MI Bayesian approach using a linear mixed-effects model such that the distribution of the parameters can be simulated by Markov chain Monte Carlo (MCMC) methods. However, full Bayesian and MCMC methods are beyond the scope of the current study. Instead, the authors confront the problem at hand by taking advantage of the asymptotic distributions of the parameters estimated from the IRT model fitted to the observed data.

Empirical Bayes Imputation Model

Let $\mathbf{Y}$ be the vector of observed binary scores (Y^o) containing missing data (Y^m) such that $\mathbf{Y} = [Y^o, Y^m]$ and assume that the MAR assumption is plausible. For simplicity, let's consider a model with crossed random item and person effects only, fitted to the observed scores Y^o such that $\text{logit}(\pi_{pi}) = \boldsymbol{\eta}_{pi} = \mathbf{x}'_{pi}\boldsymbol{\beta} + \mathbf{z}'_{pi}\boldsymbol{\nu}_{pi}$ and $Y^o_{pi} \sim \text{binomial}(1, \pi_{pi})$. The $\mathbf{x_{pi}}$ is the vector of regressors for unit (p, i) with fixed effects $\boldsymbol{\beta}$, and z_{pi} is the vector of variables having the estimated random effects, denoted $\boldsymbol{\nu}_{pi}$. The random effects are usually assumed to be independently and multivariate normally distributed as $\boldsymbol{\nu}_{pi} \sim N(0, \boldsymbol{\Omega}_\nu)$. The logit can be estimated by using estimated values of the fixed effects parameters ($\hat{\boldsymbol{\beta}}$), and the estimates of random effects ($\hat{\boldsymbol{\nu}}$) that are obtained directly from the fitted model by their conditional expected values using empirical Bayes methods. Imputations for the missing data Y^m are then generated from $(Y^m | Y^o, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\Omega}})$. Specifically, imputations are obtained by successively drawing plausible estimates from the estimated

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

asymptotic sampling distributions of individual person and item random parameters to account for parameter uncertainty, and successively combining these to the set of fixed linear predictors, to obtain plausible probabilities of success for each person–item combination. Let $\hat{\pi}^*_{pi}$ denote an estimate of the probability of a correct response of person p on item i such that

$$\hat{\pi}^*_{pi} = \frac{\exp\left(\hat{\eta}^*_{pi}\right)}{1+\exp\left(\hat{\eta}^*_{pi}\right)} \text{ and } \hat{\eta}^*_{pi} = \mathbf{x}'_{pi}\hat{\boldsymbol{\beta}} + \hat{\omega}^*_p + \hat{\varepsilon}^*_i, \tag{3}$$

with $\hat{\omega}^*_p \sim N(\hat{\omega}_p, \hat{\tau}^2_{\hat{\omega}_p})$ and $\hat{\varepsilon}^*_i \sim N(\hat{\varepsilon}_i, \hat{\tau}^2_{\hat{\varepsilon}_i})$ as the asymptotic distributions of the random effects estimates, where $\hat{\tau}^2_{\hat{\omega}_p}$ and $\hat{\tau}^2_{\hat{\varepsilon}_i}$ are respective squared estimated standard errors for the estimated person and item random effects. In this way, M values of $\hat{\pi}^*_{pi}$ are generated each being a plausible probability of success for a missing score Y^m_{pi}. Each $\hat{\pi}^*_{pi}$ is used to draw a 1 or 0 for the missing score from $Y^m_{pi} \sim B(1, \hat{\pi}^*_{pi})$, and Rubin's (1987) rules are then employed to make a single inference. By sampling repeatedly a sufficient number of times from respective distributions, the imputation model in Equation 3 reflects the uncertainty about the random effects' parameter estimates as well as the uncertainty about the imputed values. The described procedure is specific to missing binary item scores in the outcome variable of a simple cross-classification IRT data structure but can be easily extended to the multilevel structure of Equation 2.

Simulation Study

Through a simulation study, we studied the performance of the DL analysis and MI approach using the described imputation model for three explanatory IRT models: an item covariate model, a Level 1 covariate model, and a model with both covariates. The latter model is for a situation where the imputation and analysis models are similar. The models were fitted to data simulated under two situations, namely, a simple cross-classified IRT structure (Study I) and multilevel cross-classified IRT structure (Study II).

Simulating Data

Item response data were generated according to the model $\text{logit}(\pi_{kpi}) = \beta_0 + \beta_1 x_i + \beta_2 m_{pi} + \omega_p + u_k + \varepsilon_i$ with $\omega_p \sim N(0, \sigma^2_\omega), u_k \sim N(0, \sigma^2_u)$ and $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$. Population values assigned to respective parameters are given in the next subsections. A response score on item i by person p from school k (Y_{kpi}), was generated from a binomial distribution as $Y_{kpi} \sim B(1, \pi_{kpi})$. In each of the two studies, the following factors were considered for data simulation.

Clustering. For Study I, a simple crossed-classified structure was considered where item difficulties were assumed to be generated from an item bank with a standard normal distribution, $\varepsilon \sim N(0, 1)$, and abilities of persons also from a standard normal distribution, $\omega \sim N(0, 1)$, such that $\sigma^2_u = 0$ (no school clustering). For Study II, this structure was extended such that persons are nested in $K = 20$ schools to examine the effect of the multilevel structure on imputations. As in the simulation study of Kadengye et al. (2012), the authors assumed that 30% of the differences in scores was situated between schools. The total variance between persons is the same as in Study I, such that $u \sim N(0, .3)$ and $\omega \sim N(0, .7)$.

Sample size. Two item sample sizes were compared, that is, $I = 40$ and $I = 80$ while the persons' sample size was fixed at $P = 200$. For Study II, the persons are equally distributed over the schools, that is, 10 persons per school.

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

Distribution of covariates. In both simulation studies, the two covariates are assumed to be binary variables. The scores on these variables were drawn from a Bernoulli distribution with a probability parameter of .5 simply to obtain a generally balanced design.

Effect size of fixed effects. For both studies, a parameter value of $\beta_1 = \beta_2 = .5$ was assumed to be predictive of the outcome. According to Cohen's (1988) rules of thumb for effect sizes, the value of .5 is a ''medium'' effect which is considered to be big enough to be related with the outcome. The value of β_0 is set to 0 because it's not of much interest in the current study.

Missing proportions (ζ). Each sample data set was used twice: a selection of 25% and then 50% of the item scores were removed, yielding moderate and high proportions of missing scores. These values were based on the authors' experience with e-learning environments (because this is the example that is used) that tend to have high levels of nonresponse.

Missing mechanism. For all data samples, missingness at random (MAR) was induced in item scores. To induce the two missing proportions given the MAR process, probabilities of responding $P(\text{resp}_{pi} = 1)$ were dependent on the Level 1 covariate m. Person and item combinations with $m = 1$ were assigned higher values of $P(\text{resp}_{pi} = 1)$ than their counterparts: 95% versus 55% for $\zeta = 25\%$ and 70% versus 30% for $\zeta = 50\%$. Missing mechanisms MCAR and MNAR were not simulated for reasons already discussed in the section titled ''Presence of Missing Item Scores: Example of Electronic Learning Data.''

Analysis Procedure

For both simulation studies, the authors assumed that an analyst may be interested in one of three explanatory IRT models: an item covariate model, a Level 1 covariate model, or a model that contains both covariates. If the Level 1 property is not included in the analysis model (for the item covariate IRT model), we expect increased bias in the fixed effects estimates for the DL analysis, because by excluding this covariate, the MAR process is violated as this covariate is associated with the missing item scores. For the other two models, we expect that the DL analysis and the MI approach to have more or less similar results and that bias will be minimized because the analysis model corrects for the underlying missing process. For the models where one of the predictors is dropped, the expected value for the intercept is equal to the average of the expected values for the two conditions distinguished based on the ''population'' parameter values of the intercept and the particular covariate not in the model. In this case, this value is $\beta_0^* = .25$, because in the reference condition, the expected value is equal to the intercept β_0 which is 0, while in the other condition the expected value is equal to .5, this is $\beta_0 + \beta_1$ or $\beta_0 + \beta_2$, respectively, with $\beta_1 = \beta_2 = .5$.

Imputations for the missing item scores were generated using the model for simulating the data. Under MI, the basic rule of thumb is that the number of imputations is usually set to $3 \leq M \leq 5$ to get sufficient inferential accuracy. According to Rubin (1987), the efficiency of an estimate based on M imputations relative to one based on an infinite number of imputations is approximately $(1 + (\zeta/M))^{-1}$, where ζ is the fraction of missing information for the quantity being estimated. With a moderate value of ζ, for instance, $\zeta = 0.15$, an estimate based on $M = 3$, $M = 5$, and $M = 10$ will have a relative efficiency of 95.2%, 97.1%, and 98.5%, respectively. The respective standard errors of the estimate are 1.02, 1.01, 1.01 times as large as the estimate with $M = \infty$ (cf. Schafer, 1997). As such, it would appear that the resulting gain in efficiency is typically unimportant because the Monte Carlo error is a relatively small portion of the overall inferential uncertainty and that the additional resources that would be required to create and store more than a few imputations would not be well spent. Moreover, Molenberghs and

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

Kenward (2007) showed that efficiency gains rapidly diminish after the first $M = 2$ imputations for small values of ζ, and after the first $M = 5$ for large values of ζ. But Schafer (1997) argued that, with high rates of missing data, a larger number of imputations may be necessary to guarantee stable parameter estimates and better efficiency. For instance, when $\zeta = 0.30$, an estimate based on $M = 3$, $M = 5$, and $M = 10$ will have a relative efficiency of 90.9%, 94.3%, and 97.1% with respective standard errors that are 1.05, 1.03, 1.01 times as large as the estimate with $M = \infty$. Hence, as relatively large values of ζ are considered, $M = 10$ was used in both simulation studies.

In the first study, analyses involved 4 conditions: 2 (item sample sizes) × 2 (missing proportions). In the second study, only two conditions: 1 (item sample size) × 2 (missing proportions) were considered because analyses of Study II are used for comparison purposes. In all, 500 data sets were simulated for each of the six conditions. Each data set was analyzed 9 times: the three models are fitted to the completely observed data, to the incomplete data, and to the data whose missing item scores have been multiply imputed according to Equation 3, resulting in 27,000 analyses. The purpose of analyzing the completely observed data sets (denoted CD) was to compare the performance of the MI procedures with the performance of the IRT analysis in the ideal situation, this is, when there is no missingness. Note that the number of conditions, and specifically the number of replications was kept relatively small due to estimation time considerations. For instance, the 500 replications of only one condition (e.g., $P = 200$, $I = 40$, and $\zeta = 25\%$) took approximately 27 hr to estimate all parameters of interest on a computer with a CPU of Intel Core i5 with 2.67 GHz and RAM 4GB, and this time almost doubles for conditions with $I = 80$.

Outcomes of interest are the accuracy of the fixed regression weights' estimates, of the estimated standard errors of these fixed effects, and of the estimates for random effects' variance over persons (σ^2_ω), items (σ^2_ε), and schools (σ^2_u). Therefore, we look at the bias, precision, and mean squared error (MSE) of these estimates. Bias is the difference between an estimator's expected value and the population value of the parameter being estimated. Precision is the inverse of the sampling variance, and MSE refers to the expected squared deviation of the estimates from the true value of the population parameter, and is given by the sum of the variance and the squared bias of an estimator. Fixed effects' expected values for the estimates are estimated within a condition by the mean of the 500 estimates. For the variance of random effects, the median of the estimates is used rather than the mean estimate to evaluate bias, because variance estimates are bounded above zero and therefore positively skewed. Standard errors' estimates were evaluated by comparing within each condition the mean estimated standard error (denoted by Mean *SE*) of each parameter to the standard deviation (denoted by *SD*) of the 500 estimates for each parameter, and by looking at the coverage proportion of the respective confidence intervals.

Model fitting was done in R statistical software version 2.12.1, as described by De Boeck et al. (2011). For each simulated data set, model parameters were estimated using the lmer() function in the package lme4 (D. Bates & Maechler, 2010). The Laplace approximation to the likelihood is optimized to obtain approximate values of the ML or restricted maximum likelihood (REML) estimates for the parameters of interest and for the corresponding conditional modes of the random effects (Doran, Bates, Bliese, & Dowling, 2007). In random effects models, the REML estimates of variance components are generally preferred because they are usually less biased than ML estimates (D. M. Bates, 2008; Sarkar, 2008). For the MI procedure, parameter estimation and the missing-data imputation were conducted alternatively. That is, in the DL analysis, a general GLMM in Equation 3 (for Study I) or the one with an additional random school effect (for Study II) was fitted to the observed data with incomplete scores. Obtained individual person and item random REML parameter estimates were then used to successively draw $M = 10$ plausible estimates from their asymptotic sampling distributions. These

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

Table 1. Bias and MSE in Fixed Effects' Estimates by Fitted Model, Missing Proportion, Analysis Approach, and Study.

		Model 1				Model 2				Model 3					
		Bias		MSE		Bias		MSE		Bias			MSE		
ζ	Method	β_0^*	β_1	β_0^*	β_1	β_0^*	β_2	β_0^*	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Study I, 40 items															
25	CD	1	3	60	105	6	2	33	3	3	2	10	63	3	107
	DL	67	4	66	108	6	2	34	4	3	2	11	65	4	111
	MI	–4	0	60	106	–4	8	33	4	–5	8	6	64	4	109
50	DL	101	5	72	110	7	1	36	7	4	1	11	65	7	112
	MI	–9	–9	59	105	–13	9	34	7	–10	9	–2	62	7	107
Study I, 80 items															
25	CD	–11	7	29	53	0	–1	19	2	–8	–1	13	30	2	55
	DL	55	7	33	54	–1	0	20	2	–8	0	13	31	2	55
	MI	–14	3	29	53	–7	4	20	2	–13	4	9	31	2	54
50	DL	88	10	38	53	0	0	22	4	–9	0	16	33	4	55
	MI	–20	1	30	52	–16	6	22	4	–20	6	7	32	4	53
Study II, 80 items															
25	CD	13	–25	45	53	6	–1	33	1	17	–1	–19	47	1	54
	DL	80	–26	52	53	6	–2	34	2	18	–2	–21	48	2	54
	MI	11	–32	45	53	–2	3	33	2	12	3	–26	47	2	54
50	DL	115	–26	59	53	5	3	34	3	16	3	–21	48	3	53
	MI	6	–38	43	51	–12	8	33	3	6	8	–32	45	3	52

Note: $\beta_0^* = .25, \beta_0 = .00, \beta_1 = .50, \beta_2 = .50$. All entries to be multiplied by 10^{-3}. Model 1 = item covariate model; Model 2 = Level 1 covariate model; Model 3 = Level 1 + item covariate model; MSE = mean squared error; CD = complete data; DL = direct likelihood; MI = multiple imputation.

were then successively combined to the set of corresponding fixed linear predictor estimates, to obtain plausible probabilities of success for each person–item combination—the $\hat{\pi}_{pi}^*$s. Next, $M = 10$ synthetic complete data sets were generated by sampling the missing scores from Bernoulli$(1, \hat{\pi}_{pi}^*)$. Each ''completed'' data set was analyzed afresh using an explanatory IRT model of interest. For each IRT model of interest, $M = 10$ sets of estimates are obtained and combined using Rubin's (1987) rules for MI to make a single inference.

Results

Fixed Effects' Parameter Estimates

Estimated bias. Table 1 shows the bias and MSE of the fixed effects estimates by fitted model, missing proportion, analysis approach, and item sample size. For the item covariate model (Model 1, Table 1), the bias in estimates of the intercept parameter from the DL analyses is larger and not comparable with that of the CD analysis for all considered conditions. This bias increases when more item scores are missing. This is not the case when analyzing the multiply imputed data sets, where estimated bias in all parameters and across all conditions is generally comparable with that of the CD analysis. However and for the Level 1 covariate model across the two sample sizes (Model 2, Table 1), estimated bias from the DL analysis and MI approach is small and comparable with that obtained from the CD analysis, especially at 25% missing proportion. At 50% missing proportion, however, estimated bias in the intercept parameter for MI seems to be slightly larger than that of the DL analysis. Similar results are found for the

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

Table 2. Mean Standard Errors and Standard Deviations of Fixed Effects' Estimates by Fitted Model, Missing Proportion, Analysis Approach, and Study.

		Model 1				Model 2				Model 3					
		M SE		*SD*		*M SE*		*SD*		*M SE*			*SD*		
ζ	Method	β_0^*	β_1	β_0^*	β_1	β_0^*	β_2	β_0^*	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Study I, 40 items															
25	CD	230	312	246	324	179	54	181	54	234	54	316	251	54	328
	DL	231	314	246	329	183	64	185	61	237	64	318	254	61	333
	MI	229	311	246	326	181	64	183	62	235	64	315	252	62	330
50	DL	233	316	249	332	188	83	189	84	242	83	320	256	84	335
	MI	229	310	243	323	183	70	184	85	236	70	314	249	85	327
Study I, 80 items															
25	CD	172	223	171	231	137	38	140	39	175	38	226	174	39	234
	DL	172	224	172	232	139	45	142	47	177	45	227	176	47	234
	MI	172	223	171	230	139	45	140	47	176	45	226	175	47	233
50	DL	175	227	175	231	143	59	149	60	181	59	229	182	60	233
	MI	173	224	172	227	140	49	146	61	178	49	227	179	61	230
Study II, 80 items															
25	CD	205	223	212	228	177	38	182	38	208	38	226	216	38	231
	DL	205	224	213	230	179	45	183	44	210	45	226	218	44	232
	MI	204	222	211	227	180	51	181	46	210	51	225	216	46	230
50	DL	206	226	214	228	181	59	185	57	212	59	228	218	57	230
	MI	207	222	208	223	185	55	180	57	214	55	225	212	57	226

Note: $\beta_0^* = .25$, $\beta_0 = .00$, $\beta_1 = .50$, $\beta_2 = .50$. All entries to be multiplied by 10^{-3}. Model 1 = item covariate model; Model 2 = Level 1 covariate model; Model 3 = Level 1 + item covariate model; CD = complete data; DL = direct likelihood; MI = multiple imputation.

analysis models that include both covariates (Model 3, Table 1). Because the covariate related to missing scores is included in the DL analysis, and as such the MAR mechanism is not violated, both the DL analysis and MI approach have comparable biases.

MSE. For the item covariate model (Model 1, Table 1), the MSE values for the intercept when using the DL approach are larger than those of the CD analysis for both missing proportions. However, the MSE values for all the parameter estimates of the item covariate model using the MI approach are comparable with those of the CD analysis for all the considered conditions. Indeed, for these estimates, the model used to impute missing item scores corrected for the covariate related to the missing scores—much as it was not of interest to the analyst. At 25% missing proportions, results for Models 2 and 3 (Table 1) indicate that the MSE in all parameter estimates for the DL and MI approaches are comparable with those given by the CD analysis. However, the MSE for the DL analysis tends to increase for the smaller item samples in which 50% of the scores are missing which is not the case for the MI approach.

Standard errors. Table 2 shows the mean of the standard errors (mean *SE*) and standard deviations (*SD*) of the fixed effects estimates by fitted model, missing proportion, analysis approach, and item sample size. For all analysis approaches and in all conditions but especially for the smaller item samples, the *SD* of the estimates were slightly underapproximated by the estimated mean *SE*. At 50% missing proportions, the estimated mean *SE* and the respective *SD* of the estimates tend to get slightly larger both for the DL and MI approaches, but this is more pronounced in the DL approach. This is expected due to the increased uncertainty as a result of

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

Table 3. Coverage (in Percentage) of the True Fixed Effects Values by the 95% Confidence Interval by Fitted Model, Missing Proportion, Analysis Method, and Study.

		Model 1		Model 2		Model 3		
ζ	Method	β_0^*	β_1	β_0^*	β_2	β_0	β_1	β_2
Study I, 40 items								
25	CD	.95	.94	.94	.95	.94	.95	.93
	DL	.92	.92	.94	.96	.94	.96	.92
	MI	.94	.92	.94	.95	.94	.95	.92
50	DL	.90	.93	.95	.95	.94	.95	.93
	MI	.95	.93	.95	.92	.94	.92	.93
Study I, 80 items								
25	CD	.95	.94	.94	.96	.94	.96	.94
	DL	.94	.94	.95	.94	.94	.94	.94
	MI	.95	.95	.95	.94	.94	.94	.95
50	DL	.93	.94	.93	.95	.94	.95	.94
	MI	.94	.94	.94	.92	.95	.92	.94
Study II, 80 items								
25	CD	.93	.94	.94	.96	.94	.96	.95
	DL	.91	.95	.94	.96	.93	.96	.95
	MI	.93	.95	.95	.96	.93	.96	.95
50	DL	.90	.96	.94	.96	.95	.96	.96
	MI	.95	.95	.95	.96	.95	.94	.96

Note: $\beta_0^* = .25$, $\beta_0 = .00$, $\beta_1 = .50$, $\beta_2 = .50$. Model 1 = item covariate model; Model 2 = Level 1 covariate model; Model 3 = Level 1 + item covariate model; CD = complete data; DL = direct likelihood; MI = multiple imputation.

large amounts of missing scores and hence the decreased number of scores on which estimates are based. In general, for all models, there was little variation for both DL and MI approaches in the difference between the mean standard errors and their corresponding standard deviations.

Coverage percentages. Coverage percentages of the true fixed effects values by the 95% confidence intervals are shown in Table 3. Good confidence intervals should have coverage probability equal (or close) to the nominal level of 95%. For the item covariate model (Model 1, Table 3), the coverage percentages for the intercept parameter when using the DL approach are generally smaller than the nominal value. This is not the case with the MI approach for this particular analysis model. In the other two analysis models (Models 2 and 3 of Table 3) and at 25% missing, the coverage percentages are comparable with those of the CD analysis. The MI approach tends to have coverage percentages slightly below the nominal value at 50% missing for the Level 1 covariate parameter estimate.

Random Effects' Variance Estimates

Estimated bias. Table 4 shows the bias and the MSE values of random effects' variance estimates by fitted model, missing proportion, analysis approach, and item sample size. For all analysis models and for both DL and MI approaches, it can be seen that the bias of all random effects' variance estimates increases when more data are missing. However, now, the bias seems to be slightly larger for the MI approach, but still in the same direction as that of the DL analysis. This is not unexpected because by replacing missing scores with imputed values sampled from the estimates of individual persons and items, the overall variability in the data

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

Table 4. Bias and MSE in Random-Effect Variance Estimates by Fitted Model, Missing Proportion, and Analysis Approach.

		Model 1						Model 2						Model 3					
		Bias			MSE			Bias			MSE			Bias			MSE		
ζ	Method	σ^2_ω	σ^2_ε	σ^2_u	σ^2_ω	σ^2_ε	σ^2_u	σ^2_ω	σ^2_ε	σ^2_u	σ^2_ω	σ^2_ε	σ^2_u	σ^2_ω	σ^2_ε	σ^2_u	σ^2_ω	σ^2_ε	σ^2_u
Study I, 40 items and $\sigma^2_\omega = 1, \sigma^2_\varepsilon = 1$																			
25	CD	−34	−78		15	53		−8	32		14	64		−9	−58		14	53	
	DL	−41	−84		17	54		−19	29		16	66		−19	−57		16	54	
	MI	−58	−104		18	55		−33	14		17	62		−33	−77		17	53	
50	DL	−51	−84		21	57		−26	25		21	69		−26	−58		21	57	
	MI	−82	−125		24	60		−59	−15		22	60		−59	−100		22	57	
Study I, 80 items																			
25	CD	−34	−62		13	30		−9	51		12	37		−9	−40		12	30	
	DL	−33	−55		14	31		−10	46		14	38		−10	−34		14	31	
	MI	−47	−72		14	32		−23	34		14	36		−23	−46		14	31	
50	DL	−38	−52		17	34		−17	48		16	42		−17	−34		16	34	
	MI	−67	−83		18	34		−46	16		17	36		−46	−62		17	33	
Study II, 80 items and $\sigma^2_\omega = .7, \sigma^2_u = .3, \sigma^2_\varepsilon = 1$																			
25	CD	−18	−52	−32	7	30	14	0	42	−25	7	36	14	0	−24	−26	7	30	14
	DL	−17	−54	−34	7	31	14	−2	37	−28	8	36	15	−2	−30	−28	8	30	15
	MI	−28	−72	−37	8	31	14	−11	18	−30	7	33	14	−11	−50	−30	7	30	14
50	DL	−18	−54	−33	9	33	14	−2	37	−27	9	37	15	−2	−29	−28	9	32	15
	MI	−32	−96	−34	9	36	14	−13	−1	−26	9	33	14	−13	−73	−26	9	33	14

Note: All entries to be multiplied by 10^{-3}. Model 1 = item covariate model; Model 2 = Level 1 covariate model; Model 3 = Level 1 + item covariate model; MSE = mean squared error; CD = complete data; DL = direct likelihood; MI = multiple imputation.

set will decrease. This is probably also due to the shrinkage effect associated with empirical Bayes estimates from which imputations were drawn.

MSE. Table 4 shows that the obtained MSE values from both DL and MI approaches for all random effects' variance estimates are comparable with those of the CD analysis for almost all considered conditions. The MSE values of the variance estimates for item difficulties are only slightly larger at 50% missing data. In general and compared with the CD analysis, there was a similar but slight loss of prediction accuracy for both the DL and MI analysis approaches in all the considered conditions.

Example

Assuming the MAR mechanism, a tracking-and-logging data set with a relatively large proportion of missing item scores (57%) was used to compare the DL and MI approaches. MI is implemented with different values of the number of imputations, that is, for $M = 1, 3, 5, 10$, and 20 to illustrate the effect of M on the estimates and standard errors of parameters of interest.

The analyzed data set was obtained from an item-based e-learning environment which was designed for students of the bachelor programs of Educational Studies, Educational Sciences, and Speech Therapy and Audiology Sciences of the University of Leuven in Belgium and offers a regression analysis module. The environment's usage is voluntary much as students are encouraged and motivated by the authors to independently engage the exercises online at different times of their choice outside the traditional classroom-based lecture times. Each student can engage an item only once and item scores are binary coded (correct/wrong). Eighty multiple-choice exercise items were presented to students during the two semesters of the 2011/2012 academic year. One item property (item type) and one person property (course) are taken into account. In this regard, 35.5% of the items pertain to comprehension, 13.8% to knowledge, and 53.7% to practical computations. The considered data set contains only those persons (134 in number) who have got item scores on at least 5 items. Of these persons, only 8% belonged to the Educational Studies program, while 67.9% and 23.9% belonged to the bachelor programs of Educational Sciences, and Speech Therapy and Audiology Sciences. Overall, the distribution of the number of items per person is skewed to the right with a median and mean of 26 and 34 exercise items, respectively. The minimum number of persons per item is 43 with a mean of 58 and a maximum of 71.

Three explanatory IRT models—an item covariate model, a person covariate model, and a model with both covariates—were fitted to the data set. All covariates were fitted as categorical predictors. The ''course'' covariate includes only two dummy variables. The first dummy equals 1 for persons following bachelor of educational studies and 0 otherwise, while the second dummy variable equals 1 for persons following bachelor of educational sciences and 0 otherwise. Persons following the bachelor of speech therapy and audiology sciences were used as a reference category. Similarly, the ''item type'' covariate was modeled using two dummy variables with comprehension items used as a reference category. For implementation of the MI approach, similar procedures as discussed in the simulation study are followed, while the DL approach is straightforward as before. The obtained results are given in Table 5. From all DL analyses, it appears that the considered covariates are not strongly related to the missing item scores because exclusion of one covariate has little effect on the estimates of the parameters remaining in the model. However, two important observations can be noted. First, taking the DL results for the models including the covariates as a benchmark, all analyses involving $M = 1$ resulted in estimates of fixed effects and of the variances of random effects that are biased downward. The corresponding standard errors under $M = 1$ are also underestimated. Second, it

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

Table 5. Estimates and Standard Errors for the Models Fitted to the Empirical Example by Analysis Approach.

	Estimates						Standard errors					
Parameters	DL	M1	M3	M5	M10	M20	DL	M1	M3	M5	M10	M20
Item covariate model												
Fixed effects												
Intercept	0.03	0.06	0.07	0.02	0.03	0.04	0.22	0.20	0.22	0.23	0.22	0.21
Knowledge	0.30	0.32	0.29	0.36	0.32	0.30	0.38	0.36	0.39	0.38	0.39	0.37
Practice	0.52	0.42	0.50	0.51	0.53	0.50	0.27	0.25	0.27	0.27	0.27	0.26
Random effects												
Persons	0.51	0.46	0.53	0.53	0.50	0.49						
Items	1.05	0.94	1.08	0.99	1.07	0.98						
Person covariate model												
Fixed effects												
Intercept	0.81	0.69	0.87	0.79	0.76	0.76	0.19	0.17	0.18	0.19	0.19	0.18
Educational studies	−0.76	−0.57	−0.82	−0.80	−0.65	−0.68	0.29	0.24	0.29	0.33	0.26	0.28
Educational sciences	−0.58	−0.46	−0.63	−0.56	−0.51	−0.53	0.17	0.14	0.16	0.18	0.17	0.16
Random effects												
Persons	0.45	0.42	0.45	0.46	0.44	0.34						
Items	1.11	0.97	1.13	1.05	1.13	1.08						
Item and person covariate model												
Fixed effects												
Intercept	0.49	0.42	0.56	0.47	0.43	0.45	0.25	0.23	0.24	0.26	0.25	0.25
Educational studies	−0.76	−0.57	−0.83	−0.80	−0.65	−0.69	0.29	0.24	0.29	0.33	0.26	0.28
Educational sciences	−0.58	−0.47	−0.63	−0.56	−0.51	−0.53	0.17	0.14	0.16	0.18	0.17	0.16
Knowledge	0.31	0.32	0.29	0.36	0.32	0.30	0.38	0.36	0.39	0.38	0.39	0.37
Practice	0.51	0.42	0.50	0.51	0.53	0.50	0.27	0.25	0.27	0.27	0.27	0.26
Random effects												
Persons	0.45	0.42	0.45	0.46	0.44	0.44						
Items	1.05	0.94	1.08	0.99	1.07	0.98						

Note. DL = direct likelihood; M# = number of imputations for a multiple imputation approach.

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

can be observed that the estimates and standard errors under the MI approach are relatively stable and comparable for all analyses with $M = 3, 5, 10$, and 20 imputations. Moreover, all the results with $M > 1$ are comparable with those obtained under the DL approach.

Discussion and Conclusion

The main goal of this article was to compare the performance of a MI approach based on empirical Bayes techniques with a DL analysis when analyzing multilevel cross-classified item response data with missing binary item response scores using different explanatory IRT models, and when the model for generating multiply imputed data sets is more general than that used for analysis. The imputation model included covariates that are predictive of the outcome and are related to the missing scores, but any of these may be dropped from an analysis model of interest and still get unbiased estimates. But a similar analysis model using the DL approach can result in biased estimates. Given that the fitted IRT model is specified correctly, analyses of the fully observed data sets (CD) are assumed to give reliable estimates of the parameters of interest. Therefore, both the DL and MI approaches were considered to perform effectively if obtained results are similar to those of CD analysis.

From the simulations, it is observed that when an analysis model contained those covariates related to the missing scores, both the DL and MI approaches produced unbiased and generally similar parameter estimates of fixed effects, with MSE, standard errors (and standard deviations of the estimates), and confidence interval coverage probabilities comparable with those of a CD analysis. This is in conformity with Schafer (2003) who stated that

> if the user of the maximum likelihood (ML) procedure and the imputer use the same set of input data (same variables and observational units); if their models apply equivalent distributional assumptions to the variables and the relationships among them; if the sample size is large; and if the number of imputations is sufficiently large; then the results from the ML and MI procedures will be essentially identical. (p. 23)

Of course, as expected, precision of estimates tended to decline with larger proportions of missing scores—both for DL and MI analyses—and consequently, confidence intervals may not cover the true values very well.

Furthermore, these simulations illustrate that when the covariate causing missingness is not included in the analysis model, the DL approach produced biased parameter estimates of the intercept with MSE values larger than those of the CD analysis. The corresponding coverage proportions of the confidence interval are smaller than the nominal value. These observations are expected because the MAR assumption is violated in such a DL analysis. However, analyses from a similar model fitted to multiply imputed data resulted in estimates that were similar and comparable with those of the CD analysis. This illustrated the main conclusion of this article—that as long as the MAR assumption is postulated, a data set with missing item scores can first be imputed using a more general MI model, and then analysts can fit ''smaller'' models of their interest, even if these models do not include covariates associated with the missingness, but still arrive at similar conclusions as those of the CD analysis. Of course, these conclusions apply in principle only for the conditions considered in this study. It would be interesting to study larger missing proportions and other sample sizes. In the considered conditions, random effects' variance parameter estimates are slightly negatively biased for the DL approach, with the MI approach posting slightly larger biases. However, the corresponding MSE estimates for both approaches are comparable with those of the CD analysis implying no or very little loss of prediction accuracy.

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

The empirical example illustrated that results based on $M = 3$, 5, 10, and 20 are generally similar in the case of a relatively large proportion of missing item scores (i.e., 57%). This further underlines the point that one can obtain reliable and efficient estimates when using a moderate number of imputations. Note that in this particular example, single value imputation resulted in biased estimates and that all standard errors were underestimated. As discussed before, this is because single-value imputations do not account for imputation uncertainty because variability induced by not knowing the missing values is ignored. Hence, this approach is generally not recommended.

The authors conclude that MI is generally desirable when the analysis and imputation approaches differ but in case they are similar, then the DL analysis is a parsimonious approach because it can be implemented with no additional programming involved, yielding similar results as those that would be obtained if the described imputation approach is employed. However, not much is known on how or whether the implemented MI model can be applied to missing covariate data, or to settings outside the scope of the current article. These are areas of possible future research.

Acknowledgments

The authors express kind acknowledgments to the Hercules Foundation and the Flemish Government–Department EWI for funding the Flemish Supercomputer Centre (VSC) whose infrastructure was used to carry out the simulations. They also thank two anonymous reviewers and Editor in Chief Dr. Hua-Hua Chang for their insightful comments and suggestions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Lead contributor is funded under scholarship Ref. BEURSDoc/FLOF.

References

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 S4 classes (R package) [Computer software] . Available from http://cran.r-project.org

Bates, D. M. (2008). *Fitting mixed-effects models using the lme4 package in R*. Retrieved from http://www.stat.wisc.edu/~bates/PotsdamGLMM/LMMD.pdf

Beretvas, S. N. (2011). Cross-classified and multiple membership models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 313-334). Milton Park, England: Routledge.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Routledge.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (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., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

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 from http://www.editlib.org/p/23329

Doran, H., Bates, D. M., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, *20*, 1-18.

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

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

Goldstein, H. (2010). *Multilevel statistical models* (4th ed.). Chichester, England: John Wiley.

Kadengye, D. T., Cools, W., Ceulemans, E., & Van den Noortgate, W. (2012). Simple imputation methods versus direct likelihood analysis for missing item scores in multilevel educational data. *Behavior Research Methods*, *44*, 516-531. doi:10.3758/s13428-011-0157-x

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

Klinkenberg, S., Straatemeier, M., & Van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*, 1813-1824. doi:10.1016/j.compedu.2011.02.003

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John 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*. Chichester, England: John Wiley.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John 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.

Sarkar, D. (2008). *Fitting mixed-effects models using the lme4 package in R*. Retrieved from http://www.bioconductor.org/help/course-materials/2008/PHSIntro/lme4Intro-handout-6.pdf

SAS Institute Inc. (2005). *Proceedings of the Thirtieth Annual SAS Users Group International Conference*. Cary, NC: Author.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Washington, DC: Chapman and Hall/CRC Press.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, *57*, 19-35. doi:10.1111/1467-9574.00218

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

Sheng, X., & Carrière, K. C. (2005). Strategies for analyzing missing item response data with an application to lung cancer. *Biometrical Journal*, *47*, 605-615.

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. (2011). Multiple imputation of multilevel data. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 173-196). Milton Park, England: Routledge.

van Buuren, S., & Oudshoorn, C. G. M. (2000). *Multivariate imputation by chained equations: MICE V1.0 users's manual*. TNO Prevention and Health, Public Health. Retrieved from http://www.stefvanbuuren.nl/publications/mice%20v1.0%20manual%20tno00038%202000.pdf

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014

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 Ginkel, J. R., van der Ark, L. A., Sijtsma, K., & Vermunt, J. K. (2007). Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Computational Statistics & Data Analysis*, *51*, 4013-4027. doi: 10.1016/j.csda.2006.12.022

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: urn:nbn:nl:ui:12-3130382

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: Mathematical, Physical & Engineering Sciences*, *366*, 2389-2403.doi:10.1098/rsta.2008.0038

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*, 589-600.

Downloaded from apm.sagepub.com at Katholieke Univ Leuven on January 16, 2014