8

Imputing Proficiency Data under Planned Missingness in Population Models

Matthias von Davier*
Educational Testing Service

CONTENTS

Introduction

From the perspective of missing data, in most assessments the proficiency of interest is a latent variable that is missing for *all* respondents. We never observe "intelligence" or "reading skill" directly; we only observe how respondents react to certain tasks that are assumed to be indicative of these underlying variables. The responses to these tasks, however, are fallible indicators at best, and only in naïve approaches to measurement are these directly equated to the underlying variable of interest.

* I thank the anonymous reviewers, Irwin Kirsch, Kentaro Yamamoto, and Frank Rijmen for comments on previous versions of this chapter. All remaining errors are, of course, my responsibility.

Because the proficiency variable is missing, an assumption has to be made regarding how observed performance indicators are related to underlying student proficiency. In large-scale educational surveys such as the National Assessment of Educational Progress (NAEP), Trends in Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), and the Programme for International Student Assessment (PISA), an item response theory (IRT; Lord and Novick 1968) model—a special case of a latent structure model (Haberman 1974; Haberman et al. 2008)—is used to derive a model-based estimate of this missing proficiency variable for each respondent. Maximum likelihood estimates, or expected a posteriori (EAP) estimates, together with their associated measures of uncertainty—model-based standard errors or posterior variances—are used in IRT models (compare Chapter 7 of this volume) to derive estimates of proficiency.

The statistical procedures used in educational large-scale assessments extend IRT by involving a latent regression of the respondent proficiency variable(s) on a set of predictor variables (Mislevy 1991). Von Davier et al. (2006) describe this approach as well as recent extensions in some detail and show how it is implemented in NAEP. The rationale behind this combined IRT latent regression approach is the use of all available information to impute accurate distributions of student proficiencies given all available information. This approach follows the rationale developed by Rubin (1987, 1996) suggesting that all available information should be used in an imputation model if the data are missing at random (MAR, see the section on missing data below), that is, if the conditional distribution of the variable (and the indicator variable that shows whether a value is missing or not) depends on the values of observed quantities. In this sense, common practice in large-scale educational assessments follows standard recommendations in the treatment of missing data by the inclusion of all available observed variables in the imputation model.

A deviation of the customary approach to include all variables is the use of rotated background questionnaires in which some sets of variables are not presented to some respondents. Fetter (2001) treats the case of imputations in covariates, a case very similar in design to case cohort studies, and the case of incomplete (rotated) background questionnaires. His results show that if mass imputation of covariates is involved, there can be more, less, or an equal amount of bias in the estimates of relationships with other variables.

Also, the techniques to treat missingness in regression predictors so far applied to latent regressions are not entirely satisfactory. Finally, the fact that substantial parts of the students' background data that were not taken by the student are constant may introduce bias for some or all student groups due to effects related to omitted variable bias in regressions (Hanushek and Johnson 1977; Greene 2003). This paper describes the imputation model used in large-scale assessments such as NAEP, PISA, TIMSS, and PIRLS and shows how incomplete background data by design relates to issues discussed in

relevant literature on missing data and (mass) imputation. In addition, the need for a careful examination of the goals of large-scale survey assessments is pointed out, and a call for reevaluation of the desire to introduce more planned missingness for the sake of broader coverage of a more diverse set of variables is voiced.

Imputation and Missing Data

In the previous section we argued that the latent variable θ can be viewed as a variable that is missing for all respondents. We introduced the specialized imputation model based on a combination of IRT with a multiple linear regression model. In this section we continue with a more general approach to missing data and introduce an important distinction of missing data mechanisms that can be used to explain the rationale of imputation using all observed data in the larger context.

Missing Data Mechanisms

Rubin (1987) developed a system that today is the gold standard of classifying and describing missing data mechanisms. This classification also has important consequences with respect to the treatment of missingness in statistical analysis. It is important to note that the assumptions made and the categorization of missing data mechanisms operates on the level of indicator variables that show whether a value in a data set is missing, not the variables containing the data of interest. More formally, assume there are $n = 1,\ldots,N$ respondents who are expected to provide data on $Y_1,\ldots,Y_K$ variables, so that y_{nk} denotes the value of variable k obtained for respondent n. Note that we do not make a distinction any longer between item responses $X_1,\ldots,X_I$ and covariates $Z_1,\ldots,Z_L$; we use the same letter Y for all (potentially) observed variables. Rubin (1987, 1996) defines missing data indicator variables $M_1,\ldots,M_K$ with

$$m_{nk} = \begin{cases} 0: & y_{nk} \text{ observed} \\ 1: & y_{nk} \text{ missing} \end{cases}$$

so that each respondent is characterized by two vectors of data, $(y_{n1},\ldots,y_{nK})$ representing the variables of interest, and $(m_{n1},\ldots,m_{nK})$ representing the indicators that show whether each of the variables of interest was observed or is missing for respondent n. Each vector of variables $\boldsymbol{y}_n = (y_{n1},\ldots,y_{nK})$ can be split into a vector $\boldsymbol{y}_{n,\text{obs}}$ of all observed responses of respondent n and a vector $\boldsymbol{y}_{n,\text{miss}}$ of what should have been observed if the data were not missing. Note that

we do not know the values in $\boldsymbol{y}_{n,\text{miss}}$. The indicator variables $\boldsymbol{m}_n = (m_{n1},\ldots,m_{nK})$ will then tell us how to recombine the observed vector $\boldsymbol{y}_n = (y_{n1},\ldots,y_{nK})$ from $\boldsymbol{y}_{n,\text{obs}}$ by inserting the appropriate values in those positions where $m_{nk} = 0$.

Then, we can define the following three missing data mechanisms:

MCAR: Y_i is missing completely at random if $P(M_i \mid \boldsymbol{y}_{n,\text{obs}}, \boldsymbol{y}_{n,\text{miss}}) = P(M_i)$.
MAR: Y_i is missing at random if $P(M_i \mid \boldsymbol{y}_{n,\text{obs}}, \boldsymbol{y}_{n,\text{miss}}) = P(M_i \mid \boldsymbol{y}_{n,\text{obs}})$.
NMAR: Y_i is not missing at random if $P(M_i \mid \boldsymbol{y}_{n,\text{obs}}, \boldsymbol{y}_{n,\text{miss}}) \neq P(M_i \mid \boldsymbol{y}_{n,\text{obs}})$.

Note that the missingness condition NMAR states that the occurrence of missing data depends on the missing data $\boldsymbol{y}_{n,\text{miss}}$ itself. Obviously, this is the least desirable situation, which requires more complex modeling in order to find a predictive distribution for Y_i. The three definitions above do not talk about $P(Y_i)$ and potential dependencies between Y_i and the remaining variables in $\boldsymbol{Y}^{(i)}_{n,\text{obs}} = \boldsymbol{Y}_{n,\text{obs}\setminus Y_i}$ and $\boldsymbol{Y}^{(i)}_{n,\text{miss}} = \boldsymbol{Y}_{n,\text{miss}\setminus Y_i}$. This implies that examining

$$P(Y_i \mid \boldsymbol{y}^{(i)}_{n,\text{obs}}),$$

that is, the conditional distribution of Y_i given the observed data as a means to find a predictive distribution. This would be a useful approach when it is determined that missingness is MAR or MCAR. It turns out that many imputation methods indeed use this conditional distribution to derive values for Y_i if it was not observed and missingness is MAR or MCAR. A few examples of these approaches will be given in the next subsection. If missingness is NMAR, as is often the case for ability-related nonresponse in psychological and educational testing, more involved models are required that make use of the missingness indicators $\boldsymbol{m}_n = (m_{n1},\ldots,m_{nK})$ to come up with a predictive distribution for missing variables.

Imputation Methods in a Nutshell

Roughly, imputation describes the generation of (random) replacements of observations that are missing for one reason or another in a data set. Often, imputations are used to provide "complete" data sets in order to allow the use of methodologies that require complete data. Without imputations, many important standard methods that require complete data would be applicable only after list-wise deletion of all cases with missing data, which may introduce bias if the occurrence of missingness is not MCAR.

As an example, if gender is completely observed, that is, it is part of $Y_{\cdot,obs}$, and Y_1 is a variable indicating that the respondent is a smoker, and we have $P(M_1 = 1|\text{male}) > P(M_1 = 1|\text{female})$ (males omit the response to the smoking question more often than women) as well $P(Y_1 = 1|\text{male}) > P(Y_1 = 1|\text{female})$ (males are more likely to smoke), we would obtain a smaller value for $P(Y_1 = 1)$, the proportion of smokers, than expected if we delete cases listwise. That is

because we would delete more records of males who are more likely to smoke and would omit the response to the smoking question.

Imputations may help in this case if there are methods that can take these types of relationships into account. However, it may not be enough to know the gender in order to impute missing self-reports on smoking. It may also be useful to know whether the respondents' parents were smokers, or if there was other substance use, as well as knowledge about blood pressure, body-mass index, type of job, or other variables that may have a statistical relationship to smoking. When imputing, it seems, it would be good to use all the information we have, and it appears to be so when looking at standard recommendations put forward, for example, by Little and Rubin (2002). Indeed, Collins et al. (2001) show that what they call the inclusive strategy is greatly preferred (an imputation model that contains as many covariates as possible and available) compared to the restrictive strategy (an imputation model with a limited number of variables, for example, only the core variables in a rotated student questionnaire).

There are many imputation methods, and this chapter discusses only a select number of these methods. Interested readers are referred to Little (1988) as well as Little and Rubin (2002) for a discussion of a range of imputation methods. Among the ones frequently found in the literature are

1. *Hot deck imputation*—A seemingly simple way of coming up with a procedure that "imputes" values for missing observations. Under this approach, a respondent with missing data is simply compared to all other respondents (on the set of observed variables that are not missing), and a "donor" is chosen that is maximally similar to the respondent with missing values. Then, the respondent receives "copies" of the variable values that are missing from this donor.
2. *Multivariate normal distribution*—Another commonly used method for imputation of missing observations due to non-response, which allows generating conditional distributions of variables with missing observations. Obviously, this approach assumes that all observed variables are at least interval level and are (approximately) normally distributed. This approach is therefore not very useful for completion of questionnaire and test data. Instead, methods that allow for modeling multidimensional categorical data are suitable for imputations (e.g., Vermunt et al. 2008).
3. *Multiple imputation by chained equations*—If no multivariate distribution can be found to model all available data, multiple imputation by chained equations is a suitable alternative (e.g., van Buuren and Groothuis-Oudshoorn 2011). This method specifies a multivariate imputation model on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable. After an initial imputation, this method draws imputations by iterating over the

> imputation model for the different variables. Typically, a low number of iterations are used. The approach is attractive because it can be implemented by means of imputing first for the variable with the fewest missing values, using only complete variables in the model, then moving to the next variable with the second-lowest number of missing values, and so on, until values for all variables are imputed and a more comprehensive imputation model can be used for subsequent imputations.

Most imputation methods are typically considered for small amounts of missing data based on item-level non-response, and many of the recommendations about the use of only five imputed data sets come from the fact that only small percentages of missingness are considered. If large amounts of missing data are considered, this situation is known as "mass imputation" (e.g., Haslett et al. 2010), and standard methods of imputation may not yield the expected quality of the imputed values (Fetter 2001) in terms of the appropriate recovery of statistical relationships between variables in the database.

Also, as Piesse and Kalton (Westat Research Report 2009) put it: "The dominant concern with imputation is that it may affect the associations between items. In general, the association between an imputed item and another item is attenuated towards zero unless the other item is included as an auxiliary variable in the imputation scheme." As Rubin (1996) puts it: "More explicitly, when Q or U involves some variable X, then leaving X out of the imputation scheme is improper and generally leads to biased estimation and invalid survey inference. For example, if X is correlated with Y but not used to multiply-impute Y, then the multiply-imputed data set will yield estimates of the XY correlation biased towards zero."

This explains why all available observed variables customarily are included in the imputation model applied to produce multiple "completed" data sets. If mass imputation is involved, or if many variables are affected by missing data, this approach can only be followed to some extent, because for many imputations, other observations for the same respondent may be missing as well. In the case of mass imputation, when large sets of variables are jointly missing, the imputation problem becomes circular, and the result may depend on what was chosen to be imputed initially as well as the sequence in which variables were "completed" in the process.

Conditional Independence and Imputation

The categorization of the different missing data mechanisms in the above section used the concept of conditional independence casually without a formal introduction. Because this is a central concept of importance for many of the issues discussed in this chapter, we reiterate the definition and discuss a few implications of conditional independence assumptions more formally.

Let X denote a binary variable that indicates performance on a complex computer simulation task that is intended to measure computer literacy, with $x_n = 1$ indicating that respondent n solves the task, and $x_n = 0$ otherwise. Let Z_* denote a variable that indicates whether at least one of the respondent's parents has a college degree, $z_{n*} = 1$, or not, $z_{n*} = 0$. Let Z_A denote a variable that indicates whether the respondent has his or her own computer at home, $z_{nA} = 1$, or not, $z_{nA} = 0$. Then we can define the following:

The random variables X and Z_A are conditionally independent given Z_* if and only if

$$P(X, Z_A \mid Z_*) = P(X \mid Z_*)P(Z_A \mid Z_*),$$

that is, if conditional independence is given for X and Z_A, we can factorize the joint probabilities into a product of marginal probabilities. This definition leads to the following important corollary

$$P(X \mid Z_*, Z_A) = P(X \mid Z_*),$$

that is, if X and Z_A are conditionally independent given Z_*, then the conditional distribution of X given Z_* and Z_A equals the conditional distribution of X given only Z_*. By symmetry, we have

$$P(Z_A \mid Z_*, X) = P(Z_A \mid Z_*)$$

and we write

$$X \perp Z_A \mid Z_*.$$

This means that if conditional independence of two variables X and Z_A given a third Z_* is established, then we can use only the third variable Z_* and leave the second, Z_A, out of the imputation model for X, and vice versa. Of course, this assumes we keep the conditioning variable (here, only Z_*) in the imputation model, that is, the variable required to establish conditional independence.

Conditional independence can be considered a strong assumption about the relationships among observed variables. Typically, a conditional independence assumption is used in order to define idealized (latent) variables that make all other influences "vanish" or allows us to factorize the likelihood of the joint occurrence of a series of variables by assuming an underlying probabilistic "cause" (Suppes and Zanotti 1995). One of the standard examples where conditional independence (also called local independence in that context) is assumed is in latent structure models, such as IRT (see the Introduction for this chapter as well as Lord and Novick 1968) or latent class models (Lazarsfeld and Henry 1968).

In imputation models, one typically avoids such a strong assumption, mainly out of recognition that wrongly assuming conditional dependence

and eliminating variables from the imputation model may introduce bias in estimates of relationships (e.g., Rubin 1996; Collins et al. 2001; Piesse and Kalton 2009). As a consequence, if certain conditional relationships are incorrectly assumed to be nonexistent for some groups of variables, then bivariate or higher-order dependencies between these variables may not be reproduced accurately. Therefore, imputation models typically are maximally inclusive by using all available observed variables for imputation of missing observations whenever possible.

Let us now assume that we have a fourth variable, Z_B, which may indicate that a respondent reports a high ($Z_B = 1$) or low ($Z_B = 0$) level of self-efficiency in terms of computer literacy. Note that conditional independence is not a condition that can be extended to this additional variable in trivial terms. More specifically,

$$X \perp Z_A \mid Z_* \nLeftrightarrow X \perp Z_A \mid (Z_*, Z_B),$$

that is, if X and Z_A are conditionally independent given Z_*, that does not necessarily imply that X and Z_A are conditionally independent given Z_* and Z_B. As an example, consider Table 8.1, which represents the joint distribution of four binary variables.

It can be seen in Table 8.1 that the sum of all probabilities adds up to 1.0 in either rows or columns of the table. All four conditional tables, when normalized to a sum of 1.0 within the 2×2 table, describe the distribution of X and Z_A given the levels of Z_* and Z_B. Table 8.2 shows the conditional tables of X and Z_A given Z_*.

The two tables above show that the conditional distribution of X and Z_A is independent given Z_*. For example, $0.32/0.5 = (0.32/0.5 + 0.08/0.5)^2$, $0.02/0.5 = (0.02/0.5 + 0.08/0.5)^2$, and so on. This means that by collapsing tables across the two levels of Z_B we produced two tables in which conditional independence holds. Finally, Table 8.3 shows that the two variables X and Z_A are not independent when not conditioning on any variable.

TABLE 8.1

Joint Distribution for Which We Do Not Show an Association between and Z_A, but Conditional Independence Holds as in $X \perp Z_A \mid Z_*$, but $X \perp Z_A \mid (Z_*, Z_B)$ Does Not Hold

		$Z_* = 0$				$Z_* = 1$		
		0	1			0	1	
$Z_B = 0$	0	0.31	0.01		0	0.01	0.07	
	1	0.01	0.01		1	0.07	0.01	
				0.34				0.16
		0	1			0	1	
$Z_B = 1$	0	0.01	0.07		0	0.01	0.01	
	1	0.07	0.01		1	0.01	0.31	
				0.16				0.34

TABLE 8.2

Conditional Distributions with Conditional Independence for $X \perp Z_A | Z^*$

		$Z_* = 0$			$Z_* = 1$		
		0	1		0	1	
$Z_B = 0 \vee 1$	0	0.32	0.08	0	0.02	0.08	
	1	0.08	0.02	1	0.08	0.32	
			0.5				0.5

The example shows that it is not enough to find a set of variables for which conditional independence holds. *When more predictors are obtained, conditional independence relationships may change*. More specifically, it may be that the distribution $P(X|Z_*,Z_A,Z_B)$ does differ from $P(X|Z_*,Z_B)$, even if it was found that $P(X|Z_*) = P(X|Z_*,Z_A)$. Therefore, if a predictive distribution for imputations is wanted, it appears a good strategy is to include all available variables in order to ensure that the relationship between X and the other variables is maintained when values for X are imputed.

Let us also assume that the additional variable Z_B is assessed together with X and Z_* in only one half of the sample (Sample B), while Z_A is assessed in the other half (Sample A) together with X and Z_*. Let us further assume that conditional independence does NOT hold for X and Z_B given Z_*, but does hold for X and Z_A, given Z_*. We are facing the following situation: We cannot define an imputation model for X given all three variables Z_*, Z_A, Z_B because they are never jointly observed without making additional strong assumptions that allow us to construct a synthetic joint distribution (Little and Rubin 2002).

If we try to devise an imputation model without making additional assumptions, we are bound to use the variables that are used in each of the half samples. If we assess a respondent in Sample B who has a missing response on X, we are well advised to use an imputation model that contains both Z_* and Z_B because conditional independence does not hold. If, however, we have a missing response on X for a respondent from Sample A, the imputation model for this person can only contain Z_*, and optionally, but without consequence, Z_A. Z_B cannot be added because only Z_A was observed, but X and Z_A are conditionally independent given Z_*. This means that when imputing missing X responses in Sample A, we can only use Z_* but an important

TABLE 8.3

Marginal Distribution with a Positive Association between X and Z_A

		$Z_* = 0 \vee 1$	
		0	1
$Z_B = 0 \vee 1$	0	0.34	0.16
	1	0.16	0.34
			1.0

piece of the information is lost because Z_B, of which X is not conditionally independent given Z_*, was not collected in Sample A.

Just to drive the point home: If the background variable Z_B were omitted in some way (not collected, or only collected on a subset of the sample), the joint distribution under a missingness of X of these four variables would not be recovered correctly in an imputation model that contains only Z_* and Z_A. Moreover, an imputation model using the latter two variables would actually suggest that $X \perp Z_A$, and thus only $P(X|Z_*)$, could be used, and thus, information is lost and bias introduced in secondary analyses that would involve imputed values of X and variables that are statistically associated or correlated with Z_A, Z_B.

Imputation of Proficiency in Large-Scale Assessments

In large-scale educational survey assessments, 100% of the student proficiency data is imputed using a specialized imputation model based on statistical procedures that are tailored to incorporate both cognitive response data and student background data (Mislevy 1991; von Davier et al. 2006, 2009; and Chapter 7 in this volume). These procedures are referred to as latent regression models and provide EAP estimates and estimates of posterior variance of proficiency. These estimates are based on a Bayesian approach and thus are using a prior distribution of proficiency, a feature that opens an avenue to the introduction of conditional prior distributions, for example, in the shape of a multiple-group IRT model (Bock and Zimowski 1997; Xu and von Davier 2008).

Latent regression models extend the multiple-group IRT model and provide a different conditional prior distribution for each respondent's proficiency based on a set of predictor variables (Mislevy 1991; von Davier et al. 2006).

Item Response Theory

Let θ denote the latent variable of interest. This may be a variable representing mathematics, reading, or science proficiency. Let $X_1,\ldots,X_I$ denote variables that represent the responses to test items that are assumed to be governed by θ, that is, assume that for each respondent n, the responses to X_i are denoted by x_{ni} and that

$$P_i(x_{ni} = 1 \mid n) = P_{\zeta_i}(x_{ni} = 1 \mid \theta_n)$$

with some item-dependent parameters ζ_i and with θ_n denoting the n-th respondent's proficiency. While most educational assessments use multiparameter IRT models such as the 2PL and 3PL IRT model (Lord and Novick

1968) for binary data, and the generalized partial credit model (Muraki 1992) for polytomous data, PISA has been using the Rasch model (Rasch 1960/80) as the basis (Adams et al. 2007). The probability function defining the IRT models used can be written as

$$P_{\zeta_i}(x_{ni}=1\,|\,\theta_n)=c_i+(1-c_i)\frac{\exp\left(\alpha_i(\theta_n-\beta_i)\right)}{1+\exp\left(\alpha_i(\theta_n-\beta_i)\right)}$$

and $P_{\zeta_i}(x_{ni}=0\,|\,\theta_n)=1-P_{\zeta_i}(x_{ni}=1\,|\,\theta_n)$. This model may be used for binary responses $y \in 0,1$ with item parameters $\zeta_i=(\alpha_i, \beta_i, c_i)$, whereas

$$P_{\zeta_i}(X=x_i\,|\,\theta_n)=\frac{\exp\left(\sum_{z=1}^{x_i}\alpha_i(\theta_n-\beta_{iz}^{*})\right)}{1+\sum_{z=1}^{k_i}\exp\left(\sum_{w=1}^{z}\alpha_i(\theta_n-\beta_{iw}^{*})\right)}$$

may be assumed for polytomous ordinal responses $y \in \{0,\ldots,k_i\}$ with item parameters $\zeta_i=(\alpha_i,\beta_{i1},\ldots,\beta_{ik_i})$. Together with the usual assumption of local independence of responses given θ we obtain

$$P_{\zeta}(x_{n1},\ldots,x_{nI}\,|\,\theta_n)=\prod_{i=1}^{I}P_{\zeta_i}(X=x_{ni}\,|\,\theta_n).$$

Customarily, it is assumed in IRT models that the ability parameter follows a distribution $f(\theta;\eta)$ where η describes the parameters of the distribution that may be fixed or estimated from the data, depending on whether the latent scale is set and how other parameters such as ζ are constrained in order to remove indeterminacy in scale and location. The distribution $f(\theta;\eta)$ is then used to derive the marginal probability of a response pattern, that is,

$$P_{\zeta,\eta}(x_{n1},\ldots,x_{nI})=\int_{\theta}\left[\prod_{i=1}^{I}P_{\zeta_i}(X=x_{ni}\,|\,\theta)\right]f(\theta;\eta)d\theta.$$

Maximum likelihood methods are the customary approach to obtain estimates of the parameters η,ζ. For this purpose, the log-likelihood function

$$\ln L(\eta,\xi;X)=\sum_{n=1}^{N}\ln\left(\int_{\theta}\left[\prod_{i=1}^{I}P_{\zeta_i}(X=x_{ni}\,|\,\theta)\right]f(\theta;\eta)d\theta\right),$$

where X represents the $N\times I$ matrix of item responses from all respondents, is maximized with respect to η and ζ.

The item response data $\boldsymbol{X} = [(x_{11},\ldots,x_{1I}),\ldots,(x_{N1},\ldots,x_{NI})]$ modeled by IRT may contain missing responses by design (MCAR) as well as item-level nonresponses that often are assumed to be ignorable as well. While the missingness by design is introduced by means of matrix sampling of item responses (Mislevy 1991) and can be assumed to be MCAR and thus ignorable, the missingness by means of respondents choosing to not provide an answer to one or more items can be assumed to be MAR at best, but there is strong indication in real data that nonresponse is informative, that is, NMAR (e.g., Rose et al. 2010).

Usually the number of omitted responses is small, but for larger amounts of informative (NMAR) missing data, model-based approaches to handle missingness have been devised. For these approaches, we refer interested readers to Moustaki and Knott (2000), Glas and Pimentel (2008), as well as Rose et al. (2010). We will not discuss this particular topic of non-ignorable missing item response data further because it is not within the focus of this chapter.

Latent Regression

In addition to the observed item response variables X_i, latent regression models assume there are additional variables collected for each respondent. Let these variables be grouped into collections of variables $Z_* = (Z_{*1},\ldots,Z_{*I^*})$ as well as $Z_A = (Z_{A1},\ldots,Z_{AI^A})$, $Z_B = (Z_{B1},\ldots,Z_{BI^B})$, and $Z_C = (Z_{C1},\ldots,Z_{CI^C})$, or more groups of covariates. One of these variable groups may represent measures of home background and socioeconomic status (say Z_A) while another group (say Z_B) may represent student attitudes and motivational variables, and so on.

It is common practice in most national and international large-scale assessments to collect observations on all variable groups $\boldsymbol{Z} = (Z_*, Z_A, Z_B, Z_C)$ for all respondents who provide responses on the test items $X_1,\ldots,X_I$. This practice is motivated by the use of these covariates in the latent regression that we will define formally below. Using these groups of covariates, the conditional distribution of the latent trait variable θ is defined as

$$P(\theta \mid z_{n^*}, z_{nA}, z_{nB}, z_{nC}) = f(\theta; \boldsymbol{z}_n)$$

with $\boldsymbol{z}_n = (z_{n^*}, z_{nA}, z_{nB}, z_{nC})$ denoting the realizations of $\boldsymbol{Z} = (Z_*, Z_A, Z_B, Z_C)$ for respondent n. If all the covariates are continuous, one may use

$$\mu_\theta(z_{n^*}, z_{nA}, z_{nB}, z_{nC}) = \sum_{w\in\{*,A,B,C\}} \sum_{i=1}^{I^w} \gamma_{iw} z_{nwi} + \gamma_0,$$

that is, a linear predictor based on the covariates with regression parameters $\Gamma = (\gamma_{*1},\ldots,\gamma_{CI_w}, \gamma_0)$, and a common conditional variance

$$\Sigma = V(\theta \mid Z_*, Z_A, Z_B, Z_C)$$

for which an estimate can be obtained using ordinary least-squares (OLS) regression methods within an EM algorithm (Mislevy 1991; von Davier et al. 2006). To obtain estimators $\hat{\Sigma}$ and $\hat{\Gamma}$, the estimation needs to take into account that θ is unobserved (latent) and therefore the regression does not have access to the values of the dependent variable. Instead, the conditional density of θ given z_n by

$$f(\theta; z_n, \Gamma, \Sigma) = \phi(\theta; \Gamma, \Sigma)$$

is used in an iterative scheme to estimate Γ and Σ (Mislevy 1991; von Davier et al. 2006).

Imputing Plausible Values from the Posterior Distribution

While the marginal distribution of responses in ordinary IRT does not depend on any other observed variables, the marginal distribution of a response pattern in latent regressions does depend on the observed covariates z_n. We obtain

$$P(x_{n1}, .., x_{nI} \mid z_n) = \int_{\theta} \left[\prod_{i=1}^{I} P_{\zeta_i}(X = x_{ni} \mid \theta) \right] f(\theta; z_n, \Gamma, \Sigma) d\theta_\theta.$$

Then, the predictive distribution of θ given item responses and covariates $x_1 \ldots x_I$, z_n is

$$p(\theta \mid x_{n1} \ldots x_{nI}, z_n) = \frac{\left[\prod_{i=1}^{I} p(x_i \mid \theta) \right] f(\theta; z_n)}{\int_{\theta^*} \left[\prod_{i=1}^{I} p(x_i \mid \theta^*) \right] f(\theta; z_n)}$$

It is important to understand that the above expression is an imputation model for θ given $x_1, \ldots, x_I$, z_n in that it allows us to derive an expected value and a variance of the proficiency variable given the item responses $x_1, \ldots, x_I$ and the covariates $z_1, \ldots, z_D$. Even more important to understand is that the relationship between θ and $x_1, \ldots, x_I$ is model-based and prescribed to follow a strong monotone relationship given by an IRT model (Hemker et al. 1996), while the relationship between θ and $z_1, \ldots, z_D$ is estimated using a linear regression as described in the previous section.

The latent regression IRT model is employed to derive imputations for public use files that can be used by secondary analysts. More specifically,

each proficiency variable is represented in the public use data by a set of M imputations, in this context called plausible values (PVs), for each respondent in the data. These PVs depend on the item responses (proficiency indicators) as well as covariates such as student self-reports on activities, attitudes and interests, socioeconomic indicators, and other variables of interest for reporting. The rationale is to include all covariates that are available and may be of analytic interest in order to reduce bias in secondary analysis (Mislevy 1991; von Davier et al. 2009). Following the rationale put forward by Rubin (1987, 1996) and Mislevy (1991), all available background and item response data are used in the imputation in order to minimize bias in the estimates of statistical quantities involving these imputations.

Proficiency Imputation under Missingness in Background Data

In the previous section we learned that the inclusion of as many relevant variables as possible into the imputation model is superior to a strategy that includes only a minimum set of variables (Rubin 1996; Collins et al. 2001). This means that if a large amount of variables have a substantive proportion of missing data, it is to be expected that some bias in the estimation of relationships between the partially observed and other variables will be introduced.

This reasoning put forward by Rubin (1987, 1996) and others led to the convention to use as many variables as possible that will provide data about the student in the population modeling in all PISA (and other survey assessments) cycles, with the exception of PISA 2012. Therefore, use of rotated background questionnaires appears not to follow longstanding advice to use all available information on students in the population model to derive multiple imputations of student proficiency (Mislevy 1991; Rubin 1996; von Davier et al. 2006). We will illustrate two versions of planned missingness in background data that deviate from the imputation rationale put forward by Rubin (1987, 1996; Little and Rubin 2002) and others in Tables 8.4 and 8.5.

TABLE 8.4

File Matching Case of Missing Data

Rotation	X_1	X_2	X_3	θ	Z_*	Z_A	Z_B	Z_C
I	X	X			X	X		
II		X	X		X		X	
III	X		X		X			X
IV	X	X			X		X	
VI		X	X		X			X
VII	X		X		X	X		
…	…	…	…	…	…	…	…	…
??	X		X		X		X	

TABLE 8.5

Incomplete Background Data Collection Using Rotation

Rotation	X_1	X_2	X_3	θ	Z_*	Z_A	Z_B	Z_C
I	X		X		X	X	X	
II		X	X		X		X	X
III	X	X			X	X		X
IV	X	X			X		X	X
VI		X	X		X	X		X
VII	X		X		X	X	X	
…	…	…	…	…	…	…	…	…
??	X		X		X		X	X

In these tables, an empty cell denotes missing observations and a X denotes observed data. Note that all cells below θ are empty because the proficiency is missing for all respondents and has to be imputed for 100% of all cases. There is also planned missingness in the observed item responses $X_1,\ldots,X_I$, of which we will assume for simplicity, but without limiting generality, that these observed variables are collected in three blocks X_1, X_2, X_3 that are combined into three booklets I, II, III. The IRT model assumption does provide a conditional independence assumption that allows us to model the joint distribution of the observed responses given ability θ even though they are not jointly observed. By means of the IRT model assumption of an underlying *probabilistic cause* (Suppes and Zanotti 1995)—which we refer to as latent variable θ in IRT—we are able to derive the joint conditional distribution even if not all observed variables were collected jointly on all respondents (see, e.g., Zermelo 1929).

It is important to note that the same reasoning does not apply with respect to the sets of covariates Z_*, Z_A, Z_B, Z_C unless another set of model assumption is made and all variables collected in these covariates are assumed to be related to one or more underlying latent variables. While this may be a reasonable assumption for some subsets of the covariates, it does not appear trivial to develop a comprehensive latent variable model that is defensible and contains explanatory relationships between a set of latent variables and all covariates in Z_*, Z_A, Z_B, Z_C.

If some significant part of the background information is not collected on a large portion of the students, as is true in rotated designs discussed and implemented so far, this violates a basic imputation rationale (Rubin 1987, 1996) used in the majority of large-scale assessments such as PISA, TIMSS, PIRLS, and NAEP. Also, as Little and Rubin (2002, p. 7) point out in their introduction to missing data problems, if there are two (or more) sets of variables that are never jointly observed, associations between these two sets of variables cannot be estimated accurately. Little and Rubin (2002, p. 7) call this missing data design the *file matching* case, and the data collection design that leads to this file matching can be illustrated as in Table 8.4. We will use the same notation as in previous sections, but will now look at the case where

different patterns of missingness are found in the covariates Z_*, Z_A, Z_B, Z_C. Table 8.4 shows a balanced block design for the item responses X_1, X_2, X_3 but a file matching data collection design for Z_*, Z_A, Z_B, Z_C.

In the case depicted in Table 8.4, only the covariates in Z_* are collected in all cases, but the variables collected in Z_A are never jointly observed with variables in Z_B or Z_C. Given this setup, an imputation model for Z_A would need to assume that Z_A given Z_* is conditionally independent of Z_B, Z_C, that is, an imputation model would produce a conditional distribution $P(Z_A|Z_*)$, which involves neither Z_B nor Z_C even though it may likely be that these are not conditionally independent given Z_*. As Little and Rubin (2002, p. 7) point out, the partial associations between the rotated, never jointly observed, variable sets are unknown, and that "in practice, the analysis of data with this pattern typically makes the strong assumption that these partial associations are zero." While this case would probably not be considered for obvious reasons stated above, it is important to note that the missingness introduced by this design requires a strong independence assumption as pointed out by Little and Rubin (2002). A less extreme case that could be described as a relaxed file matching case is illustrated in Table 8.5.

In the example depicted in Table 8.5, an imputation model for Z_A could use the conditional distribution $P(Z_A|Z_*, Z_C)$, assuming conditional independence of Z_A and Z_B given Z_*, Z_C. Alternatively, one could use the conditional distribution $P(Z_A|Z_*, Z_B)$, assuming conditional independence from Z_C, to impute values in Z_A. This would lead to two competing imputation models that may produce different sets of predictive distributions for the same set of "true" but unobserved values of Z_*, Z_B, Z_C. If all variables would have been jointly observed for some subsample, or (strong) model assumptions are used to come up with a joint distribution, a predictive distribution $P(Z_A|Z_*, Z_B, Z_C)$ that involves all other variable groups except Z_A could be used instead. Without this, students with identical background profiles, but from different background questionnaire rotation groups, may receive different predicted imputed values based on the rotation that was received.

The design in Table 8.5, while less sparse than the case depicted in Table 8.4, will still need additional (strong) model assumptions, or sequential (chained) methods in order to derive an imputation model that involves all covariates. Note that chained methods have to start with only observed variables, and usually choose the variables that are complete for all cases (here Z_*). Thus, they would be penalized by a design like the one in the tables because this would essentially amount to a mass imputation problem that has been shown to potentially lead to unpredictable bias in estimates involving imputations (Fetter 2001).

Assumptions Needed for Imputation with Incomplete Background Data

The underlying issue to address is whether there is evidence that the different rotation schemes in background questionnaires may introduce bias

due to omitted variables in the latent regression models used for population modeling in large-scale assessment. Given the above discussions of customary imputation methods that tend to include all available data as well as the description of the latent regression model that provides imputation of proficiency for 100% of all respondents, there are some central questions to be answered before a decision about using incomplete questionnaires.

1. When imputing proficiency values for θ, what are the implications of using incomplete questionnaire data? If there are questionnaire and test booklet rotations such as I, II, III, IV, ... as illustrated in the table, there are different options for the use of these data as they are introduced into the latent regression model. We list a range of theoretically possible approaches to imputing θ and mention some of the assumptions and consequences on sample size and uniqueness of the imputation models involved:
   a. One option would be to use a model that incorporates only observed covariates that are available for all respondents: $P(\theta \mid X_2, X_3, Z_*; \zeta, \Gamma, \Sigma)$ or $P(\theta \mid X_1, X_3, Z_*; \zeta, \Gamma, \Sigma)$ would be the three possible combinations of background variables Z_* and blocks of item responses X_1, X_2, X_3 in the example design introduced above. Note that these could be estimated in one comprehensive latent regression model because the IRT model can handle MCAR and MAR observations in item response data X_1, X_2, X_3. In this case we would make the strong (Little and Rubin 2002) and untestable assumption that the posterior distribution of θ does not depend on Z_A or Z_B or Z_C or any combination of these, that is, θ and any combination of the omitted covariate blocks are indeed conditionally independent given one of the three combinations of X variables and Z_*. The assumption is untestable because the four sets of covariates are not observed jointly.
   b. It could be considered using different latent regression imputation models for different subpopulations: All combinations of cognitive items X_1, X_2, X_3 may be combined in the IRT part, while three different latent regressions would be estimated that include only those respondents who were assessed with either the jointly occurring Z_*, Z_A, Z_B; Z_*, Z_A, Z_C; or Z_*, Z_B, Z_C background rotations. The imputation models would be $P(\theta \mid X_1, X_2, X_3, Z_*, Z_A Z_B; \zeta, \Gamma^{(C)}, \Sigma^{(C)})$, $P(\theta \mid X_1, X_2, X_3, Z_*, Z_A Z_C; \zeta, \Gamma^{(B)}, \Sigma^{(B)})$, or $P(\theta \mid X_1, X_2, X_3, Z_*, Z_B Z_C; \zeta, \Gamma^{(A)}, \Sigma^{(A)})$, each of which is defined by the inclusion of all available background data, that is, each is missing only one of the rotated sets of background variables. This separate estimation of the imputation model parameters $\Gamma^{(C)}$, $\Sigma^{(C)}$ (those without parameters for Z_C), as well as $\Gamma^{(B)}$, $\Sigma^{(B)}$ (those without Z_B and $\Gamma^{(A)}$, $\Sigma^{(A)}$

(without Z_A) would cut the sample size for the latent regression models by a factor of 3, and would produce three latent regression estimates, one for each of the three subsamples. Again, certain conditional independence assumptions would be necessary, without being testable, for the same reason as above.

c. Another option would be a joint estimation using an imputation model that includes all cognitive variables and all sets of background variables when observed, and a placeholder if missing by design. Instead of separately estimating an imputation model for each of the three rotations, one joint model will be defined. For rotations I, placeholder values (mode if the variables are categorical, mean if the variables are continuous) are defined and set to be constant for all respondents who have missing data. Additionally, an indicator variable will be defined that shows which variables are missing and were replaced by modal or mean values. Let us denote this constant vector Z_C^{miss} if values in Z_C are missing by design, and Z_B^{miss} and Z_A^{miss} likewise. Now the data matrix that is used to estimate $P(\theta \mid X_1, X_2, X_3, Z_*, Z_A Z_B, Z_C; \zeta, \Gamma, \Sigma)$ is defined by replacing missing values in $Z_A Z_B$, Z_C by the *constant* vectors $Z_A^{\text{miss}} Z_B^{\text{miss}}, Z_C^{\text{miss}}$ and the imputation model parameters $\Gamma^{(\text{miss})}, \Sigma^{(\text{miss})}$ would be estimated. Note that it is a well-known fact that in linear regressions, this approach does not lead to consistent estimates and will underestimate the correlations of the background variables and the dependent variable θ (e.g., Jones 1996). More specifically, we will have $E\left(\Gamma^{(\text{miss})}\right) \neq \Gamma$ and $E\left(\Sigma^{(\text{miss})}\right) \neq \Sigma$, that is, the replacement of data missing by design will introduce data that are constant for large parts of the sample, which in turn will tend to bias the relationship between proficiency variable θ and predictors Z_A, Z_B, Z_C downward. Two different simulation studies (Rutkowski 2011; PISA TAG(1204)6b) have shown that this is a result that can also be seen in latent regressions. If in addition there are domains to be imputed without associated cognitive data, this effect is likely to be even more pronounced. The basic underlying problem is that data that were vectors "completed" using the constant vectors $Z_A^{\text{miss}}, Z_B^{\text{miss}}, Z_C^{\text{miss}}$ do not represent the statistical dependencies in the data accurately, and models estimated based on this replacement will underestimate the relationships in the data.

d. Finally, using different but overlapping latent regression models could be considered: Different combinations of cognitive items X_1, X_2, X_3 may be present in the IRT part, while the latent regressions would be based on two-thirds of the sample: All respondents who took either Z_*, Z_A, Z_B or Z_*, Z_A, Z_C would be modeled by a regression model that predicts θ by $P(\theta \mid X_1, X_2, X_3, Z_*, Z_A; \zeta, \Gamma^{(B,C)}, \Sigma^{(B,C)})$, and

another two-thirds of the sample would be used in a regression that allows prediction using $P(\theta \mid X_1, X_2, X_3, Z_*, Z_B; \zeta, \Gamma^{(A,C)}, \Sigma^{(A,C)})$ and $P(\theta \mid X_1, X_2, X_3, Z_*, Z_C; \zeta, \Gamma^{(A,B)}, \Sigma^{(A,B)})$. In this case, each respondent would obtain two (potentially different) predictive distributions. A respondent who took Z_*, Z_B, Z_C would be assigned imputations based on the two different imputation models $P(\theta \mid X_1, X_2, X_3, Z_*, Z_B; \zeta, \Gamma^{(A,C)}, \Sigma^{(A,C)})$ and $P(\theta \mid X_1, X_2, X_3, Z_*, Z_C; \zeta, \Gamma^{(A,B)}, \Sigma^{(A,B)})$. Each of these imputation models is using two-thirds of the sample and would assume that conditional independence holds for θ and two out of the four sets of BQ covariates.

2. Prior to imputing θ, a procedure could be devised that allows completing the data for the covariates Z_*, Z_A, Z_B, Z_C. The model size and dependency on model assumptions (e.g., normality of the latent trait) will require some decisions to make the imputation problem tractable. For example,
   a. A model that allows us to specify a joint distribution $P(Z_*, Z_A, Z_B, Z_C \mid \omega)$ given certain distributional assumptions could be employed. This would ignore the cognitive responses X_1, X_2, X_3 and thus imply certain strong assumptions regarding the conditional independence of the imputed Z sets and X_1, X_2, X_3 given the other background variable (Z) sets. The question is whether this can be justified without the use of the cognitive data, which are typically significantly correlated with many of the background variables assessed in the Z sets of variables.
   b. Alternatively, a model that contains the cognitive data could be used as well, for example, in the form of a joint distribution $P(X_1, X_2, X_3, Z_*, Z_A, Z_B, Z_C \mid \omega, \rho)$. This is a model of quite impressive proportions. Many large-scale assessments use item pools of 120 items and more. PISA uses about 170–200 items in three dimensions—reading, science, and mathematics—so that $\theta = (\theta_R, \theta_{M,} \theta_S)$. In addition, the data are quite sparse, with only a fraction of the X variables observed, because sparse balanced incomplete block designs are employed that use 13 or more instead of only three test forms in our example. Once a decision has been made on which approach to take, it is possible to develop an imputation model for each of the missing variables. As an example, if variable set Z_A is missing, the imputation model

$$P(Z_A \mid Z_*, Z_B, Z_C, X_1, X_2, X_3; \omega, \rho)$$

could be used to generate multiple draws for each respondent who was not given Z_A. The same rationale applies to the

imputation of other sets of variables Z_B, Z_C, Z_*. While there is not a question about the ability to mechanically execute such a procedure, the question remains whether bias can be avoided in this type of mass imputation (Fetter 2001; Haslett et al. 2010).

The main question that emerges is which conditional independence assumptions can be justified, if any. More specifically, can the generation of data for public use files using imputation models be justified to be carried out using a predictive distribution that involves only a subset of the background variables? From the enumeration of cases above, it should be clear that it is not only the choice of variables that is limited when using incomplete rotated background questionnaires, but it is also the sample sizes available for the different imputation models that are limited, and the generality of the imputed values may be limited due to the noninclusion of potentially important predictors.

We have so far not discussed the case where certain background characteristics are collected only in conjunction with certain cognitive domains. As an example, assume that there are two cognitive domains θ_1 and θ_2 and each is assessed by presenting different sets of tasks, say X_{11}, X_{12} for θ_1 and X_{21}, X_{22} for θ_2. In addition, assume that there are specific covariates Z_1 and Z_2, each of which will be assessed only when the task sets for θ_1 and θ_2 are tested. In this case, Z_1, X_{11}, X_{12} and Z_2, X_{21}, X_{22} are jointly observed or missing. This means that the imputation model for θ_1 and θ_2, assuming both will be imputed for all respondents, would be based on a set of predictors that includes task responses and covariates for one, but not the other, skill variable. If only Z_1, X_{11}, X_{12} are assessed and θ_1 as well as θ_2 are both imputed, then we will base our imputation model on the assumption that θ_2 and Z_2, X_{21}, X_{22} are conditionally independent given only Z_1, X_{11}, X_{12}. More concretely, if θ_1 represents reading proficiency and θ_2 is math proficiency, we would end up imputing math skills having only indicators of reading skills X_{11}, X_{12} and reading interest and self-reported reading practices Z_1 without either math skill indicators X_{21}, X_{22} or self-reported interest or practices in math.

This may seem a somewhat far-fetched case, but it is one that illustrates the extent to which assumptions are being made about what can be left out of the imputation model if important context variables are collected in a rotated data collection design.

Implications

The following condensed list of implications can be compiled from the discussion of the literature on missing data and imputation, as well as the discussion of the imputation model necessary for reporting proficiency data by means of PVs:

1. We have shown above that conditional independence is not a safe bet if it was established on a limited set of variables. More specifically, if X and Z_A are conditionally independent given Z_*, that is

$$P(X \mid Z_*) = P(X, Z_A, Z_*)$$

this does not imply that X and Z_A are conditionally independent given Z_*, Z_B. That is, it may be that

$$P(X \mid Z_*, Z_B) \neq P(X \mid Z_A, Z_*, Z_B) \quad \text{while } P(X \mid Z_*) = P(X, Z_A, Z_*).$$

This suggests that Z_A should NOT be left out of an imputation model if, by adding more predictors, the conditional independence established using a limited set of predictors will be violated. This is the basis for the commonly used imputation rationale that all available variables should be included in the predictive model used to impute missing observations in X.

2. In addition we have shown that whether or not a rotated questionnaire design is used, the imputation of proficiency variables θ is necessary for 100% of the sample. The proficiency variables θ are latent variables, since they are not directly observed. In all large-scale assessments, these variables are derived by means of a complex imputation model (called conditioning model or latent regression model) that contains an item response theory model and a latent regression. In that sense, even without rotations of the background questionnaire, the model used in large-scale assessments depends on accurately collected covariates for a successful imputation of proficiency values. When rotated background questionnaire sections are introduced, this approach is based on an even sparser foundation of observed variables, and thus becomes more vulnerable for potential bias in the estimates.
3. In PISA, the sparseness is further increased by the use of imputations of proficiency variables for which no item responses have been collected. In this assessment, students may take only a subset of the observed item response sets X_1,X_2,X_3 where X_i is the observation base for θ_i. Each student may only take booklets from two (say X_1, X_2) out of the three sets of X variables, and for the missing X_3 variable it is assumed that θ_3 can be imputed by assuming conditional independence of θ_3 and X_3 given X_1,X_2 and the background variables Z_A, Z_*, Z_{cB}. If background variables are rotated, the set of the Z variables is also incomplete, making the observed data for the imputation more sparse.
4. Following Little and Rubin (2002) we have shown that a case where nonoverlapping rotations are used requires strong assumptions

with respect to conditional independence of two variable sets given a third. These strong assumptions cannot be tested unless all variables are jointly observed (e.g., in a separate sample), so they have to be made without empirical evidence if only rotated questionnaires are used. The rotation design that uses K rotation blocks and administers $K–M$ of these two for each student also makes conditional independence assumptions when attempting to build an imputation model. More specifically, without a complete data matrix, the M blocks left out have to be assumed to be conditionally independent given the observed responses X and the $K–M$ blocks of background data.

5. A design that rotates test questions is already applied in the cognitive part. The X_1, X_2, X_3 variable sets are not completely observed and observations are collected in blocks of test items that are combined in a way that maximizes the association with the proficiency of interest across blocks. That is, each of the blocks within X_1 is targeted to assess the same underlying variable θ_1. Put another way, the blocks of test questions defined based on X_1 are designed so that each maximizes the correlation with the proficiency of interest, θ_1. The same holds for blocks based on variable sets X_2, X_3 that are designed to maximize statistical association with θ_2, θ_3. This follows the same rationale as pointed out in the discussion of conditional independence and is compatible with the common assumption made in IRT models. The idea is basically the same as the one put forward in Rässler et al. (2002): Define rotated sections of variables that minimize correlations within sections, while maximizing correlations across sections. Rässler et al. produce variable sets in which essentially each is a replicate of the other by ensuring that each variable Q gets a *sister variable* Q' in another rotated block that is highly associated with Q. Along the same lines, Graham et al. (1996, 2006) report that distributing response variables that belong to the scale across rotated sections "was clearly better from a statistical standpoint." While Graham (2006) amends this statement by saying that "the strategy that is better in a statistical sense turns out to be a big headache in a logistical sense: For every analysis, one must use the best FIML or multiple imputation procedures," he concludes that a recommendation to avoid this headache by keeping scales intact would "change as the missing data analysis procedures advance to the point of being able to handle large numbers of variables." In our view, this stage is reached by advances in computational statistics, latent variable modeling, and imputation methods. A split of variables belonging to the same scale across rotation sections would be possible in background questionnaire rotations as well, but it would require not relying on a file matching (or the relaxed versions with

multiple sets per respondent) design where different sets of variables end up in different rotations. Quite the contrary, the design suggestion put forward by Rässler et al. (2002) would discourage such an approach and would rather distribute indicators of the same underlying construct into different rotations, just as it is customary in the design of the cognitive part of the assessment since 2000 in PISA, 1995 in TIMSS, and 1983 or earlier in NAEP. These types of rotated designs can be handled by general latent variable models (e.g., Moustaki 2003; Rijmen et al. 2003; von Davier 2005, 2009).

Based on these implications it appears that rotations that violate the imputation rationale or are versions of the file-matching missing data design are not suitable data collection designs for the necessary imputation of proficiency variables. A design that balances the rotated sections in a way that each section contains indicators of the same underlying variables seems more in line with the need to minimize the use of un-testable conditional independence assumptions since it would, rather than kicking variables out of the imputation model, retain all variables by means of measurement of these variables with somewhat different sets of indicators in each rotation.

Summary

We hope to have provided enough evidence that a discussion is needed on how much we are willing to base our inferences on untested conditional independence assumptions. What are the unintended consequences if a faulty assumption was made with respect to conditional independence that leads to ignoring certain background variables in 50% or more of the imputed θ values? Recall that 100% of all proficiency values are imputed, typically with the use of item responses (cognitive indicators) X. Note, however, that these indicators sometimes may not be available as illustrated in the above example. In PISA, for example, some of the core domains of reading, mathematics, and science are not observed for all students, that is, some respondents may receive booklets that assess mathematics and science, but not reading. If imputations on all three domains are desired, the imputation of reading proficiency would be based on math and science item responses, and on the background data, of which both sets may be incomplete. This puts an even heavier burden on the imputation model (whichever is chosen) when some of the proficiencies are not only indirectly observed, but not observed at all and rather imputed by proxy through other correlated assessment domains and incomplete background data.

We have shown in this paper that the imputation of proficiencies in order to generate proficiency distributions in large-scale assessments relies heavily on the availability of reliable item response data and background questionnaire

data that are suitable for inclusion in a prediction model. Proficiency estimation in large-scale educational assessments use a complex imputation model based on an extension of IRT methods by means of a latent regression model, in which typically as many variables as possible are used for imputation in order to avoid potentially faulty conditional independence assumptions.

It was also shown by example that incomplete representation of the available data in the prediction model may implicitly make the false assumption of conditional independence, while conditional dependency may emerge once more when conditioning variables are included in the imputation model. If some variables are indeed conditionally independent of the proficiencies of interest, these could be left out of the imputation model, given the remainder of the background data. However, if that is the case and some of the background variables are indeed conditionally independent of the proficiency of interest, why should they be collected at all in an assessment that is focused on reporting distributions of proficiency and the relationships of proficiencies with policy-relevant background variables?

In addition, various options of defining an imputation model under incomplete background data and their underlying conditional independence assumptions were discussed. It was explained how either sample size would be decreased when using certain models, or how different imputation models with potentially conflicting predictions and assumptions would be obtained if respondents were pooled to augment and background data that were jointly observed in samples used to define the prediction model for imputation.

Also, prior research (ACER PISA document TAG(1204)6b) has shown that correlations between imputed values and background variables are reduced when incomplete background data are used. Jakubowski (2011) has presented results that show increased uncertainty under imputation with incomplete background data. Note that none of these reports provide strong evidence in favor of a general recommendation with respect to rotated designs. Both papers show the implications of a rotated design in a simulation based on real data from only a single cycle of PISA, and both studies show that associations between outcome variables and background variables are reduced under the rotated design, and/or that the uncertainty of comparisons is increased when introducing rotated designs.

At the very least, rotation schemes should use a similar design based on assessing constructs with incomplete representation of observables, as is already in place in the cognitive assessment (for a questionnaire design procedure that follows many of the design principles applied in cognitive parts of large-scale assessments, see Rässler et al. 2002). Leaving constructs completely out of the design in some rotations, as in the relaxed file matching design, should be avoided because strong and often untestable assumptions have to be made (Little and Rubin 2002), and different rotations would then include the same constructs, but be assessed with different subsets of observed response data.

Finally, some results obtained in studies on mass imputation for survey data indicate that, depending on the imputation model and the selection of variables, correlations of variables that involve imputations can be biased upward, downward, or almost unaffected (Fetter 2001). This leaves the question of whether any decision rule can be devised with respect to which imputation model to use and what variables to include in cases where substantial amounts of missingness are present due to incomplete collection designs.

Because the imputations from large-scale assessments enter the public use files as one of the main data products in the form of PVs required for online reporting tools and secondary analysis, we believe that a careful and diligent reevaluation of the use of incomplete background data collections is in order. The complexity of the models used to generate these data products, and the sparseness of the item response data collected, warrants more thorough examination of the risks involved when using incomplete background data for mass imputation of proficiency data.

References

Adams, R. J., Wu, M. L., and Carstensen, C. H. 2007. Application of multivariate Rasch models in international large scale educational assessment. In M. von Davier and C. H. Carstensen (Eds.) *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications* (pp. 271–280). New York, NY: Springer-Verlag.

Bock, R. D. and Zimowski, M. F. 1997. Multiple group IRT. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*. New York, NY: Springer-Verlag, 433–448.

Collins, L. M., Schafer, J. L., and Kam, C. M. 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6, 330–351.

Fetter, M. 2001. Mass imputation of agricultural economic data missing by design: A simulation study of two regression based techniques. *Federal Conference on Survey Methodology*. http://www.fcsm.gov/01papers/Fetter.pdf, downloaded 07/23/2012.

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

Graham, J. W., Hofer, S. M., and MacKinnon, D. P. 1996. Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, 31, 197–218.

Graham, J. W., Taylor, B. J., Olchowski, A. E., and Cumsille, P. E. 2006. Planned missing data designs in psychological research. *Psychological Methods*, 11(4), 323–343.

Greene, W. H. 2003. *Econometric Analysis*. 5th ed. New Jersey: Prentice-Hall.

Haberman, S. J. 1974. *The Analysis of Frequency Data*. Chicago: University of Chicago Press.

Haberman, S. J., von Davier, M., and Lee, Y. 2008. Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions. RR-08-45. ETS Research Report.

Hanushek, Eric A. and John E. Jackson. 1977. *Statistical Methods for Social Scientists*. New York: Academic Press, Inc.

Haslett, S. J., Jones, G., Noble, A. D., and Ballas, D. 2010. More for less? Comparing small-area estimation, spatial microsimulation, and mass imputation. Paper presented at JSM 2010.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., and Junker, B. W. 1996. Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika*, 61, 679–693.

Jakubowski, M. 2011. Implications of the student background questionnaire rotation on secondary analysis of PISA data. Unpublished manuscript.

Jones, M. P. 1996. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *J Am Stat Assn*. 1996; 91:222–230.

Lazarsfeld, P. F. and Henry, N. W. 1968. *Latent Structure Analysis*. Boston: Houghton Mifflin.

Little, R. J. A. 1988. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 63, 287–296.

Little, R. J. A. and Rubin, D. B. 2002. *Statistical Analysis With Missing Data* (2nd edition). New York, NY: Wiley.

Lord, F. M. and Novick, M. R. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Mislevy, R. J. 1991. Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177–196.

Moustaki, I. and Knott, M. 2000. Weighting for item non-response in attitude scales using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A*, 163(3), 445–459.

Moustaki, I. 2003. A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology*, 56(2), 337–357.

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

Piesse, A. and Kalton, G. 2009. A Strategy for Handling Missing Data in the Longitudinal Study of Young People in England (LSYPE), Westat Research Report No. DCSF-RW086.

PISA TAG(1204)6b. 2012. On the use of rotated context questionnaires in conjunction with multilevel item response models.

Rässler, S., Koller, F., and Mäenpää, C. 2002. A Split Questionnaire Survey Design applied to German Media and Consumer Surveys. *Proceedings of the International Conference on Improving Surveys, ICIS 2002*, Copenhagen.

Rijmen, F., Tuerlinckx, F., De Boeck, P., and Kuppens, P. 2003. A nonlinear mixed model framework for item response theory. *Psychological Methods*, 8, 185–205.

Rose, N., von Davier, M., and Xu, X. 2010. *Modeling Non-Ignorable Missing Data with IRT.* ETS-RR-10-10. Princeton: ETS Research Report Series.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons.

Rubin, D. B. 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473–489.

Rutkowski, L. 2011. The impact of missing background data on subpopulation estimation. *Journal of Educational Measurement*, 48(3), 293–312.

Suppes, P. and Zanotti, M. 1995. *Foundations of Probability with Applications: Selected Papers*, pp. 1974–1995. Cambridge: Cambridge University Press.

van Buuren, S. and Groothuis-Oudshoorn, K. 2011. MICE: Multivariate Imputation by Chained equations in R. *Journal of Statistical Software*, 45(3), 1–67.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., and Sijtsma, K. 2008. Multiple imputation of categorical data using latent class analysis. *Sociological Methodology*, 33, 369–297.

von Davier, M. 2005. A general diagnostic model applied to language testing data (ETS Research Report No. RR-05-16). Princeton, NJ: ETS.

von Davier, M. 2009. Mixture distribution item response theory, latent class analysis, and diagnostic mixture models. In: S. Embretson (Ed.), *Measuring Psychological Constructs: Advances in Model-Based Approaches*. ISBN: 978-1-4338-0691-9. pp. 11–34. Washington, DC: APA Press.

von Davier, M., Sinharay, S., Oranje, A., and Beaton, A. 2006 Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao and S. Sinharay (Eds.), *Handbook of Statistics (vol. 26): Psychometrics*. Amsterdam: Elsevier.

von Davier, M., Gonzalez, E., and Mislevy, R. 2009. What are plausible values and why are they useful? In M. von Davier and D. Hastedt (Eds.), *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments (Vol. 2)*. IEA-ETS Research Institute.

Xu, X. and von Davier, M. 2008. Comparing multiple-group multinomial loglinear models for multidimensional skill distributions in the general diagnostic model. RR-08-35, ETS Research Report.

Zermelo, E. 1929. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. (The calculation of tournament results as a maximum problem of probability calculus: in German). *Mathematische Zeitschrift*, 29, 436–460.

This page intentionally left blank