Simple and Effective Methods to Treat Missing Item Responses

Mark Huisman *†

Abstract

In this paper imputation models to impute missing responses to test items are presented. In a first simulation study, the performance of some simple techniques that are easy to implement are investigated (item and person mean substitution, hot deck imputation). Improvements of these techniques lead via imputation of a corrected item mean to imputation models from item response theory (IRT). Specifically, the one parameter logistic model (OPLM) and the model proposed by Mokken are used to impute missing item responses. Their performance is investigated in a second simulation study. The parametric OPLM allows multiply imputing the missing data, and a multiple imputation procedure is presented. Also the performance of hot deck imputation within adjustment cells is investigated. Adjustment cells are related to poststratification and are used to correct for nonrandom missingness.

Key words: missing item responses, imputation, corrected item mean, item response theory, nonignorable nonresponse.

*Mark Huisman, Department of Statistics, Measurement Theory, & Information Technology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, the Netherlands, telephone +31 50 3636193, e-mail: M.Huisman@ppsw.rug.nl.

†This research was supported by the Netherlands Research Council (NWO), Grant 575-67-048. The author thanks Ivo Molenaar for suggested improvements on the design of the study and comments on earlier versions of the paper, as well as the reviewer for his helpful comments.

1 Introduction

Among the wide variety of procedures to handle missing data, imputing the missing values is a popular strategy to deal with missing item responses. With imputation procedures estimates of the missing values are obtained to fill-in the blanks in the data set. Because these techniques result in completed data and standard statistical techniques can be used, the researcher is apt to forget that some values are missing. As a consequence, the danger that standard estimators are substantially biased due to nonrandom missing data mechanisms is easily forgotten. Naive imputations can even be worse than doing nothing, so care is needed while imputing missing values (see Little, 1988).

In this paper the performance of some imputation techniques is investigated. The techniques are used to handle missing responses to some kind of test or scale. These tests are often used in behavioral sciences, and consist of items which measure a latent trait of individuals (like emotional well-being or some aspect of a person's personality). In Section 2, the topic of missing responses to test items and imputation is discussed. In the third section, the design and results of a simulation study are presented, in which the performance of some simple imputation techniques is investigated. Potential improvements of these techniques are presented in Section 4. These improvements include imputing with models from item response theory (IRT), especially the one parameter logistic model (Verhelst & Glas, 1995) and the model of Mokken (1971). Also methods which (partially) adjust for nonrandom missingness mechanisms are investigated. In a second simulation study the performance of these techniques is investigated. In Section 5 the results of the two simulation studies are summarized and discussed.

2 Missing item responses

In the behavioral sciences, inferences about a latent property of persons are usually made by analyzing the responses of these persons to a set of items. Each single item does not cover all aspects of the latent trait, and a person's position on the trait can only be inferred indirectly by investigating the responses to all items in the scale. With the help of measurement models, the location of a person on the latent trait is determined. Frequently used models are factor analytic models, models based on classical test theory, and IRT models. IRT models have the advantage that latent abilities, θ, can be estimated using incomplete testing designs. In such designs different subsets of items are administered to different

subgroups of respondents, which means that the researcher specifies beforehand the mechanism that causes the incompleteness of the item responses. For a detailed discussion of incomplete testing designs and ignorability of the generating missingness mechanisms see Eggen & Verhelst (1992) and Mislevy & Wu (1996).

If the data are not missing by design, inferences for person abilities are only valid in the case of ignorable nonresponse. Mislevy & Wu (1996) show that this is generally not true when respondents decide for themselves not to respond to a presented item. In this case a suitable model for the missingness process should be included in the analysis to prevent the inferences to become biased (Greenlees, Reece, & Zieschang, 1982). Finding such a model is a difficult task.

The data used in this paper to investigate the performance of some imputation techniques consists of responses to items $\boldsymbol{X} = [x_{vi}]$ ($v = 1, \cdots, n$ respondents, $i = 1, \cdots, k$ items) and some covariates $\boldsymbol{Z} = [z_{vh}]$ ($h = 1, \cdots, q$). The k items form one scale measuring a latent property of the respondents. All items have a fixed number of ordered response categories, and the weighted sum of the item responses can be used as an estimate of the latent ability θ_v of person v:

$$r_v = \sum_{i=1}^{k} w_i x_{vi} \,,$$

where w_i is the weight of item i. In the sequel it is assumed that the covariates are completely observed and the missingness only occurs in the item responses. The missingness mechanism can be modeled with an indicator matrix $\boldsymbol{M} = [m_{vi}]$, where $m_{vi} = 0$ if person v responded to item i, and $m_{vi} = 1$ otherwise.

Two specific IRT models are used in this paper, i.e., the one parameter logistic model (OPLM, Verhelst & Glas, 1995) and the model proposed by Mokken (1971). In the former model, being a generalization of the well-known Rasch model, the sum score, r_v, is a sufficient statistic for the latent ability θ_v. In the Mokken model, r_v itself is used as estimate of the ability. Aspects of the quality of measurement can be assessed via Cronbach's *alpha*, the classical concept of reliability (see Lord & Novick, 1968) and Loevinger's H-coefficient, a measure of the scalability of items used in Mokken scaling (see Mokken, 1971; Molenaar 1991).

2.1 Imputation

With imputation techniques estimates of the missing item responses are made and substituted for the missing entries. Sande (1982) extensively discussed the

problems an imputer is faced with, and concluded that a procedure is needed which 1) will impute plausibly and consistently with the edits, 2) will reduce the bias and preserve the relationship between the items as far as possible, 3) will work for (almost) any pattern of missing items, 4) can be set up ahead of time, and 5) can be evaluated in terms of impact on the bias and precision of the estimates. She states that "*particular techniques of imputation vary in their ability to meet these requirements*" (p. 147).

There are a number of different ways in which a researcher may want to impute the missing values. A first distinction which can be made is that between naive, ad hoc methods and more principled ones. Although the naive methods are quick options (e.g., unconditional mean imputation), they often lead to biased results. More principled approaches use models for both the observed and missing data to reduce the bias. This leads to a second distinction, that between explicit and implicit imputation models. Explicit models are usually parametric models used in mathematical statistics, e.g., linear regression. Implicit models are models that underlie procedures for fixing up data structures in practice and often have a nonparametric flavor, e.g., hot deck techniques that use donor cases to impute missing responses. A combination of both kinds of models is predictive mean matching (Little, 1988), where an explicit model is used to find a suitable donor case of which the observed values are used to impute the missing data (implicit method).

A last distinction is that between deterministic and stochastic techniques. In the first group of procedures, imputed values are uniquely determined and result in identical estimates when repeated. Stochastics methods use some kind of randomization process to impute missing values, and therefore reduce the bias due to the overestimation of precision by deterministic methods.

3 Simulation study I

In this paper the results of two simulation studies are presented, with which the performance of imputation techniques for imputing missing answers to test items is investigated. Both simulation studies are also presented by Huisman (1999), to which the reader is referred to for a more detailed description of the design and results. In the first study, techniques belonging to the group of naive, implicit, and deterministic methods are investigated. Six of these techniques are presented here. Some of the techniques produce noninteger imputed values. These are rounded to the nearest integer to create completed data consistent with the edits.

3.1 Simple techniques

Random Draw Substitution (RDS). This ad hoc method replaces a missing value with a random draw from the permitted response options. This method is not particularly recommended and will only be used as negative benchmark for the performance of other techniques.

Mean Substitution. When dealing with item responses, there are several possibilities to impute a mean. The first possibility is *item mean substitution* (IMS), where the mean item score is substituted for every missing value of a particular item. A second method is *person mean substitution* (PMS). Here the mean scale score over the observed items is used to impute missing values of a person. It should be noted that replacing missing values by a mean value causes the scores of the respondents to move towards the center of the distribution. Variances and covariances are therefore systematically underestimated. There are some remedies for this, like adjusting the degrees of freedom, or adding a small random quantity (see e.g. Little & Rubin, 1987, or Bello, 1993).

Corrected Item Mean substitution (CIM). In this third mean substitution method, missing values are replaced by an item mean which is corrected for the ability of the individual:

$$\mathrm{CIM}_{vi} = w_v \bar{x}_{.i}^{(i)} = \left(\frac{\sum_{h \in obs(v)} x_{vh}}{\sum_{h \in obs(v)} \bar{x}_{.h}^{(h)}} \right) \bar{x}_{.i}^{(i)} = \left(\frac{k^{(v)}\ \mathrm{PMS}_v}{\sum_{h \in obs(v)} \mathrm{IMS}_h} \right) \mathrm{IMS}_i \,,$$

where $\bar{x}_{.i}^{(i)}$ is the mean score on item i for the nonmissing cases, $obs(v)$ is the collection of observed items and $k^{(v)}$ the number of observed items for person v. In this combination of both item and person mean, the item mean is multiplied with a weight which equals the ratio of the sum of the observed items of a person and the sum of the item means of the same set of items as observed for that individual. CIM can therefore be regarded as a very simple IRT (imputation) model.

Item Substitution. Instead of imputing the mean scale score of a person, an observed response of this person to another item in the scale can also be used to substitute a missing value. The method *interitem correlation substitution*

(ICS) replaces a missing value by the observed response on another item for which the interitem correlation with the missing item is highest. The correlation matrix is computed from the complete cases. This can result in biased estimates of the correlations, but only the order of the correlations is used here, which probably will not be affected too much even if there is a considerable amount of nonrandomly missing data.

Hot Deck imputation (HD). Hot deck imputation techniques use a completely observed donor case for the imputation of an incomplete case (see e.g. Sande, 1983). The missing values are replaced by the corresponding values of a 'nearest neighbour' which most resembles the incomplete case with respect to the observed items. This donor is found by minimizing the distance function

$$d_{v,v'} = \sum_{i \in obs(v)} (x_{vi} - x_{v'i})^2 ,$$

where v is the incomplete and v' a complete case. The case v' for which $d_{v,v'}$ is minimized is used as donor case. When several complete cases are at the same minimal distance of the currently considered incomplete case, the complete case which is nearest to the incomplete case with respect to its position in the data matrix is used as a donor case.

3.2 Simulation design

Four independent factors are used in the simulation study, i.e., data set (d), sample size (n), missing data mechanism (m), and proportion missing values (p). With these factors, incomplete data matrices are generated which will be imputed repeatedly (100 replications). For more details see Huisman (1999).

Data (d) The data come from empirical data sets from the behavioral sciences. The original data sets consist of a large number of cases from which a fixed number n of complete cases is randomly drawn. Data encountered in actual field research are realistic and may therefore provide a better picture of the accurateness and effectiveness of the imputation techniques than simulations based on data generated with some theoretical distribution (see also Kromrey & Hines, 1994). Table 1 presents the four data sets used in this study.

Sample size (n) From the data sets introduced in Table 1, samples of n = 100, 200, and 400 cases are drawn in which missing values are generated.

Table 1: *Data sets serving as basis for data used in simulation study I (names, scale, number of items (k), number of response options (Cat.), item weight (w_i), average interitem correlation (Corr.), and Cronbach's alpha).*

Data	Scale	k	Cat.	w_i	Corr.	*alpha*
*d*1 FFPI(E)	Five Factor Personality Inventory: Extraversion (Hendriks, 1997)	20	1 – 5	1.25	0.39	0.93
*d*2 RAND(PF)	RAND-36 Item Health Survey: Physical Functioning (van der Zee & Sanderman, 1993)	10	0 – 2	5	0.61	0.94
*d*3 RAND(MH)	RAND-36 Item Health Survey: Mental Health (van der Zee & Sanderman, 1993)	5	0 – 5	4	0.58	0.85
*d*4 NHPR(SI)	Nottingham Health Profile: Social Isolation (Hunt, McKenna, & McEwen, 1993)	5	0 – 1	20[a]	0.41	0.74

[a] *Average value of item weights.*

Missing data mechanism (m) The mechanisms used to create the missing data are labeled:

($m1$) MCAR—Missing Completely At Random. The probability of response of person v on item i is a random number between 0 and 1.

($m2$) NRX—NonRandomness depending on $\boldsymbol{X}$. The probability of response of person v on item i is a logistic function of the scale score r_v and the mean item score $\bar{x}_i$:

$$P(M_{vi} = 0 \mid r_v, \bar{x}_{.i}) = \frac{\exp(\alpha_0 + \alpha_1 r_v + \delta \bar{x}_{.i})}{1 + \exp(\alpha_0 + \alpha_1 r_v + \delta \bar{x}_{.i})},$$

where $M_{vi} = 0$ indicates an observed response and α_0, α_1 and δ are scalar parameters.

($m3$) NRXZ—NonRandomness depending on $\boldsymbol{X}$ and $\boldsymbol{Z}$. The probability of response depends on the scale score r_v, the mean item score $\bar{x}_{.i}$, and the covariates sex, z_{1v}, and age, z_{2v}, in the same way as in mechanism NRX.

An observation is classified missing ($M_{vi} = 1$) when the probability of response is small, i.e., smaller than a randomly drawn number from a uniform distribution.

Proportion missing values (p) The parameters in the logistic functions, in case of nonrandomly missing data, were set at specified levels (see Huisman 1999) to create three proportions of missing values: $p = 0.05$, 0.12, and 0.20. These are the proportions of missing cells in the data matrices.

3.3 Performance of the imputation techniques

An imputation technique performs well if it is able to obtain unbiased estimates of missing values. However, more important is the ability of preserving the relationship among items and the reducing the bias caused by the missing data (Sande, 1982). The effectiveness of an imputation technique should be evaluated against a criterion that is common in applied research (see Kromrey & Hines, 1994). Because the latent ability of the respondents is the topic of interest, the performance of the imputation techniques is investigated by comparing the scale scores after imputation with the original scores before data points were deleted. To judge the ability of the techniques to preserve the relationships between the items, two overall measures of these relations are investigated: Cronbach's *alpha* and Loevinger's H.

Specifically, the distribution of the deviation $d_v(t) = r_v(t) - r_v$ is used to judge the performance of the imputation techniques, where r_v is the scale score of person v in the original complete data and $r_v(t)$ the scale score of person v from the data matrix imputed with technique t. This distribution is summarized by the mean $\bar{d}(t)$, the standard deviation $s_{d(t)}$, and the *Root Mean-Squared Deviation*:

$$\text{RMSD}(t) = \sqrt{\frac{\sum_{v=1}^{n_{mis}} d_v(t)^2}{n_{mis}}} = \sqrt{\frac{\sum_{v=1}^{n_{mis}} \left(\sum_{i \in mis(v)} w_i (x_{vi}(t) - x_{vi}) \right)^2}{n_{mis}}},$$

where $x_{vi}(t)$ is the imputed value for person v and item i, $mis(v)$ is the collection of missing item responses for person v, and n_{mis} is the number of persons with missing data in the data matrix.

From this definition it follows that the RMSD(t) is dependent on the item weights w_i. These weights are used to transform the scale scores of the four scales to have equal range, i.e., $0 - 100$. The item weights equal $\frac{100}{k \times max}$, with k the number of items and max the difference between the largest and smallest response option[‡] (see Table 1). This results in a change in the interpretation of

[‡]For the NHPR(SI) this only holds for the average item weights

the RMSD(t), reflecting that 1) incorrectly imputing an item in a short scale (k small) causes more bias in $r_v(t)$ than an item in a long scale, and 2) incorrectly imputing an item with few response options (max small) causes more bias in $r_v(t)$ than an item with many response options. Incorrectly imputing an item of the FFPI(E), for instance, causes a change in $r_v(t)$ of 1.25 when the difference between the imputed value and the original value equals 1, while the same error in the NHPR(SI) causes on average a change in r_v of 20.

On scale level, the two measures of internal consistency of a test are investigated which are mentioned earlier: Cronbach's *alpha*, (reliability) and Loevinger's H-coefficient (scalability). Both measures are compared before and after deletion of data points and imputation, $d_a(t) = alpha(t) - alpha$ and $d_H(t) = H(t) - H$. These measures of internal consistency heavily depend on the covariance matrix of the items, which means that bias in estimating variances and covariances caused by imputation will lead to biases in *alpha* and H.

3.4 Results

3.4.1 Recovering r_v

The simulation design results in 108 cells ($d \times n \times m \times p$), in each of which 100 data matrices with missing data are generated and imputed with the six imputation techniques. The main results of the simulations are presented in Table 2.

In Table 2 the average values of RMSD(t) of the imputation techniques are presented for all factors of the design. From the table it follows that across all factors CIM is the best technique. For each independent variable separately, CIM also performs best, closely followed by ICS, PMS, and HD in varying order. The benchmark method RDS performs worst, as was expected. Imputing an item mean (IMS) is on average always worse than imputing a person mean (PMS), although the RMSD(PMS) more rapidly increases than that of IMS when the number of missing values increases. In many cases IMS is almost as bad as RDS, and sometimes even worse, especially for nonrandomly missing data.

The same picture emerges when the means and standard deviations of $d_v(t)$ are inspected (results not presented here). The former expresses how well a technique on average is able to recover the scale score, the latter the dispersion of the errors made by imputation. CIM and ICS both have the smallest values for $\bar{d}_v(t)$. IMS has the largest values, even larger (absolute values) than RDS which has largely negative values, indicating that on average the imputed values are smaller than the original. The other techniques all show positive values for

Table 2: *Main effects of RMSD(t) for the imputation techniques across all factors in simulations study I.*

		RDS	IMS	PMS	CIM	ICS	HD
	all	11.96	11.17	8.37	7.17	8.05	8.82
FFPI(E)	$d1$	4.51	4.23	2.57	2.26	3.26	3.57
RAND(PF)	$d2$	11.52	11.27	7.40	5.62	5.94	7.21
RAND(MH)	$d3$	12.66	10.57	8.29	6.46	8.36	8.26
NHPR(SI)	$d4$	19.14	18.61	15.23	14.33	14.63	16.26
$n = 100$	$n1$	11.89	11.09	8.28	7.09	8.16	9.02
$n = 200$	$n2$	11.98	11.18	8.39	7.18	8.07	8.79
$n = 400$	$n3$	12.01	11.23	8.44	7.23	7.91	8.66
MCAR	$m1$	10.87	6.11	5.13	4.58	5.37	5.44
NRX	$m2$	11.52	12.89	8.98	7.68	8.75	10.53
NRXZ	$m3$	13.49	14.51	11.01	9.25	10.02	10.50
$p = 0.05$	$p1$	10.52	10.34	7.26	6.19	6.79	7.31
$p = 0.12$	$p2$	11.85	11.19	8.45	7.25	8.01	8.79
$p = 0.20$	$p3$	13.51	11.98	9.40	8.06	9.34	10.36

$\bar{d}(t)$, and only when the missing data mechanism is random ($m1$), the deviations are close to zero.

When looking at the effect of the four factors the following can be said. First, the differences between the performances in the four data sets clearly show the influence of the item weights on $r_v(t)$. This reflects the 'punishment' of incorrectly imputing a short scale ($d3$ and $d4$), or a scale with few response categories ($d2$ and $d4$). Second, sample size has a marginal effect on the performance of an imputation technique, except for the hot deck method. Third, the higher the percentage missing, the larger the number of errors. Finally, the more complicated the missing data mechanism, the more values are incorrectly imputed.

Interactions When looking at three-way interactions $d \times m \times p$ (not reported here) CIM also proves to be the best method, in almost every cell. Competitors are ICS (in the RAND(PF) and NHPR(SI) data), PMS (FFPI(E) data), but they are almost never better than CIM (only in some cases with small proportions missing). In situations with medium to large proportions of nonrandomly missing data in the long scales, RDS performs better than IMS.

3.4.2 Assessment of scale quality

To assess the ability of the imputation techniques to preserve the relationships between the items, the change in Cronbach's *alpha* and Loevinger's H is investigated. For a detailed description of the results see Huisman (1999), here only the main results are presented.

The most noticeable result is the overestimation of both *alpha* and H by the techniques PMS and CIM. ICS only overestimates the scale quality in the short scales, the other three techniques lead to lower values of *alpha* and H. The overestimation by CIM is not surprising, because this technique was called earlier a very simple IRT model. When such a model is used for both the imputation of missing values and the analysis of the imputed data, the model fit will increase. *Alpha* and H can be seen as goodness-of-fit measures of measurement models, and will therefore be overestimated when missing values are imputed with an IRT model.

There is no overall best method according to the two criteria, although ICS seems to do reasonable well. Hot deck imputation underestimates the quality of the scale, but the absolute deviates are larger than those for CIM. The RDS procedure performs worst in practically all situations. Only in case of dichotomous, nonignorable missing items [NHPR(SI)], IMS is worse.

The deviations in *alpha* are largest for the short scales. They can be fairly large, for example, in case of m = NRX, $p = 0.20$, and d = NHPR(SI), *alpha* equals 0.29 after imputation of item means (IMS), where in the original data *alpha* is 0.74. The deviations in H are generally larger than those in *alpha*, except in the short scales where they are of equal size. Longer scales are more robust against imputation according to Cronbach's *alpha* (the absolute deviations are never larger than 0.15, and often almost zero). Because of the dependence of *alpha* on k this is not surprising. This positive effect of test length does not occur for the H-coefficient. However, the overestimation of H by CIM and PMS is much smaller compared to *alpha*, or does not even occur in some cases.

4 Simulation study II

The techniques investigated in the first simulation study were characterized earlier as being naive and simple methods. There are three ways in which the quality of the imputation of missing item responses can be improved: 1) adjusting for

nonrandom missing data mechanisms, 2) using more advanced imputation models, and 3) both. The first two potential improvements are studied in this second simulation study. The adjustments for nonrandom missingness are based on imputation within so-called adjustment cells (poststratification), the more advanced imputation models are item response models, specifically, the model of Mokken and the OPLM. With this latter model also a multiple imputation procedure is constructed. For a more detailed presentation of the design and results of this study, see Huisman (1999).

4.1 Adjustment cells

Adjustment cells are strata based on completely observed covariates, and contain cases with both observed and missing item responses. Within each adjustment cell, response is assumed to be independent of the items and the sampling selection process, and the subpopulation is assumed to be homogeneous with respect to the missingness mechanism. In this way the procedure is related to poststratification. For example, if age is known to influence the probability of response and age is a covariate that is observed for every person, age groups can be formed within which missing item responses are imputed. This conditioning on observed covariates reduces nonresponse bias in case of data Missing at Random (MAR) and also to some degree in case of nonignorable nonresponse as long as nonresponse is limited and good covariate information is available (Rubin, Stern, & Vehovar, 1995).

Corrected Item Mean substitution (CIMA). Within each adjustment cell the corrected item mean, as defined in the first study, is imputed for every missing item response. The values of CIM are rounded to the nearest integer.

Hot Deck imputation (HDA). Within adjustment cells a completely observed donor case is sought by minimizing the distance function given earlier. The response pattern of the complete case that most resembles the incomplete case (nearest neighbour) is used as donor, and the observed values are substituted for the missing ones.

4.2 Item response models

As was discussed earlier, all imputations are based on some kind of model, either implicit or explicit (see also Sande, 1982). Examples of implicit models are given earlier (hot deck or mean imputation). Explicit models are the parametric

models usually used in statistical analysis, like regression. Because imputation is a form of prediction, imputations should be based on the predictive distribution of the missing values given the observed values, and Little (1988) argues that "*a model underlies this distribution, so a systematic approach to imputation requires modeling the data*" (p. 288).

The obvious class of models to use for imputation of missing item responses in the case of measurement of latent abilities, is the class of measurement models. In particular IRT models seem most appropriate, because of their benefits with incomplete testing designs (Eggen & Verhelst, 1992; Mislevy & Wu, 1996), and as a continuation of the procedure CIM from the first study, which can be viewed as a very simple IRT model.

Mokken scaling (MOK). In Mokken scale analysis the cumulative response mechanism of the scale is used to order the persons and the items on the latent trait θ (Mokken, 1971). Laros & Tellegen (1991) use Mokken scaling to impute missing item responses in an adaptive testing situation. The idea is that the items should be ordered according the percentage correct responses (decreasing), from 'easy' to 'difficult'. Rules are made for imputing the missing responses with as few errors as possible. The steps in the procedure are the following:

1. Compute the proportion of correct responses for every item based on the available cases. In case of polytomous items the responses should be transformed into dichotomous item steps (Molenaar, 1997). These item steps represent imaginary thresholds between two adjacent response options, where the value 1 indicates that the respondent crossed the threshold, and 0 if not.
2. For every missing data entry the following five rules are applied:
   2.1. If the response 1 (correct) follows the missing response 9, impute the value 1.
   2.2. If not, then if the response 0 (incorrect) precedes the missing response 9, impute the value 0.
   2.3. If not, then define a_{00} as the number of incorrect responses (0's) preceding a missing response, and a_{01} as the number of correct responses (1's) preceding a missing response. If $a_{00} \geq a_{01}$ impute the value 0 (incorrect).

2.4. If not, then define a_{10} as the number of incorrect responses (0's) following a missing response, and a_{11} as the number correct responses (1's) following a missing response. If $a_{10} \leq a_{11}$ impute the value 1 (correct).

2.5. In all other cases impute a random draw from the empirical distribution of the dichotomous items, based on their proportion correct.

3. After imputation, transform the item steps into the original response options in case of polytomous items.

Example: Imputation steps
The items are ordered according to decreasing proportion correct.

9 1 1 1 1 0 9 1 0 0	$\Rightarrow$					1 1 1 1 1 0 1 1 0 0
1 1 1 1 0 0 9 1 0 9	$\Rightarrow$					1 1 1 1 0 0 1 1 0 0
1 0 0 0 0 1 9 9 0 0	$\Rightarrow$	$a_{00} = 4$	$\geq$	$a_{01} = 2$	$\Rightarrow$	1 0 0 0 0 1 0 0 0 0
1 1 0 1 9 9 0 1 1 0	$\Rightarrow$	$a_{10} = 2$	$\leq$	$a_{11} = 2$	$\Rightarrow$	1 1 0 1 1 1 0 1 1 0

Note that the rules, especially 2.3 and 2.4, are somewhat arbitrary and other rules could be formulated. However, the rules used here perform reasonably well, because they keep the number of errors small.

The One Parameter Logistic Model—OPMISS. With the program OPLM (Verhelst, Glas, & Verstralen, 1995) the one parameter logistic model can be fitted to the data set. Based on the observed responses, estimates of the item and person locations on the latent trait θ can be computed. With these parameter estimates, the distribution of the responses in every cell (v, i) of the data matrix can be computed; $\hat{\pi}_{ij}(\hat{\theta}_v)$ is the estimated probability of person v giving the response option j for item i. An ad hoc module called OPMISS was added to OPLM to impute missing item responses with the estimated response probabilities (see Nap, 1994a).

With OPMISS different imputation techniques can be used. For instance, the expected value of the distribution or a random draw from the distribution can be imputed. Nap (1994a, 1994b) showed in several simulation studies that the best results were obtained when the random draw option was used. Two versions of this procedure are investigated here: 1) replace every missing value once by a draw from the estimated response distribution, and 2) multiple imputation of each data set by repeated draws from the distribution (Rubin, 1987).

OPMISS—Expected Value substitution (OEV). The expected response of person v on item i is substituted for the missing value,

$$\text{OEV}_{vi} = \sum_{j=0}^{c} j\hat{\pi}_{ij}(\hat{\theta}_v) ,$$

where j represents the response categories with values $0, \cdots, c$ (all items in a scale are assumed to have the same number of response options). The values OEV_{vi} are rounded to the nearest integer.

OPMISS—Single Draw substitution (OSD). The values used for the imputation of the missing item responses are randomly drawn from the estimated distribution of each cell, with probabilities $\hat{\pi}_{i0}(\hat{\theta}_v), \cdots, \hat{\pi}_{ic}(\hat{\theta}_v)$. Every missing entry is imputed only once.

OPMISS—Multiple Draws substitution (OMD). For every incomplete data matrix the OSD procedure is repeated five times, i.e., every missing entry is imputed five times. This results in five completed data matrices. Each imputed data matrix is investigated separately with the criteria that will be defined in the next section to judge the performance of the imputation procedures. The five results are combined to represent the overall result of this multiple imputation technique.

4.3 Simulation design

The design of this second simulation study is almost the same as that of the first one. There are two important differences. First, due to technical problems the data sets RAND(MH) and FFPI(E) could not be imputed with OPMISS (see Huisman, 1999). The items of first data set are recoded (less response options) to solve this problem, and the latter data set is replaced by an artificial one. Moreover, the artificial data, which is generated with the OPLM to follow a Rasch scale with equal discrimination parameters, allows comparison of the performance of the OPMISS procedures under ideal circumstances. These two new data sets can be found in Table 3. The data sets used in the second simulation study are RAND(PF), RAND(MH)-recoded, NHPR(SI), and RM(D) (or $d2$, $d3r$, $d4$, and $d5$, respectively; see also Table 1).

The second difference is the number of cases drawn from the data sets; instead of 100, 200, and 400 cases, here $n = 200$, 400, and 800 cases are drawn in which missing data are created. The other two independent factors are the same as

Table 3: *Extra data sets serving as basis for data used in simulation study II.*

Data	Scale	k	Cat.	w_i	Corr.	*Alpha*
d3r RAND(MH)	RAND-36 Item Health Survey: Mental Health recoded	5	$0-3$	6.67	0.59	0.87
d5 RM(D)	Data following Rasch Model: Dichotomous items	10	$0-1$	10	0.14	0.63
RM(D) item parameters: -2.0, -1.7, -0.9, -0.5, -0.1, 0.2, 0.6, 1.1, 1.5, 1.8						

in the first study: missingness mechanisms (m) MCAR, NRX, and NRXZ, and proportions missing (p) 0.05, 0.12, and 0.20. The performance of the techniques is investigated by inspecting RMSD(t), and the differences in Cronbach's *alpha* and Loevinger's H: $d_a(t)$ and $d_H(t)$.

Adjustment cells Based on the covariates sex and age, adjustment cells are created in every data set. Because age is a continuous variable, categories are made such that the data are split into two groups (young/old). This means that four adjustment cells are made: two sex groups and two age groups. Huisman (1999) also presents the results of a more refined categorization, with two sex groups and four age groups. These results are, however, never better than the rough categorization with two age groups presented here.

4.4 Results

4.4.1 Recovering r_v

The design of the second study results in 108 cells ($d \times n \times m \times p$). Due to technical problems, an amount of 100 incomplete data matrices could not be obtained for every cell of the design (see Huisman, 1999). However, in every cell at least 50 incomplete data matrices are generated and imputed (in most cells the number is close to or equals 100). The main results of the simulation are presented in Table 4.

In Table 4 the average values of RMSD(t) are presented for all factors of the design. It follows that across all factors OMD performs best when estimating the scale score. The second best after OMD is OEV, except for the RM(D) data where CIMA is better. More specific, in 83 out of 108 cells in the design OMD was the best technique and in 20 cells it was second best. In 62 cells OEV was the second best technique. The third technique using the OPLM, OSD, does not

Table 4: *Main effects of RMSD(t) for the imputation techniques across all factors in simulation study II. The data sets are ordered according to the item weights, reflecting the difficulty to impute the data.*

		CIMA	HDA	MOK	OEV	OSD	OMD
	all	8.93	11.58	9.15	8.60	9.27	8.28
RAND(PF)	$d2$	5.59	8.57	5.63	4.93	5.50	4.78
RAND(MH)[r]	$d3r$	8.00	10.45	8.18	7.78	8.79	7.71
RM(D)	$d5$	7.68	10.40	8.41	7.91	8.61	7.25
NHPR(SI)	$d4$	14.46	16.88	14.38	13.77	14.20	13.38
$n = 200$	$n1$	8.92	12.00	9.12	8.54	9.25	8.23
$n = 400$	$n2$	8.92	11.54	9.15	8.59	9.26	8.27
$n = 800$	$n3$	8.96	11.18	9.18	8.65	9.31	8.34
MCAR	$m1$	6.50	7.41	6.71	6.27	6.97	5.94
NRX	$m2$	9.11	12.90	9.30	8.47	9.21	8.18
NRXZ	$m3$	11.19	14.41	11.43	11.04	11.65	10.72
$p = 0.05$	$p1$	7.54	9.56	7.73	7.02	7.79	6.76
$p = 0.12$	$p2$	9.00	11.68	8.61	8.61	9.27	8.29
$p = 0.20$	$p3$	10.26	13.48	10.56	10.16	10.76	9.80

[r] *Recoded data.*

perform as well as the other two, and most of the time CIMA and MOK perform better. The hot deck technique, however, is in practically all cells (98 of 108) the least accurate. The construction of adjustment cells does not seem to improve the main effects of the techniques.

The effects of the four independent factors are as expected. First, incorrect imputation in a short scale is more severe than in a long scale, as is the case with a scale consisting of dichotomous items compared with polytomous items. Second, the effect of sample size is negligible. Only for the hot deck technique there is a slight improvement. Third, the more complicated the missingness mechanism, the harder it is to impute correctly, as is the same for higher proportions of missing responses.

Interactions Investigating interactions (not reported here) reveals the same general picture: when the situation becomes 'uglier', the performance worsens. 'Uglier' is here defined as a scale that is more difficult to impute (higher weights),

more complicated missing data mechanisms, and/or more missing values. OMD generally proves to be the best technique. However, the differences between OMD and OEV, CIMA, and in some cases MOK are not large and not in all cases in favor of OMD. The HDA procedure performs worst, but the performance gets better with increasing levels of n and when the mechanism is a function of the covariates (NRXZ). For the RM(D) data the performance of the IRT imputation techniques (including CIMA) also improves when the mechanism is nonignorable. More specifically, for m = NRX, the methods perform better than when the data are MCAR.

4.4.2 Assessment of scale quality

The main results of the investigation of Cronbach's *alpha* and Loevinger's H for and after imputation, show the same picture as was found in the first simulation study. Imputation techniques based on IRT models result in an overestimation of the reliability and scalability of the items. The overestimation is largest for OEV. Even if the data are MCAR, the coefficients are severely overestimated with OEV. For example, in case of m = MCAR, p = 0.12, and d = RM(D), H equals 0.34, which classifies the items as weakly scalable (see Mokken, 1971), where the original value equals 0.25, indicating nonscalable items.

The overestimation is generally smallest for OMD and OSD, followed by MOK and CIMA. For the dichotomous data [NHPR(SI) and RM(D)], MOK performs better than CIMA, for the polytomous items CIMA is better. The hot deck procedure, on the other hand, underestimates the quality of the scale. The absolute deviations are generally smallest when the data are MCAR, but generally largest for mechanism NRX. For NRXZ they are as large as those found with CIMA. Again, the small values of the deviations in *alpha* show that long scales are more robust against imputation according to this criterion.

5 Discussion

As expected, from the results of the simulations studies it follows that the proportion of missing data, the missingness mechanism, and characteristics of the scale under investigation, all largely influence the success of imputing missing item responses. Sample size, on the other hand, does not seem to affect the performance of the imputation techniques very much. Exceptions are the hot deck procedure, for which the performance slightly improves when the sample size increases, and

imputation within adjustment cells. The effect of sample size is still small in these cases.

Although the two simulation studies cannot strictly be compared, due to minor differences in the designs, some general remarks can be made. When the sum score r_v of the respondents is investigated, the OPMISS procedures generally prove to be the best performing techniques. If only the mean sum scores are considered, imputing the expected value from the estimated response distributions (OEV) is best, if the variation between the sum scores is also taken into account, OMD (average of five times multiply imputing random draws from the estimated response distributions) performs best. A drawback of the multiple imputation procedure is that every data set has to be imputed and analyzed five times, and the results have to be combined to obtain a final result. The results of imputing only one draw from the response distribution, however, are not as good as the other two methods. Imputing a corrected item mean is in most cases actually even better. Also the simpler methods ICS and PMS work reasonable well in some specific situations.

Investigation of the ability of the imputation techniques to recover the relationships that exist between the items, by looking at differences between reliability of the scale (Cronbach's *alpha*) and scalability of the items (Loevinger's H) before and after creating and imputation of missing values, shows a disadvantage of using IRT models for imputation: the IRT imputation methods cause an overestimation of the quality of the scale. This overestimation is due to using IRT models for both imputation and analysis of the completed data, which causes an increased model fit and therefore a too optimistic presentation of the internal structure of the test. The increased model fit results in higher values of *alpha* and H, and can be very severe. The same is true for imputing a corrected item mean, which can be considered a very simple IRT imputation model. The overestimation is largest for OEV, and smallest for OMD.

The success of the imputations not only depends on the factors in the design, but is also dependent on the quality of the imputation model. If a good fit can be accomplished, the parameter estimates, and therefore the estimates of the response distributions will be better. This will result in better estimates of r_v. There are, however, two drawbacks. First, improving the model fit will cause the overestimation of the internal consistency of the scale to become even larger. Second, more time and effort are needed to obtain estimates of the missing values when more advanced imputation models are used. In this respect, the computational simplicity of the methods CIM and MOK (or even ICS and PMS)

may prevail the slightly better results obtained with the OPMISS procedures, which are more difficult to use.

Using IRT models for the imputation of scale (i.e., IRT) data results in good estimates of the missing values, especially when the missing data mechanism is in some way dependent on the item responses. The case of the artificially created Rasch data illustrates this result. Here, the performance of the IRT procedures improves when the mechanism changes from missing completely at random to a nonignorable mechanism in which the missingness is a logistic function of the item responses. The missing data mechanism in this case resembles the data generating model, and is implicitly included in the analysis when IRT imputation models are used to handle the missing data. This results in better predictions of the missing responses. When covariate information is used for the generation of the missing values, the performance of some procedures can be slightly improved by post-stratification on the covariates. The construction of adjustment cells, however, only results in a better performance when large proportions of item responses are missing, and only the hot deck procedures showed some improvement.

How to deal with missing values is a problem that is hard to tackle and for which no generally best solution exists. Imputation is one way in which item non-response can be treated, although it can never completely solve the missing data problem. Still, trying to find good imputation methods is worthwhile, because complete data sets are rare in empirical research and missing data will not cease to exist.

References

Bello, A.L. (1993). Choosing among imputation techniques for incomplete multivariate data: A simulation study. *Communications in Statistics A, 22*, 853-877.

Eggen, T.J.H.M. & Verhelst, N.D. (1992). *Item calibration in incomplete testing designs*. (Measurement and Research Department reports, 92-3.) Arnhem: CITO.

Greenlees, J.S., Reece, W.S., & Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association, 77*, 251-261.

Hendriks, A.A.J. (1997). *The construction of the Five-Factor Personality Inventory (FFPI)*. (Ph.D. thesis.) Groningen: University of Groningen.

Huisman, M. (1999). *Item nonresponse: Occurrence, causes, and imputation of missing answers to test items*. Leiden: DSWO Press.

Hunt, S.M., McKenna, S.P. & McEwen, J. (1993). Nottingham Health Profile (NHP). In C. Koenig-Zahn, J.W. Furer, & B. Tax (Eds.), *Het meten van de gezondheidstoestand; 1-Algemene gezondheid* (pp. 100-114). Assen: Van Gorcum.

Kromrey, J.D. & Hines, C.V. (1994). Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments. *Educational and Psychological Measurement, 54*, 573-593.

Laros, J.A. & Tellegen, P.J. (1991). *Construction and validation of the SON-R* $5\frac{1}{2}$*-17, the Snijders-Oomen non-verbal intelligence test.* Groningen: Wolters-Noordhoff.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics, 6*, 287-296.

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

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading: Addison-Wesley.

Mislevy, R.J. & Wu, P-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing.* (Research report RR-96-30-ONR.) Princeton, NJ: Educational Testing Service.

Mokken, R.J. (1971). *A theory and procedure of scale analysis with applications in political research.* New York/Berlin: de Gruyter, Mouton.

Molenaar, I.W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 37*, 97-117.

Molenaar, I.W. (1997). Nonparametric models for polytomous responses. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369-380). New York: Springer-Verlag.

Nap, R.E. (1994a). *OPMISS: Handling missing data in OPLM.* (Heymans Bulletin HB-94-1173-IN.) Groningen: Department of Statistics & Measurement Theory, University of Groningen.

Nap, R.E. (1994b). *Missing Data: different forms of imputation methods and their application to empirical data sets.* (Research report VSM-94-01-SW.) Groningen: Department of Statistics & Measurement Theory, University of Groningen.

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

Rubin, D.B., Stern H.S., & Vehovar, V. (1995). Handling "don't know" survey responses: The case of the Slovenian plebicite. *Journal of the American Statistical Association, 90*, 822-828.

Sande, I.G. (1982). Imputation in surveys: Coping with reality. *The American Statistician, 36*, 145-152.

Sande, I.G. (1983). Hot-deck imputation procedures. In W.G. Madow, I. Olkin & D.B. Rubin (Eds.), *Incomplete data in sample surveys, Vol. III: Proceedings of the symposium* (pp. 339-349). New York: Academic Press.

Van der Zee, K.I. & Sanderman, R. (1993). *Het meten van de algemene gezondheidstoestand met de RAND-36: Een handleiding.* [Measuring the general state of health with the RAND-36: A manual.] (NCG reeks meetinstrumenten; 3.) Groningen: Northern Center for Healthcare Research (NCG).

Verhelst, N.D. & Glas, C.A.W. (1995). The one parameter logistic model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 215-237). New York: Springer-Verlag.

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). *One-Parameter Logistic Model OPLM*. Arnhem: CITO, National Institute for Educational Measurement, The Netherlands.

Ontvangen: 15-02-1999
Geaccepteerd: 10-05-1999