The effect of missing data imputation on Mokken scale analysis

van der Ark, L.A.; Sijtsma, K.

Published in:
New developments in categorical data analysis for the social and behavioral sciences.

Publication date:
2005

Chapter 8

The Effect of Missing Data Imputation on Mokken Scale Analysis

L. Andries van der Ark[1] and Klaas Sijtsma
Tilburg University

8.1 Introduction

Tests and questionnaires can be constructed mainly in two ways. The first is exploratory. This means that the final test is selected from the initial set of items so as to optimize psychometric criteria. For example, the test constructor may want to select a subset of items so as to satisfy a lower bound for the reliability of person ordering. The second way of test construction is confirmatory. This means that the set of items is considered to be fixed and the psychometric properties of this set are determined under a particular model without changing the composition of the item set. For example, after fifteen years of use the test constructor may decide that the norms for interpretation of test results need to be updated. The stand-alone software

[1]The first author's research has been supported by the Netherlands Research Council (NWO), Grant No. 400.20.011. Thanks are due to Liesbeth van den Munckhof for her assistance with the MSP analyses and Joost van Ginkel for correcting an error in the initial computation of the statistic MIN.

package MSP (Molenaar & Sijtsma, 2000) allows both possibilities. A well known problem in data analysis for test and questionnaire construction is that some of the N respondents did not supply an answer to some of the J items, so that the data matrix $\mathbf{X}$ is incomplete. MSP only offers listwise deletion to handle the missing data problem. This may result in the loss of many cases, biased estimates of parameters of interest, and reduced accuracy of estimates. The topic of this chapter is the comparison of imputation methods with respect to the outcomes of exploratory and confirmatory test construction as implemented in MSP.

8.1.1 Missing Data Mechanisms

Missing item scores may be due to many reasons. Often these reasons are unknown to the researcher. For example, the respondent may have missed a particular item (e.g., due to inattention or time pressure), missed a whole page of items, saved the item for later and then forgot about it, did not know the answer and then left it open, became bored while taking the test or questionnaire and skipped a few items, felt the item was embarrassing (e.g., questions about one's sexual habits), threatening (questions about the relationship with one's children), or intrusive to privacy (questions about one's income and consumer habits), or felt otherwise uneasy and reluctant to answer.

Rubin (1976; also, see Little & Rubin, 1987; Schafer, 1997) formalized mechanisms of missing data into three classes. Let i denote the respondent index and j the item index, and let x_{ij} be the integer score of respondent i on item j. Let $\mathbf{M}$ be an $N \times J$ indicator matrix of with elements $m_{ij} = 1$ if score x_{ij} is missing, and $m_{ij} = 0$ if score x_{ij} is observed. The observed part of $\mathbf{X}$ is denoted $\mathbf{X}_{obs}$ and the missing part is denoted $\mathbf{X}_{mis}$. Thus, $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$. Let β be a set of parameters governing the data, $\mathbf{X}_{obs}$ and $\mathbf{X}_{mis}$, and ξ a set of parameters governing the missingness, $\mathbf{M}$. We may model the distribution of the missing data as $P(\mathbf{M}|\mathbf{X}_{mis}, \mathbf{X}_{obs}, \beta, \xi)$.

The missing data are called *missing at random* (MAR) when the distribution of the missing data does not depend on the missing item scores; that is

$$P(\mathbf{M}|\mathbf{X}_{mis}, \mathbf{X}_{obs}, \beta, \xi) = P(\mathbf{M}|\mathbf{X}_{obs}, \xi)P(\mathbf{X}_{obs}|\beta).$$

An example of MAR is that missing item scores depend on other observed items or covariates. Such a covariate may be gender. For example, for men it may be more difficult to admit to the item 'I cry at weddings' than for women (item taken from questionnaire by Vingerhoets & Cornelius, 2001). Therefore, a larger proportion of the male respondents may decide not to respond to this item.

A special case of MAR is *missing completely at random* (MCAR). Data are MCAR when the missing data values are a simple random sample of all data values; that is,

$$P(\mathbf{M}|\mathbf{X}_{mis}, \mathbf{X}_{obs}, \beta, \xi) = P(\mathbf{M}|\xi).$$

For MCAR the parameters in ξ only affect the proportion of missing values, but not the pattern of missingness.

Missing data are called *nonignorable* when their distribution $P(\mathbf{M}|\mathbf{X}_{mis}, \mathbf{X}_{obs}, \beta, \xi)$ depends on $\mathbf{X}_{obs}$, $\mathbf{X}_{mis}$, and ξ, and indirectly on β since these parameters govern $\mathbf{X}_{obs}$ and $\mathbf{X}_{mis}$. One example of a nonignorable missingness mechanism is that the distribution of the missing data depends on values of variables that were not part of the investigation. For example, in a personality inventory missingness may depend on general intelligence or reading ability. Another example of a nonignorable missingness mechanism is that the distribution of the missing data depends on the missing item scores; for example, respondents who cry at weddings have a higher probability of not answering the item 'I cry at weddings' than respondents who never cry at weddings. Consequently, any missing data method based on available item scores would underestimate the missing value.

8.1.2 Test Construction

Exploratory and confirmatory test construction

Our frame of reference in this study is nonparametric item response theory (NIRT; Boomsma, Van Duijn, & Snijders, 2001; Mokken, 1971; Sijtsma & Molenaar, 2002; Van der Linden & Hambleton, 1997). Following NIRT, we define a latent trait θ that stands for a psychological property or a collection of psychological properties measured by the J items. For example, the item "I cry at weddings" may be indicative of the latent trait "tendency to cry". Parameter θ thus governs the data and replaces parameter vector β. Let X_j be the random variable for the score on item j. Item scores may be dichotomous or polytomous. For example, the item "I cry at weddings" may have only two answer categories, "applies" and "does not apply", which may be dichotomously scored $x_j = 1$ and $x_j = 0$ with respect to latent trait "tendency to cry", respectively. Another possibility is that the respondent indicates on an ordered rating scale the degree to which the item applies to him/her, and the corresponding polytomous scoring then may be $x_j = 0, \ldots, g$. Latent trait θ is estimated by means of $X_+ = \sum_j X_j$ (Hemker, Van der Ark, & Sijtsma, 2001; Junker, 1991; Stout, 1990). Note that X_+ may either estimate a unidimensional θ or a multidimensional θ.

The construction of a test or questionnaire mainly follows two possibilities. The first possibility is that one starts from scratch, defining the construct of interest and a useful operationalization, and then defines a collection of experimental items. Then a clustering method from MSP may be used to determine the structure of the data in terms of the underlying latent traits. A cluster is a set of items that measure the same latent trait. This is an exploratory approach because the dimensionality structure was not hypothesized prior to the application of the clustering method but found by the program. The second possibility is that one starts with an existing instrument and wants to know whether it can be used in another population or at a later point in time. This entails drawing a new sample of respondents to which the existing item set is administered, or administering the item set to the same respondents once more. Then MSP may be used to analyze the item set as one cluster and determine its psychometric properties. Because the item set is considered to be fixed, we consider this kind of item analysis to be confirmatory in the sense that for this set it is determined whether or not it is a useful instrument in a new context.

Test construction according to MSP

Scalability coefficients. Both for exploratory and confirmatory test construction, MSP uses the scalability coefficient H (Mokken, 1971, pp. 148-153; 1997; Sijtsma & Molenaar, 2002, pp. 49-64) as a scaling criterion. For two items j and k, $Cov(X_j, X_k)$ defines their covariance and $Cov(X_j, X_k)_{\max}$ defines their maximum covariance given the marginal distributions of their bivariate frequency table. The scalability coefficient for these two items is defined as

$$H_{jk} = \frac{Cov(X_j, X_k)}{Cov(X_j, X_k)_{\max}}.$$

Coefficient H_{jk} is the basis for the scalability coefficient of one item with respect to the other $J-1$ items; this coefficient is denoted H_j and defined as

$$H_j = \frac{\sum_{k \neq j}^{J} Cov(X_j, X_k)}{\sum_{k \neq j}^{J} Cov(X_j, X_k)_{\max}}.$$

Finally, scalability coefficient H for all J items is defined as

$$H = \frac{\sum_{j=1}^{J-1} \sum_{k=j+1}^{J} Cov(X_j, X_k)}{\sum_{j=1}^{J-1} \sum_{k=j+1}^{J} Cov(X_j, X_k)_{\max}}.$$

Monotone homogeneity model. The use of scalability coefficients H_{jk}, H_j, and H is related to the monotone homogeneity model (MHM; Mokken, 1971, p. 118). The MHM assumes a unidimensional latent trait θ, local independence of the item scores given θ, and a monotone nondecreasing relationship between $P(X_j \geq x_j|\theta)$ and θ. For scores $x_j = 1, \ldots, g$, the conditional probabilities $P(X_j \geq x_j|\theta)$ are the item step response functions (ISRFs) (for $x_j = 0$ the ISRF equals 1 by definition). For dichotomous items ($g = 1$) the only relevant ISRF is $P(X_j \geq 1|\theta) = P(X_j = 1|\theta)$. This is the item response function (IRF). Together, the assumptions of unidimensionality, local independence, and monotonicity define the MHM. For dichotomous items, the MHM implies the stochastic ordering of latent trait θ by means of observable summary score X_+; that is, for any t, we have that $P(\theta > t|X_+)$ is nondecreasing in X_+ (based on Grayson, 1988; also, see Hemker, Sijtsma, Molenaar, & Junker, 1997). Thus, the MHM implies ordinal person measurement on θ using X_+. The more complicated case for polytomous items is treated by Van der Ark (in press).

Relationship between MHM and coefficient H. The MHM implies that $H_{jk} \geq 0$ (Holland & Rosenbaum, 1986; Mokken, 1971, pp. 149-150). By implication, we have that $H_j \geq 0$ and $H \geq 0$. Based on these implications, Mokken (1971, p. 184; Sijtsma & Molenaar, 2002, pp. 67-68) defined a scale as a set of dichotomously scored items for which, for a suitably chosen positive constant c, and for product-moment correlation ρ,

$$\rho_{jk} > 0, \text{ for all item pairs } (j,k); \tag{8.1}$$

and

$$H_j \geq c > 0, \text{ for all items } j. \tag{8.2}$$

Equation 8.1 implies that $H_{jk} > 0$. Equation 8.1 also implies that $H_j > 0$ and $H > 0$. In addition, by specifying that $H_j \geq c$, Equation 8.2 poses minimum requirements on the slope of the IRF. That is, constant c forces a minimum level of discrimination power on the individual items. This is not implied by the MHM, but because this model allows weakly sloped IRFs and even flat IRFs as a borderline case, the addition of a minimum discrimination requirement is a practical measure for reliable person ordering. Finally, the definition of a scale can be extended readily to polytomous items (Sijtsma & Molenaar, 2002, p. 127).

Automated item selection. For exploratory test construction, MSP selects items according to the definition of a scale (Equations 8.1 and 8.2). The default option for item selection, to be used here, has the following steps (Mokken, 1971, pp. 190-194).

1. From the J available items, MSP selects from the item pairs which have a H_{jk} that is significantly greater than 0, that pair which has the highest H_{jk} that is greater than c. This is the start set for item selection.

2. From the remaining $J-2$ items, that item is added to the start set that (a) has a positive covariance with both selected items (Equation 8.1); (b) has an H_j value with the selected items that is at least c (Equation 8.2); and (c) has the highest common H value with the selected items, given all candidate items for selection.

3. The next items are selected following the logic of Step 2. The item selection for the first scale ends when no more items satisfy the criteria mentioned in Step 2.

4. If items remain unselected after the first scale has been formed, from the unselected items MSP tries to form a second scale, a third scale, and so on, until no more items remain or no more items satisfy the criterion in Step 1.

For confirmatory test construction, the MHM is fitted to the data corresponding to the a priori defined test consisting of J items using methods implemented in MSP (Molenaar & Sijtsma, 2000; Sijtsma & Molenaar, 2002). This includes calculating and evaluating the H_j and H coefficients.

8.2 Methods for Missing Data Imputation

We introduce four methods for the imputation of item scores for missing observations in a data matrix $\mathbf{X}$, plus listwise deletion. Listwise deletion is the only method currently implemented in MSP. It was used as a benchmark for the other methods. For each of the five methods it was investigated how they influence the results of the automated item selection procedure in MSP (exploratory test construction) and how they influence the results of fitting the MHM to an a priori defined scale (confirmatory test construction). The five missing data handling methods are discussed next.

Listwise Deletion. Listwise deletion (LD) deletes from the analysis all cases that have at least one missing item score. Because for data matrices that contained at least ten percent missing item scores it was found that

LD led to the rejection of almost the whole data matrix, in these cases we used the imputation of a random item score as an alternative (called *Random Imputation*; abbreviated RI).

Two-Way Imputation. Because in a unidimensional test or questionnaire all item scores measure the same latent trait, the scores on the available items can be used for imputing scores for missing data. Let PM_i be the mean item score of person i calculated across his/her available item scores; let IM_j be the mean score on item j calculated across the item scores available in the sample of N persons; and let OM be the mean item score calculated across all available item scores in $\mathbf{X}$. Then for missing item score (i, j), we calculate

$$TW_{ij} = PM_i + IM_j - OM;\ TW_{ij} \in \mathbb{R}.$$

The item score to be imputed is obtained by rounding TW_{ij} to the nearest feasible integer. Two-way imputation (TW) was proposed by Bernaards and Sijtsma (2000; see Huisman & Molenaar, 2001, for a related method).

Response Function Imputation. Response function imputation (RF; Sijtsma & Van der Ark, 2003) is based on the idea to impute item scores x_{ij} as random draws from the distribution $P(X_j = x_j|\theta_i)$. The steps in this procedure are the following.

- First, estimate θ_i by means of restscore $R_{i(-j)} = X_{i+} - X_{ij}$ (e.g., Hemker, et al., 1997; Junker, 1993; Sijtsma & Molenaar, 2002, p. 40). This is done as follows. Due to missing data, the number of available item scores on the remaining $J - 1$ items may vary across respondents. This number is denoted J_i ($J_i \leq J - 1$). Restscore $R_{i(-j)}$ is computed as the sum of these available item scores. Because different respondents may have different numbers of available item scores, to have all restscores on the same scale each restscore is multiplied by $(J - 1)/J_i$.

- Second, estimate $P(X_j = x_j|\theta_i)$ by means of $P[X_j = x_j|R_{i(-j)}]$, for $x_j = 0, \ldots, m$. The latter probability is computed in the subgroup having an observed score on X_j. Each respondent's X_j is weighted by the accuracy with which his/her restscore, $R_{i(-j)}$, estimates its expectation, $E_i[R_{i(-j)}]$. Because for each respondent one restscore is available, the determination of its accuracy is based on its constituent J_i item scores. Let the mean item score of respondent i be denoted $\overline{X}_i = \frac{R_{i(-j)}}{J_i}$. Let σ_i^2 denote the variance of the item scores of respondent i, estimated by $S_i^2 = \frac{\sum_j (X_{ij} - \overline{X}_i)^2}{J_i}$. The inaccuracy of $\overline{X}_i$ is given by $SE(\overline{X}_i) = \sqrt{S_i^2/J_i}$. The weight for respondent i in computing $P[X_j = x_j|R_{i(-j)}]$ is $1/SE(\overline{X}_i)$.

- Third, for a missing score in cell (i, j) we impute a random draw from $P[X_j|R_{i(-j)}]$. In the subgroup of people having a missing score on item j, restscores may exist that did not exist in the group with X_j observed that was used for estimating $P[X_j|R_{i(-j)}]$. For example, among the latter group $R_{i(-j)} = 2$ may not have been observed; thus, $P[X_j|R_{i(-j)} = 2]$ was not estimated. In that case, item score probabilities are obtained by linear interpolation between the two nearest restscores from the group with X_j observed. If restscore groups are too small for an accurate estimate of $P[X_j|R_{i(-j)}]$, adjacent restscore groups may be joined. See Sijtsma and Van der Ark (2003) for more details.

Multiple Response Function Imputation. Multiple response function imputation (MRF) entails five times the application of the RF procedure. This involves five random draws from $P[X_j|R_{i(-j)}]$, which yields five different completed data matrices. Each completed data matrix is analyzed separately, and the results are combined later using Rubin's rules (see, e.g., Schafer, 1997, pp. 109-110) or a variation to be discussed later.

Multiple multivariate normal imputation. An imputation method for categorical data proposed by Schafer (1997, pp. 257-275) and implemented in publicly available software (program CAT; Schafer, 1998a) was considered for item score imputation. This method requires a frequency table based on J items with $m+1$ answer categories, which thus has $(m+1)^J$ entries. In our applications, this number was too large for maximum likelihood estimation of the imputation model. Thus, CAT could not be used. Instead we assumed a multivariate normal imputation model as suggested by Schafer (1997, p. 148; program NORM, Schafer, 1998b). The method is called multiple multivariate normal imputation (MMNI). Method MMNI assumes that the item scores have a J-variate normal distribution. In an initial step the model parameters, the mean vector and the covariance matrix, are estimated using an EM algorithm. Then an iterative procedure called *data augmentation* is used to obtain the distribution of the missing item scores given the observed item scores and the model parameters. The missing values are imputed by random draws from this conditional distribution. Since these random draws are real-valued and our data integer-valued, the random draws were rounded to the nearest feasible integer. For more detailed information on data augmentation we refer to Tanner and Wong (1987) and for the implementation of EM and data augmentation in NORM to Schafer (1997, chap. 5 and chap. 6).

8.3 Method

We investigated the influence of each of the five imputation methods on the results of confirmatory and exploratory item analysis using the program MSP. Three real data sets (first design factor) were used. These data sets are referred to as *original* data sets.

- **Verbal analogies data** (Meijer, Sijtsma, & Smid, 1990). For this data set, $N = 990$ and $J = 32$, with $g + 1 = 2$. This test measures verbal intelligence in adults. Meijer et al. (1990) found that 31 items together formed one scale (each $H_j > 0$). This was the basis for the confirmatory analysis. All 32 items were used in the exploratory analysis.

- **Coping data** (Cavalini, 1992). For this data set, $N = 828$ and $J = 17$, with $g + 1 = 4$. This questionnaire measures coping styles in response to industrial malodors. Cavalini (1992, pp. 53-54) found four item subsets (17 items in total) measuring different coping styles. Each of these subsets was used separately in the confirmatory analysis. The set of 17 items was the input for the exploratory analysis.

- **Crying data** (Vingerhoets & Cornelius, 2001). Here, $N = 3965$ and $J = 54$, with $g + 1 = 7$. This questionnaire measures determinants of adult crying behavior. Scheirs and Sijtsma (2001) found three subsets of items (54 items in total), representing three psychological states. Each subset was the basis of the confirmatory analysis. All 54 items together were subjected to the exploratory analysis.

Each data set was complete. In each original data set item scores were deleted using procedures that resulted in either MCAR, MAR, or nonignorable missingness (second design factor). The percentage of missing item scores was either 5%, 10%, or 20% (third design factor). The data sets containing missing data are referred to as *incomplete* data sets. Missingness was simulated as follows

- **MCAR**. The probability of a missing score was the same for each entry in the data set.

- **MAR**. Let $L = \text{trunc}(J/2)$ be a cut-off value that splits the item set into a first half (items $1, \ldots, L$) and a second half (items $L+1, \ldots, J$). When the missing item scores were MAR, the probability of a missing item score in the second half was twice the probability of a missing item score in the first half.

- **Nonignorable missingness.** When missingness was nonignorable, the missing item scores were MAR in combination with the following mechanism: Let $G = \text{trunc}(g/2)$ be a cut-off value that splits the item scores into low item scores $(0, \ldots, G)$ and high item scores $(G + 1, \ldots, g)$. The probability of a missing value for high item scores was twice the probability of a missing value for low item scores.

The incomplete data sets were imputed using *listwise deletion* (5% missing item scores) or *random imputation* (10% and 20% missing item scores), *two-way imputation*, *response function imputation*, *multiple response function imputation*, and *multiple multivariate normal imputation* (fourth design factor). These data sets are referred to as *completed* data sets. Both the original and the completed data sets were subjected to exploratory and confirmatory data analysis (fifth design factor).

Exploratory analysis. For the single imputation methods (RI, TW, and RF), for each incomplete data set, the MCAR, MAR, and nonignorable missingness conditions were used to construct three different completed data sets. For each completed data set, MSP found a cluster solution, which was compared with the original data cluster solution. Assume that an item set consists of five items, indexed $j = 1, \ldots, 5$, then the original-data clustering might be $(1, 2, 2, 0, 1)$: The 1 scores indicate that items 1 and 5 were in the same cluster, the 2 scores that items 2 and 3 were in another cluster, and the 0 score that item 4 remained unselected. Now, assume that the completed-data clustering is $(1, 1, 1, 0, 0)$; then, ignoring the cluster numbering (which is nominal) the smallest number of items to be moved to reobtain the original-data solution is sought. Here, items 1 and 5 need to be moved to a separate cluster. Denote the minimum number of items to be moved by MIN (with realization min), then for this example $MIN = 2$.

For the multiple imputation methods (MRF and MMNI), for each incomplete data set five completed data sets were generated. The five completed-data cluster solutions were combined to one by taking the mode of the cluster indices for each item. For example, let the five cluster solutions found be $(1, 2, 2, 0, 1)$, $(2, 2, 1, 0, 1)$, $(1, 2, 1, 1, 2)$, $(1, 2, 2, 0, 1)$, and $(0, 2, 2, 0, 0)$; then, the modal solution is $(1, 2, 2, 0, 1)$ and the MIN value with respect to the original-data clustering, which was $(1, 2, 2, 0, 1)$ (previous example), is determined to be 0.

Confirmatory analysis. The H values of the completed data were compared with the H values of the corresponding original data. For multiple imputation the mean H of the five completed data matrices was taken.

The design was completely crossed with 3 (original data matrices) $\times$ 3 (missingness mechanisms) $\times$ 3 (percentages of missing item scores) $\times$ 5

Table 8.1: Number of Verbal Analogies Items Incorrectly Clustered in Exploratory Analysis, for Five Imputation Methods, Three Missingness Mechanisms, and Three Percentages (5, 10, and 20) of Imputed Item Scores [$J = 32$; $\max(MIN) = 18$].

Method	Missingness Mechanism								
	MCAR			MAR			Nonignorable		
	5	10	20	5	10	20	5	10	20
LD/ RI	13	18	18	10	18	16	8	18	18
TW	8	14	16	5	15	16	4	9	16
RF	4	3	8	5	3	7	3	8	4
MRF	2	2	7	5	6	9	3	3	4
MMNI	10	17	17	12	11	16	6	12	16

(imputation methods) $\times$ 2 (exploratory vs. confirmatory analysis) = 270 cells. The study was programmed in S-Plus 6 for Windows (2001); the exploratory and confirmatory analyses were done using MSP (Molenaar & Sijtsma, 2000).

8.4 Results

8.4.1 Exploratory Analyses

Table 8.1 (Verbal Analogies data), Table 8.2 (Coping data), and Table 8.3 (Crying data) give the value of MIN for the complete design. An unscalable set of items is one in which each item forms a unique cluster; for this setup MIN was determined, and the result was called $\max(MIN)$. The value of $\max(MIN)$ was used as a benchmark.

Verbal analogies data. Methods LD and RI always led to almost one half to all items incorrectly clustered ($8 \leq min \leq 18$). Method TW led to a misclassification of almost all items for 10% and 20% imputed item scores. Methods RF and MRF performed best ($2 \leq min \leq 8$). Method MMNI led to high MIN-values ($6 \leq min \leq 17$). This result was not expected and may be related to convergence to a local optimum. This is further elaborated in the Discussion.

Coping data. For 5% imputed item scores, all methods performed well. For 10% and 20% imputed item scores, method RI led to large values of MIN. Methods TW, RF, and MRF led to the misclassification of approximately one-fifth of the items for 10% imputed item scores, and to

Table 8.2: Number of Coping Data Items Incorrectly Clustered in Exploratory Analysis, for Five Imputation Methods, Three Missingness Mechanisms, and Three Percentages (5, 10, and 20) of Imputed Item Scores [$J = 17$; $\max(MIN) = 12$].

Method	Missingness Mechanism								
	MCAR			MAR			Nonignorable		
	5	10	20	5	10	20	5	10	20
LD/ RI	1	6	10	1	6	10	1	7	10
TW	0	3	6	1	3	5	0	1	4
RF	0	2	6	0	2	5	0	4	4
MRF	0	1	6	0	2	4	0	3	5
MMNI	0	0	0	0	0	1	0	0	0

Table 8.3: Number of Crying Data Items Incorrectly Clustered in Exploratory Analysis, for Five Imputation Methods, Three Missingness Mechanisms, and Three Percentages (5, 10, and 20) of Imputed Item Scores [$J = 54$; $\max(MIN) = 45$].

Method	Missingness Mechanism								
	MCAR			MAR			Nonignorable		
	5	10	20	5	10	20	5	10	20
LD/ RI	10	16	29	9	17	34	11	21	38
TW	5	3	10	2	7	5	3	3	12
RF	5	4	7	2	4	6	3	6	10
MRF	3	4	6	3	5	7	1	6	10
MMNI	21	16	44	25	36	44	16	32	44

Table 8.4: Bias in H (in hundredths; i.e., -2 stands for $-.02$) for One Cluster of Verbal Analogies Items, for Five Imputation Methods, Three Missingness Mechanisms, and Three Percentages (5, 10, and 20) of Imputed Item Scores ($J = 31, H = .25$).

Method	Missingness Mechanism								
	MCAR			MAR			Nonignorable		
	5	10	20	5	10	20	5	10	20
LD/ RI	1	9	5	−1	−16	−20	0	−15	−19
TW	−3	−5	−9	−2	−9	−10	−2	−4	−7
RF	0	0	−1	0	0	−1	0	0	−1
MRF	0	0	−1	0	0	−1	0	0	−1
MMNI	−2	−5	−10	−2	−7	−10	−2	−5	−9

the misclassification of approximately one-third of the items for 20% imputed item scores. Method MMNI led to a correct clustering except for 20% item scores that were MAR. Only small differences were found among the missing data mechanisms MCAR, MAR and nonignorable.

Crying data. Method LD/RI led to a misclassification of approximately one-fifth (5% missing item scores, $min = 9$) to two-thirds (20% missing item scores, $min = 38$) of the items. Method MMNI resulted in even higher MIN-values ($16 \leq min \leq 44$). Similar to the results for the Verbal Analogies data (Table 8.1), this is probably due to a bad model-fit. Methods TW, RF, and MRF performed best and yielded misclassifications of approximately one-tenth (5% and 10% imputed item scores) to one-fifth (20% imputed item scores) of the items. Only small differences were found among the missing data mechanisms MCAR, MAR and nonignorable.

8.4.2 Confirmatory Analysis

Table 8.4 (Verbal Analogies data), Table 8.5 (Coping data), and Table 8.6 (Crying data) give the bias in H for the entire design of a single predefined cluster of a data set. The bias is defined as H of the completed data minus H of the original data. For notational convenience the fractional divisions and leading zeros are omitted. Thus, a bias notation of -2 stands for -0.02.

Verbal analogies data. For 5% imputed item scores all imputation methods led to a small bias (Table 8.4). For 10% and 20% imputed item scores, methods TW and MMNI led to a negative bias between $-.10$ and

Table 8.5: Bias in H (in hundredths; i.e., -2 stands for $-.02$) for Four Clusters of Coping Data Items, for Five Imputation Methods, Three Missingness Mechanisms, and Three Percentages (5, 10, and 20) of Imputed Item Scores (Cluster I: $J = 7$, $H = .31$; Cluster II: $J = 4$, $H = .50$; Cluster III: $J = 3$, $H = .56$; Cluster IV: $J = 3$, $H = .35$).

Method	Missingness Mechanism								
	MCAR			MAR			Nonignorable		
	5	10	20	5	10	20	5	10	20
	Cluster I								
LD/ RI	1	−7	−17	−1	−10	−16	−2	−9	−17
TW	0	0	2	1	0	0	0	1	2
RF	1	0	−2	0	−1	−3	−1	0	−3
MRF	0	0	−2	0	−1	−3	0	−1	−2
MMNI	0	1	−1	0	0	0	0	0	−1
	Cluster II								
LD/ RI	−1	−18	−27	−2	−20	−29	1	−16	−31
TW	−1	−6	−2	−3	−7	−7	−2	−6	−7
RF	0	−3	−6	−2	−2	−11	−2	−3	−7
MRF	−1	−3	−7	−2	−4	−10	−1	−4	−9
MMNI	1	−2	−1	0	−1	−1	0	−2	−3
	Cluster III								
LD/ RI	−2	−13	−21	1	−9	−13	−2	−8	−16
TW	1	3	3	1	1	2	1	1	4
RF	−2	−4	−14	0	−1	−3	−1	−1	−5
MRF	−2	−3	−13	0	−1	−3	−1	−1	−3
MMNI	−2	−2	−1	0	0	−2	−1	0	−2
	Cluster IV								
LD/ RI	1	−9	−14	2	−9	−14	0	−9	−16
TW	3	4	7	4	6	13	3	9	16
RF	0	−2	−1	2	−3	−5	−3	−2	−3
MRF	0	−2	−3	0	−3	−4	1	−2	−6
MMNI	0	−1	0	1	0	1	1	1	3

Table 8.6: Bias in H (in hundredths; i.e., -2 stands for $-.02$) for Three Clusters of Crying Data Items, for Five Imputation Methods, Three Missingness Mechanisms, and Three Percentages (5, 10, and 20) of Imputed Item Scores (Cluster I: $J = 22$, $H = .43$; Cluster II: $J = 14$, $H = .41$; Cluster III: $J = 18$, $H = .30$).

Method	Missingness Mechanism								
	MCAR			MAR			Nonignorable		
	5	10	20	5	10	20	5	10	20
	Cluster I								
LD/ RI	1	−12	−20	0	−13	−22	−2	−12	−22
TW	−1	−2	−4	−1	−2	−4	−2	−4	−6
RF	−1	−1	−3	0	−1	−3	−1	−2	−5
MRF	−1	−1	−3	−1	−1	−3	−1	−2	−5
MMNI	0	0	0	0	−1	0	0	−1	−1
	Cluster II								
LD/ RI	−1	−9	−16	2	−9	−16	0	−10	−17
TW	−2	−4	−7	−2	−4	−7	−3	−6	−9
RF	0	−1	−2	0	−1	−2	−1	−1	−4
MRF	0	0	−2	0	−1	−2	0	−1	−4
MMNI	0	0	0	0	0	0	0	0	−1
	Cluster III								
LD/ RI	0	−10	−17	0	−10	−16	−1	−10	−16
TW	0	0	−1	0	0	−1	0	−1	−1
RF	−1	−1	−3	−1	−1	−3	−1	−2	−4
MRF	−1	−1	−4	−1	−1	−3	−1	−1	−4
MMNI	0	0	−1	0	0	−1	0	−1	−1

−.04. Methods RF and MRF performed best yielding unbiased or almost unbiased results in all cases.

Coping data. The results for the four clusters of the Coping data are presented in Table 8.5. For Cluster I, all methods except LD/RI yielded a small bias in H in all conditions; method MMNI gave the best results.

For Cluster II, method LD/RI had a small bias for 5% missing item scores and a large negative bias for 10% and 20% missing item scores. Methods TW, RF, and MRF had a small negative bias within the range $[-.07, .00]$, for 5% and 10% imputed item scores, and a larger negative bias within the range $[-.11, -.02]$, for 20% imputed item scores. Method MMNI was the most successful method, the largest bias in H being $-.03$.

Similar to Cluster II, for Cluster III method LD/RI showed a large negative bias for 10% and 20% imputed item scores. Method TW led to a small positive bias in H, and method MMNI led to a small negative bias. Methods RF and MRF showed a large negative bias ($-.14$) in H when applied to data with 20% item scores that were MCAR. This unexpected result may be related to the small number of items in Cluster III. This is further elaborated in the Discussion.

Similar to Cluster II and Cluster III, for Cluster IV method LD/RI showed a large negative bias for 10% and 20% imputed item scores. Methods RF, MRF, and MMNI gave the best bias results, which were between $-.06$ and .03. Method TW showed large positive bias (.07, .13, and .16) when applied to data with 20% imputed item scores. This unexpected result may also be related to the small number of items in Cluster IV.

For all item clusters it was found that there were only small differences among MCAR, MAR, and nonignorable missingness. It was also found for all clusters that methods RF and MRF produced approximately the same results.

Crying data. The results for the three clusters of the Crying data are presented in Table 8.6. The results were similar for the three clusters. For 5% imputed item scores all methods led to a small bias in H within the range $[-.03, 02]$. For 10% and 20% imputed item scores, methods TW, RF, MRF, and MMNI produced satisfactory results although, when applied to Cluster II, method TW produced a bias that was a little higher (within the range $[-.04, -.09]$). Method MMNI performed best. There were only small differences among MCAR, MAR, and nonignorable missingness.

8.5 Discussion

This chapter showed that using method LD in Mokken scale analysis can result in cluster solutions that deviate much from the cluster solutions that

would have been obtained had the data been complete. For 10% and 20% missingness, the number of cases left may be so small that Mokken scale analysis becomes impossible. These results are in line with earlier studies on method LD (e.g., Schafer, 1997, p. 23). The alternative benchmark, method RI, led to large values of MIN and large biases in H.

By using total scores on the J items, methods TW, RF, and MRF make use of the property that all items are indicators of the same latent variable. The advantage of method TW is its simplicity, which makes the method easy to use for researchers. The values of MIN and the bias in H resulting from method TW were large for the Verbal Analogies data and smaller for the Coping data and the Crying data.

The results for methods RF and MRF were similar. The main reason for choosing multiple imputation over single imputation is to obtain more stable results and correct standard errors. For Mokken scale analysis the standard errors of H usually do not play an important role, and the bias and the values of H produced by methods RF and MRF were similar. Thus, we could not demonstrate the advantage of method MRF over method RF. Methods RF and MRF are not as simple as method TW and involve some computational decisions, such as the sample size of the restscore-groups and the weight given to each restscore. In general, methods RF and MRF performed a little better than method TW with respect to MIN values and bias.

We found a large bias in H for imputation methods RF and MRF, for a cluster of 3 items (Coping data, Cluster III), 20% missingness, and missingness mechanism MCAR. When $J = 3$, the restscore is based on two items. Given these conditions, theoretically under MCAR it is expected that 32% of the sample has a missing score on one item and 4% of the sample has missing scores on both items. This may have caused inaccurate rest-score estimates which led to the large bias.

Method MMNI yielded the lowest MIN-values and the smallest bias of all methods when the number of items was less than 23 (Crying data, Cluster I). For larger item sets (Verbal Analogies data [$J = 31$], and the Crying data [$J = 54$]), the results for method MMNI were worse than the results for method LD/RI. The reason may be the EM-algorithm in program NORM reached a local optimum for which the fit was much worse than the required fit. The algorithm then kept iterating (without improvement) until the maximum number of iterations was reached, yielding a badly fitting model. Consulting the auxiliary statistics provided by NORM and keeping track of the number of iterations may prevent the researcher from using these wrong estimates. The successor of NORM, which is incorporated in the software package S-plus 6 for Windows (2001), gives an error message in these situations without supplying completed data.

Currently, a more systematic investigation (Van Ginkel, Van der Ark, & Sijtsma, 2004) is conducted to determine the effect of multiple imputation using the methods discussed here on results of Mokken scaling and several other psychometric methods. Using simulated data, several comprehensive designs were analyzed to obtain a more definitive impression about the usefulness of our (multiple) imputation methods.

References

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research, 35*, 321-364.

Boomsma, A., Van Duijn, M. A. J., & Snijders, T. A. B. (Eds.) (2001). *Essays on item response theory.* New York: Springer.

Cavalini, P. M. (1992). *It's an ill wind that brings no good: Studies on odour annoyance and the dispersion of odour concentrations from industries.* Unpublished doctoral dissertation. University of Groningen, The Netherlands.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53*, 383-392.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*, 331-347.

Hemker, B. T., Van der Ark, L. A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika, 66*, 487-506.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*, 1523-1543.

Huisman, J. M. E., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In A. Boomsma, M. A. J. van Duijn & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 221-244). New York: Springer.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimations for polytomous items. *Psychometrika, 56*, 255-278.

Junker, B. W. (1993). Conditional association, essential independence, and monotone unidimensional item response models. *The Annals of Statistics, 21*, 1359-1378.

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

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14*, 283-298.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* The Hague: Mouton/Berlin: De Gruyter.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 352–367). New York: Springer

Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows.* Groningen, The Netherlands: iecProGAMMA.

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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London: Chapman & Hall.

Schafer, J. L. (1998a). CAT. Software for S-PLUS Version 4.0 for Windows. Retrieved from *http://www.stat.psu.edu/~jls/sp40.html.*

Schafer, J. L. (1998b). NORM. Software for S-PLUS Version 4.0 for Windows. Retrieved from *http://www.stat.psu.edu/~jls/sp40.html.*

Scheirs, J. G. M., & Sijtsma, K. (2001). The study of crying: Some methodological considerations and a comparison of methods for analyzing questionnaires. In A. J. J. M. Vingerhoets & R. R. Cornelius (Eds.), *Adult Crying. A Biopsychosocial Approach* (pp. 279–298). Hove, UK: Brunner-Routledge.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

Sijtsma, K., & Van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research, 38*, 505-528.

S-Plus 6 for Windows. [Computer software.] (2001). Seattle, WA: Insightful Corporation.

Stout, W. F. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association, 82*, 528-550.

Van der Ark, L. A. (in press). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika.*

Van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory.* New York: Springer.

Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2004). *Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results.* Manuscript submitted for publication.

Vingerhoets, A. J. J. M., & Cornelius, R. R. (Eds.) (2001). *Adult Crying: A Biopsychosocial Approach.* Hove, UK: Brunner-Routledge.