Tilburg University

Investigation and treatment of missing scores in test and questionnaire data

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

Link to publication in Tilburg University Research Portal

Multivariate Behavioral Research

Investigation and Treatment of Missing Item Scores in Test and Questionnaire Data

Klaas Sijtsma [a]; L. Andries van der Ark [a]
[a] Tilburg University.

PLEASE SCROLL DOWN FOR ARTICLE

Investigation and Treatment of Missing Item Scores in Test and Questionnaire Data

Klaas Sijtsma and L. Andries van der Ark
Tilburg University

This article first discusses a statistical test for investigating whether or not the pattern of missing scores in a respondent-by-item data matrix is random. Since this is an asymptotic test, we investigate whether it is useful in small but realistic sample sizes. Then, we discuss two known simple imputation methods, person mean (PM) and two-way (TW) imputation, and we propose two new imputation methods, response-function (RF) and mean response-function (MRF) imputation. These methods are based on few assumptions about the data structure. An empirical data example with simulated missing item scores shows that the new method RF was superior to the methods PM, TW, and MRF in recovering from incomplete data several statistical properties of the original complete data. Methods TW and RF are useful both when item score missingness is ignorable and nonignorable.

Introduction

A well known problem in data collection using tests and questionnaires is that several item scores may be missing from the *n* respondents by *J* items data matrix, **X**. This may occur for several reasons, often unknown to the researcher. For example, the respondent may have missed a particular item, 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 making 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.

The literature is abundant with methods for handling missing data. For example, Little and Schenker (1995) and Smits, Mellenbergh, and Vorst (2002) discuss and compare a large number of simple and more advanced methods. Several methods are rather involved and, as a result, sometimes perhaps beyond the reach of individual psychological and educational researchers who are not trained statisticians or psychometricians. One

Correspondence concerning this article should be addressed to Klaas Sijtsma, Department of Methodology and Statistics, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands; e-mail: k.sijtsma@uvt.nl

example is the EM method (Dempster, Laird, & Rubin, 1977; Rubin, 1991) that alternately estimates the missing data, then updates the parameter estimates of interest, uses these to re-estimate the missing data, and so on, until the algorithm converges to, for example, maximum likelihood estimates. Another example is multiple imputation (e.g., Little & Rubin, 1987). Here, w complete data matrices are estimated by imputing for a respondent having missing data, for example, scores of sets of other respondents with complete data that are similar to the respondent's available data. Then, statistics based on the w (usually a surprisingly small number; see Rubin, 1991) complete data matrices, are averaged to obtain parameter estimates and standard errors. Data augmentation (Schafer, 1997; Tanner & Wong, 1987) is an iterative Bayesian procedure that resembles the EM method and also incorporates features of multiple imputation (Little & Schenker, 1995).

Our starting point was that many researchers do not have a statistician or a psychometrician in their vicinity who is available to help them implement these superior but complex and involved missing data handling methods. Those researchers may be better off using simpler methods, that are easy to implement and lead to results approaching the quality of EM and multiple imputation. A circumstance favorable for these simpler methods to succeed is that the items in a test measure the same underlying ability or trait and, thus, the observed item scores contain much information about the missing item scores. This helps to obtain reasonable estimates of missing item scores, even with simple methods.

However, first we investigated whether an asymptotic statistical test (Huisman, 1999) for the hypothesis that the pattern of missing item scores in a data matrix $\mathbf{X}$ is random (to be explained later on), is useful in small but realistic sample sizes. This test may be seen as a useful precursor for item score imputation: When its conclusion is that item score missingness is random, the researcher can safely use a sensible item score imputation method to produce a complete data matrix. When item score missingness is not random, imputation methods must be robust so as to produce a data matrix that is not heavily biased. We investigated this robustness issue in a real data example for four imputation methods. Two simple methods were known (e.g., Bernaards & Sijtsma, 2000), and two others were new proposals based on concepts from item response theory (IRT), but without using strong assumptions about the data structure.

Before we continue, it may be noted that a purely statistical approach of the missing data problem may be too simple in some cases. For example, when one item produces most of the missing scores then, depending on the research context, the item may simply be deleted from further research (e.g., it was printed on the back of the page and therefore missed by many), it may be reformulated (e.g., positively worded instead of negatively, which caused

confusion) in future research, or it may be replaced (e.g., respondents did not understand what was asked of them). Thus, the statistical treatment of missing item scores should be considered in combination with other courses of action.

Types of Missing Item Scores

The next example item was taken from a questionnaire that measures people's tendency to cry (Vingerhoets & Cornelius, 2001):

I cry when I experience opposition from someone else
Never ☐ ☐ ☐ ☐ ☐ ☐ ☐ Always

In general, for a particular respondent or group of respondents nonresponse may depend on:

1. The missing value on that item. For example, belonging to the right-most "Always" group may imply a stronger nonresponse tendency than belonging to the left-most "Never" group. Consequently, any missing data method based on available item scores would underestimate the missing value.

2. Values of the other observed items or covariates. For example, for men it may be more difficult to give a rating in the three boxes to the right (showing endorsement or partial endorsement) than for women. Thus, gender has a relation with item score missingness and this can be used for estimating the missing item scores.

3. Values of variables that were not part of the investigation. For example, nonresponse may depend on the unobserved verbal comprehension level of the respondents or on their general intelligence. This kind of missingness is relevant only if the unobserved variables are related to the observed variables, and have an impact on the answers to the items in the test.

Item scores are missing completely at random (MCAR; see Little & Rubin, 1987, pp. 14-17) if the cause of missingness is unrelated to the missing values themselves, the scores on the other observed items and the observed covariates, and the scores on unobserved variables. Thus, item score missingness is ignorable because the observed data are a random sample from the complete data. After listwise deletion, statistical analysis of the resulting smaller data set results in less statistical accuracy and less power when testing hypotheses, but unbiased parameter estimates.

When nonresponse depends on another variable from the data set, but not on values of the item itself or on unobserved variables, item scores are missing at random (MAR; see Little & Rubin, 1987, pp. 14-17). For example, men may find it more difficult to answer "always" to the example item than women, resulting in more missing item scores for men. The distributions of

item scores are different between men and women, but the distributions are the same for respondents and nonrespondents in both groups. Note that within the groups of men and women we have MCAR (given that no other variables relate to item score missingness). This means that if, for example, a regression analysis contains gender as a dummy variable the estimates of the regression coefficients for both groups are unbiased. Thus, when missingness is of the MAR type it is also ignorable.

When missingness is not MCAR or MAR, the observed data are not a random sample from the original sample or from subsamples. Thus, the missingness is nonignorable. In practice, a researcher can only observe that item scores are missing. To decide whether item score missingness is ignorable or nonignorable, he/she has to rely on the pattern of item score missingness in the data matrix, **X**. When he/she finds no relationships to other observed variables, he/she may decide that the missingness is of the MCAR type. When a relationship to other observed variables is found, he/she may use these variables as covariates in multivariate analyses or to impute scores. When a more complex pattern of relationships is found, item score missingness may be considered nonignorable. A reasonable solution is to impute scores when the imputation method is backed up by robustness studies (e.g., Bernaards & Sijtsma, 2000, for factor analysis of rating scale data; and Huisman & Molenaar, 2001, in the context of test construction).

Missing Item Score Analysis

Theory for Analysis of the Whole Data Matrix

The scores on the J items are collected in J random variables X_j, $j = 1, ..., J$. For respondent i ($i = 1, ..., n$), the J item scores, X_{ij}, have realizations x_{ij}. Let M_{ij} be an indicator of a missing score with realization m_{ij}; $m_{ij} = 0$ if X_{ij} is observed and $m_{ij} = 1$ if X_{ij} is missing. These missingness indicators are collected in an $n \times J$ matrix **M**.

Huisman (1999; Kim & Curry, 1978) investigated whether or not the pattern of missingness in the data matrix **X** is unrelated among items. This is called *random missingness* and is defined as follows. Frequency counts of observed missing scores and expected missing scores are compared, given statistical independence of the missingness between the items. Thus, whether a respondent misses the score on item j is unrelated to whether he (or she) misses the score on item k. Items j and k may have different proportions of missing scores. A more restricted assumption, to be used later on, is that the proportions for all J items are equal, as is typical of MCAR. It may be noted that MCAR implies random missingness.

Huisman (1999) classifies each respondent in the sample into one of $J + 2$ classes: (a) *NM* (No Missing): none of the item scores in a pattern are missing; (b) M_j (Missing on item j): a score is missing only on item j; and (c) *MM* (Multiple Missings): scores are missing on at least two items.

Let $q_j = \Sigma_i M_{ij}/n$ be the proportion of missing values on item j in the sample and let $p_j = 1 - q_j$ be the proportion of observed values on item j. Then, under the assumption of random missingness (as defined above), the expected values for *NM*, M_j, and *MM* are

$$E(NM) = n\prod_{j=1}^{J} p_j;$$

$$E(M_j) = \frac{q_j}{p_j} E(NM); \text{ and}$$

$$E(MM) = n - E(NM) - \sum_{j=1}^{J} E(M_j).$$

The observed frequencies in these $J + 2$ classes are denoted by $O(NM)$, $O(M_j)$, and $O(MM)$. Under the assumption of random missingness Pearson's chi-squared statistic,

$$(1) \quad X^2 = \frac{[O(NM) - E(NM)]^2}{E(NM)} + \sum_{j=1}^{J} \frac{\left[O(M_j) - E(M_j)\right]^2}{E(M_j)} + \frac{[O(MM) - E(MM)]^2}{E(MM)},$$

has a χ^2 distribution with $J + 1$ degrees of freedom as $n \rightarrow \infty$ (see, e.g., Agresti, 1990, pp. 44-45). For $n = 8$, Table 1 shows an incomplete data matrix **X** and the corresponding missingness indicator matrix, **M**. This example is used to calculate the X^2 statistic (Equation 1). Because $p_2 = 1$, we have that $E(M_2) = 0$; this is a structural zero, which is ignored in the computation of X^2 at the cost of one degree of freedom. Table 2 shows the observed and the expected frequencies that result in $X^2 = 1.65$ ($df = 5$). Given the small sample size, it makes no sense to draw any inferences on the basis of the outcome.

Robustness of X^2 Statistic for Small Samples

Problem Definition. The robustness of Huisman's (1999) asymptotic test for small (realistic) samples is important. For similar expected frequencies in each of the $J + 1$ classes, Koehler and Larntz (1980) found that

Table 1

Artificial Data Matrix **X** Containing Missing Scores (Blanks), and Corresponding Missingness Indicator Matrix **M**

Case	Variables					Missingness Indicators				
	X_1	X_2	X_3	X_4	X_5	M_1	M_2	M_3	M_4	M_5
1	2	1	1			0	0	0	1	1
2	3	5	4	5	5	0	0	0	0	0
3	4	3		3	4	0	0	1	0	0
4	1	1	1	3	2	0	0	0	0	0
5		3	3		4	1	0	0	1	0
6	5	5	3		5	0	0	0	1	0
7	1	3	2	2	2	0	0	0	0	0
8	3	3	1	2		0	0	0	0	1
q_j						.125	.0	.125	.375	.25
p_j						.875	1.0	.875	.625	.75

Table 2

Expected and Observed Frequencies for the Data in Table 1

Frequency	Expected	Observed
NM	2.87	3
M_1	0.41	0
M_3	0.41	1
M_4	1.72	1
M_5	0.96	1
MM	1.63	2

statistic X^2 approximates a chi-squared distribution when $n > \sqrt{10 \times (J + 1)}$, given that $n > 10$ and $J > 2$. This rule does not apply when expected frequencies are dissimilar, as in Huisman's derivation of the expected frequencies assuming random missingness. Now, if we assume the stronger null-hypothesis of MCAR, under Huisman's classification the expected frequencies depend on the mean proportion of missing values, $\bar{q} = \sum q_j / J$, and test length, J, resulting in

$$
\begin{aligned}
E(NM) &= n(1 - \bar{q})^J, \\
E(M_j) &= n\bar{q}(1 - \bar{q})^{J-1}, \text{ and} \\
E(MM) &= n\left[1 - (1 - \bar{q})^J - J\bar{q}(1 - \bar{q})^{J-1}\right].
\end{aligned}
$$

(2)

Note that as with Koehler and Larntz's study the $E(M_j)$s are all equal, but that the other two expected frequencies are different from this value. Because of this dissimilarity, we investigated whether the conditions given by Koehler and Larntz for X^2 to approximate a chi-squared statistic also hold here.

Simulation Study on Robustness. For different combinations of n, $\bar{q}$, and J (i.e., $n = 10, 20, 50, 100, 200, 500, 1000, 2000$; $\bar{q} = 0.01, 0.05, 0.10$; and $J = 10, 20$), missingness indicator matrices, **M**, were simulated. The elements of **M** were drawn from the multinomial distribution with probabilities based on Equation 2. Table 3 shows the multinomial distributions of the expected scores for $\bar{q} = 0.01, 0.05, 0.10$; and $J = 10, 20$ (these distributions are the same for different n). The last two rows give evenly distributed classes, corresponding to Koehler and Larntz's (1980) study. The last two columns give the sample sizes needed such that the Type I error rate approximates well the nominal significance level, $\alpha = 0.05$, under a chi-squared distribution. Column $n_{accurate}$ gives the sample sizes that resulted in a relatively close approximation (Type I error rates between 0.050 and 0.055), and Column $n_{inaccurate}$ gives the sample sizes that resulted in less accurate Type I error rates (between 0.050 and 0.080). If the sample size was smaller than indicated in the last two columns, the Type I error rate was less accurate and always exceeded 0.05. This means that for smaller sample sizes MCAR was supported too often. Table 3 shows that the required sample size for X^2 is smallest when the expected proportions are evenly distributed, as in Koehler and Larntz's study. Moreover, if the $E(M_j)$s are small (e.g., when $\bar{q} = 0.01$) the required sample size increases rapidly.

K. Sijtsma and L. van der Ark

Table 3

Distribution of the Multinomial Resulting from Huisman's Classification, and Sample Sizes Needed to Approximate the Correct Nominal Type I Error Rate

$\bar{q}$	J	$E(NM)/n$	$E(M_j)/n$	$E(MM)/n$	n_{accurate}	$n_{\text{inaccurate}}$
.01	10	.9044	.0091	.0046	1000	100
	20	.8179	.0083	.0161	1000	100
.05	10	.5987	.0315	.0863	100	20
	20	.3585	.0187	.2675	500	50
.10	10	.3487	.0387	.2543	100	20
	20	.1216	.0135	.6084	500	100
	10	.0833	.0833	.0833	50	10
	20	.0455	.0455	.0455	100	20

Discussion. For a test of reasonable length ($J = 20$) and for little nonresponse ($\bar{q} = 0.01$, as in a rather well-controlled data collection procedure), $n = 1000$ is needed for the Type I error rate to match the nominal error rate. For higher percentages of nonresponse, smaller samples ($n = 500$) will yield this result. Given the limitations of this simulation, as a rule of the thumb for trusting the p-values of the chi-squared statistics one can compute various power divergence statistics (Cressie & Read, 1984) and compare the differences. Power divergence statistics for Huisman's classification are given by,

$$S = \frac{2}{\lambda(\lambda+1)}\left[\sum_{j=1}^{J}O\left(M_j\right)\left\{\left[\frac{O\left(M_j\right)}{E\left(M_j\right)}\right]^{\lambda}\right\} + O(NM)\left\{\left[\frac{O(NM)}{E(NM)}\right]^{\lambda}\right\} + O(MM)\left\{\left[\frac{O(MM)}{E(MM)}\right]^{\lambda}\right\}\right].$$

The power divergence statistic S equals X^2 for $\lambda = 1$, the likelihood ratio statistic G^2 for $\lambda \to 0$, Neyman's modified X^2 for $\lambda = -2$, the Cressie-Read statistic (CR) for $\lambda = 2/3$, and the Freeman-Tukey statistic for $\lambda = -1/2$ (see, e.g., Agresti, 1990, p. 249). Asymptotically, all power divergence statistics converge to a chi-squared distribution. Differences between the various power divergence statistics may occur when the sample size is too small, and then the resulting p-values should be mistrusted. Koehler and Larntz (1980;

also, see Von Davier, 1997) noted that for sparse multinomials X^2 converges faster to a chi-squared distribution than G^2.

Analysis of Missingness for Individual Items

Knowing which items in particular caused nonignorable nonresponse may lead to the rejection of such items. Huisman (1999) suggested to first split the sample into respondents with $m_j = 0$ and $m_j = 1$, and then compare these subgroups with respect to the distributions of item scores on each of the other $J - 1$ items using χ^2 tests, or the item means using t-tests or nonparametric tests. Another possibility, assuming MAR, is to check the expectation that the correlation matrix of the missingness indicator matrix $\mathbf{M}$, $\mathbf{R}_M$, is an identity matrix. Non-zero correlations provide evidence of nonignorable missingness for (some of) the items involved. Significant correlations of covariates with missingness variables, M_j, may provide indications of the causes of nonresponse, and this may help to remedy the missingness. In general, nonsignificant correlations and differences between distributions indicate MAR, and significant results indicate nonignorability.

Treatment of Missing Item Scores

Simple Imputation Methods

Person Mean Imputation. Huisman (1999) and Bernaards and Sijtsma (1999) imputed for all missing item scores of respondent i his/her mean on the available items, denoted PM_i. Suppose that for respondent i, J_i items ($J_i < J$) are available of which the indices are collected in set $A_{(i)}$; then,

$$PM_i = \frac{\displaystyle\sum_{j \in A_{(i)}} X_{ij}}{J_i}; \; PM_i \in \mathbb{R}.$$

For binary (0/1) item scores, we impute for each missing value another random draw from the Bernoulli distribution with parameter PM_i. For ordered polytomous (0, ..., k) item scores, for example, for $k = 4$ and $PM_i = 2.56$, we impute item score 2 if the value of the random draw from the Bernoulli distribution with parameter 0.56 was 0 and item score 3 otherwise. Method PM corrects for score differences between respondents but not for score differences between items.

Two-Way Imputation. Bernaards and Sijtsma (2000) corrected method PM for the item mean score and the overall score level of the group. The item mean, IM_j, is defined as the mean score of the observed scores on item j, and the overall mean, OM, is defined as the mean of all observed scores in the data matrix, $\mathbf{X}$. Then for missing item score (i, j),

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

Integer scores are imputed following the procedure outlined for method PM.

New Imputation Methods Using Nonparametric Regression

General Introduction. Let $\mathbf{\theta}$ denote the vector of latent trait parameters necessary to describe the data structure in data matrix $\mathbf{X}$, and let $\mathbf{\zeta}_j$ be a vector of possibly multidimensional item parameters, such as the item locations and discriminations. IRT models all have the form $P(X_j = x_j|\mathbf{\theta}; \mathbf{\zeta}_j) = f(\mathbf{\theta}; \mathbf{\zeta}_j)$; that is, the probability of having a score, x_j, on item j, known as the item response function (IRF), depends on respondent and item parameters. By choosing a particular function for $f(\mathbf{\theta}; \mathbf{\zeta}_j)$, such as a logistic regression function (e.g., Baker, 1992; Fischer & Molenaar, 1995), even for incomplete data, $\mathbf{X}$, the item parameters may be estimated from the likelihood of the model,

$$L(\text{model}) = P(\mathbf{X} \mid \text{model}) = \prod_{i=1}^{n} \prod_{j=1}^{J} P\left(X_{ij} = x_{ij} \mid \mathbf{\theta}_i; \mathbf{\zeta}_j\right).$$

Assuming that the estimates $\hat{\mathbf{\zeta}}_j$ are the true parameters, the respondent parameters, $\mathbf{\theta}_i$, are estimated next (e.g., Baker, 1992). Suppose, imputation is used to produce a complete data matrix for further analysis. First, the estimates $\hat{\mathbf{\theta}}_i$ and $\hat{\mathbf{\zeta}}_j$ are inserted in the IRT model, such that $P(X_{ij} = x_{ij}|\hat{\mathbf{\theta}}_i; \hat{\mathbf{\zeta}}_j)$ is obtained. Then, for binary scores, a draw from a Bernoulli distribution with estimated probability $P(X_{ij} = 1|\hat{\mathbf{\theta}}_i; \hat{\mathbf{\zeta}}_j)$ can be imputed for missing value (i, j); and for polytomous items, a draw from a multinomial distribution with parameters $P(X_{ij} = x_j|\hat{\mathbf{\theta}}_i; \hat{\mathbf{\zeta}}_j)$, $x_j = 1, ..., k$, can be imputed for missing value (i, j). This is called model-based imputation.

Obviously, if a particular IRT model represents the hypothesis of interest and is also used for imputation, the resulting data set is biased in favor of this hypothesis. Here, we propose two imputation methods based on the IRF, that are based on nonparametric regression, and do not impose restrictions on the

shape of the IRF and not explicitly on the dimensionality of measurement. For example, if a researcher wants to fit the Rasch (1960) model (with $\theta = \theta$, a scalar; and $\zeta_j = \delta_j$, a location parameter) to his/her data, and he/she uses one of our item score imputation methods, the resulting complete data matrix is not explicitly biased in favor of the Rasch model as it would be if that model itself were used for item score imputation.

Two remarks are in order. First, although the two methods to be proposed do not explicitly make assumptions about the dimensionality of the data, they are likely to be more successful when the data are unidimensional. The reason is that, like methods PM and TW, they use total person scores like PM_i based on the summation of the items. Strong multidimensionality produces a correlation structure among the items (with many 0 or almost 0 correlations) that renders such total scores inadequate summaries of the information available. Second, more than, say, linear regression, an IRT context is suited for missing item score imputation in tests and questionnaires because it models data from variables that are allowed to correlate highly, thus avoiding multicollinearity. Further, IRT models are flexible in that the error component of the model is heteroscedastic. Also, given the highly discrete nature of item scores the nonlinearity of IRT is helpful.

Response-Function Imputation. In the nonparametric IRT context adopted here, for convenience we assume that the IRF is a function of a scalar latent trait θ, and that it varies across items, but we do not assume a latent item parameter vector, ζ_j, that can be estimated from the likelihood. See Van der Ark and Sijtsma (in press) for the use of several of the methods discussed here when data are explicitly multidimensional.

Define a person summary score $X_+ = \sum_{j=1}^{J} X_j$. Let the restscore, $R_{(-j)} = X_+ - X_j$, be the total score on $J - 1$ binary items from the test except item j (Junker & Sijtsma, 2000). Restscore $R_{(-j)}$ is used as a proxy for θ (e.g., Hemker, Sijtsma, Molenaar, & Junker, 1997; Junker, 1993; Sijtsma & Molenaar, 2002). We estimate $P(X_j = 1|\theta)$ by means of $P[X_j = 1|R_{(-j)}]$, or $P_j[R_{(-j)}]$, for short. This observable probability is the item-rest regression (Junker & Sijtsma, 2000). Using only those respondents that have completely observed data, probability $P_j[R_{(-j)} = r]$ can be estimated as the fraction of the subgroup with rest score $R_{(-j)} = r$, that have item j correct. We use this fraction to impute scores as follows.

1. Consider a respondent who has missing scores on item j and possibly on other items as well. As before, the indices of the J_i available items are collected in set $A_{(i)}$. Multiplying PM_i by $J - 1$, we obtain a real, $\hat{R}_{(-j)i}$, that estimates respondent is integer restscore, $R_{(-j)i}$, based on complete data; that is,

K. Sijtsma and L. van der Ark

$$\hat{R}_{(-j)i} = PM_i \times (J - 1); \ \hat{R}_{(-j)i} \in \mathbb{R}.$$

2. Insert $\hat{R}_{(-j)i}$ in the ordering, $R_{(-j)} = 0, ..., J - 1$. If estimate $\hat{R}_{(-j)i}$ is an integer, probability $\hat{P}_j\left[\hat{R}_{(-j)i}\right]$ can be obtained as the fraction of respondents with restscore $\hat{R}_{(-j)i}$ that have item j correct. If estimate $\hat{R}_{(-j)i}$ is a real, it has a left neighbor, $R_{(-j)}^{\text{left}}$, and a right neighbor, $R_{(-j)}^{\text{right}}$. From the sample of completely observed respondents we have the corresponding probabilities $P_j[R_{(-j)}^{\text{left}}]$ and $P_j[R_{(-j)}^{\text{right}}]$. For respondent i, the probability $P_j[\hat{R}_{(-j)i}]$ is estimated by linear interpolation between $P_j[R_{(-j)}^{\text{left}}]$ and $P_j[R_{(-j)}^{\text{right}}]$. Noting that $R_{(-j)}^{\text{right}} - R_{(-j)}^{\text{left}} = 1$, the linear interpolation formula is

$$\hat{P}_j\left[\hat{R}_{(-j)i}\right] = P_j[R_{(-j)}^{\text{left}}] + \{P_j[R_{(-j)}^{\text{right}}] - P_j[R_{(-j)}^{\text{left}}]\} \times [\hat{R}_{(-j)i} - R_{(-j)}^{\text{left}}].$$

3. Impute a score in cell (i, j) by randomly drawing from a Bernoulli distribution with parameter $\hat{P}_j\left[\hat{R}_{(-j)i}\right]$.

These three steps are repeated for all missing item scores in **X**. For example, for $J = 5$ let Carol have missing scores on items 1 and 3, and let her have two items correct. Then, Carol's estimated restscore for item 1 (Figure 1, upper panel) equals

$$\hat{R}_{(-1)Carol} = \frac{2}{3} \times (5 - 1) = 2\frac{2}{3}.$$

Assume that $P_1[R_{(-1)}^{\text{left}} = 2] = 0.7$ and that $P_1[R_{(-1)}^{\text{right}} = 3] = 0.85$; then

$$\hat{P}_1\left[\hat{R}_{(-1)Carol} = 2\frac{2}{3}\right] = 0.7 + 0.15 \times \frac{2}{3} = 0.8.$$

This method is called *Response-Function* (RF) imputation. The algorithm contains reasonable provisions to take care of small or even empty rest score groups (following a methodology used by Molenaar & Sijtsma, 2000, p. 67), and other data problems. Explaining them in detail would take too much space. Note that method RF takes differences between respondents into account through the rest score groups and differences between items through the item-rest regressions (cf. method TW).
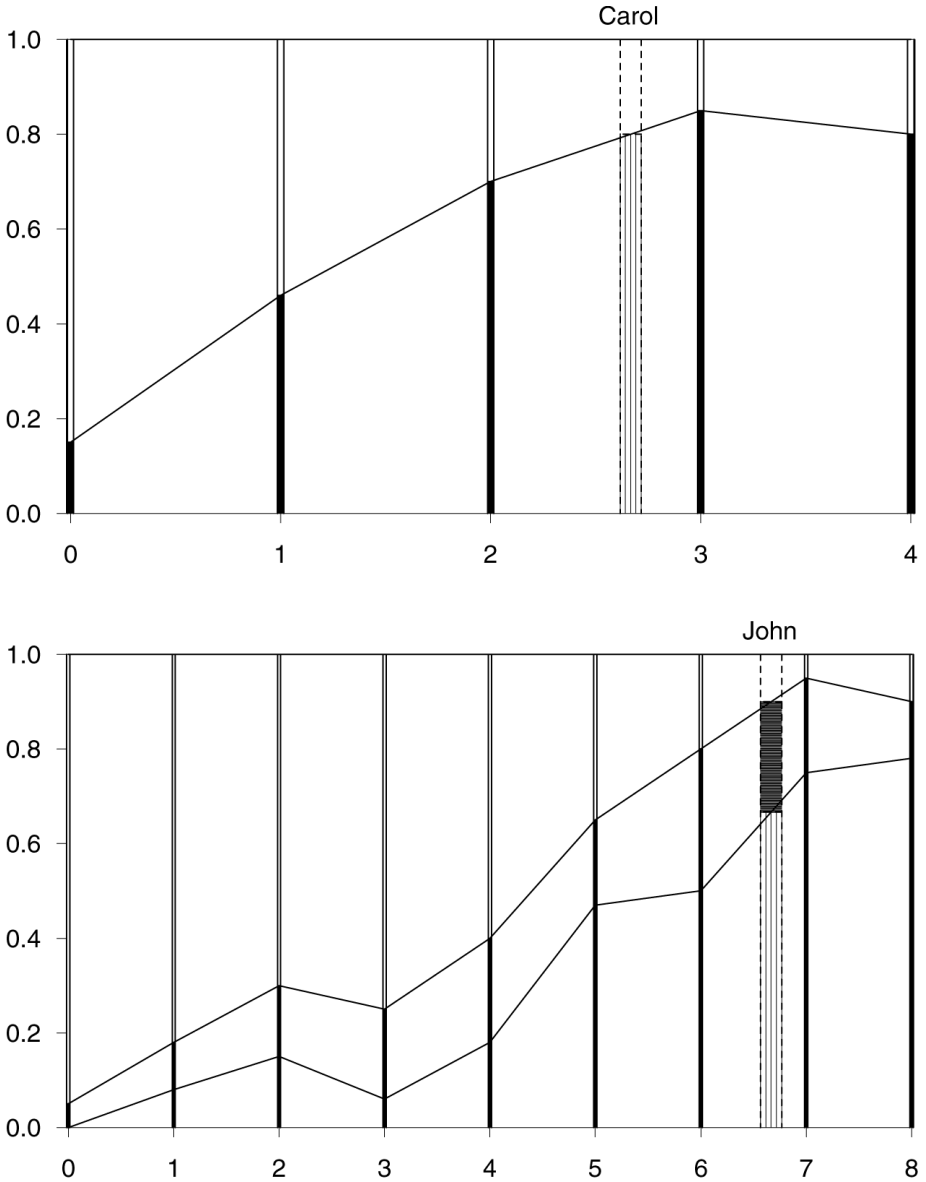
Figure 1

Item-rest regressions for dichotomous items (upper panel) and polytomous items ($k = 2$; lower panel), and linearly interpolated response probabilities (corresponding to differently marked columns) for Carol (upper panel; scores 0, 1) and John (lower panel; scores 0, 1, 2)

For polytomous items, response probabilities, $P(X_j \geq x_j|\theta)$, $x_j = 0, ..., k$, are estimated using procedures outlined above for dichotomous items. Figure 1 (lower panel) illustrates how method RF can be generalized to an item with three ordered answer categories. For each item, we have response functions $P(X_j \geq 1|\theta)$ and $P(X_j \geq 2|\theta)$, that are estimated using $P[X_j \geq 1|R_{(-j)}]$ and $P[X_j \geq 2|R_{(-j)}]$, respectively (Junker, 1993; Molenaar & Sijtsma, 2000).

For example, for $J = 5$ let John have missing scores on items 1 and 3, and scores 2, 2, 1 on the three remaining items. Then, John's estimated restscore for item 1 is

$$\hat{R}_{(-1)John} = \frac{5}{3} \times (5-1) = 6\frac{2}{3}.$$

Because for each item there are two response functions, interpolation has to be done twice. Let $P[X_1 \geq 1|R_{(-1)} = 6] = 0.80$, $P[X_1 \geq 2|R_{(-1)} = 6] = 0.50$, $P[X_1 \geq 1|R_{(-1)} = 7] = 0.95$, and $P[X_1 \geq 2|R_{(-1)} = 7] = 0.75$; then

$$\hat{P}\left[X_1 \geq 1 | \hat{R}_{(-1)John} = 6\frac{2}{3}\right] = 0.80 + 0.15 \times \frac{2}{3} = 0.9$$

$$\hat{P}\left[X_1 \geq 2 | \hat{R}_{(-1)John} = 6\frac{2}{3}\right] = 0.50 + 0.25 \times \frac{2}{3} = 0.67.$$

Figure 1 (lower panel) shows RF imputation of John's score on item 1. The response probabilities are shown by the bars (white bar for $x = 0$; black bar for $x = 1$; and grey bar for $x = 2$). Integer item scores are drawn from a multinomial distribution with category probabilities corresponding to the length of the bars in Figure 1.

Mean Response-Function Imputation. The second new imputation method uses the means of the J item-rest regressions and thus ignores item differences (cf. method PM). It is denoted mean response-function imputation (method MRF). Because joining small restscore groups for one item (e.g., the groups $R_{(-j)} = 0, 1, 2$) may render the resulting joined group incomparable to restscore groups of other items (e.g., the joint groups $R_{[-(j+1)]} = 2, 3$), we avoid this problem by following the next steps.

1. Estimate all J item-rest regressions, each based on all J rest-score groups (unless a group is empty; then it is ignored). The restscore group-size for group $R_{(-j)} = r$ is denoted n_{rj}.

2. For each rest-score value, $R_{(-j)} = r$, take the mean of the J success probabilities, $P_j[R_{(-j)} = r], j = 1, ..., J$ (or a number smaller than J: see step 1); and weigh each success probability by

$$n_{rj} / \sum_{j=1}^{J} n_{rj} .$$

Denote this mean by P_r, defined as,

$$P_r = \sum_{j=1}^{J} \frac{n_{rj} \times P_j \left[R_{(-j)} = r \right]}{\sum_{j=1}^{J} n_{rj}}, \; r = 0,1,\ldots,J.$$

3. The estimate P_r of the mean of the item-rest regressions is used for imputing scores.

Note that once we have estimated the restscore $\hat{R}_{(-j)i}$ and determined the corresponding success probability using one of the two methods outlined previously, we may impute missing values by repeatedly drawing from the same Bernoulli distribution that has that particular success probability as a parameter. Generalization to polytomous items can be done similarly to the generalization of method RF.

An Empirical Data Example

Method

Example Data. We used data from a questionnaire ($J = 23$) asking people how they responded to determinants (memories, thoughts, images, experiences, situations) that could make them cry or weep (Vingerhoets & Cornelius, 2001). Respondents were either Australians, Belgians or Indians. Each item was scored 0 (determinant does not or rarely elicit crying) or 1 (determinant more often or almost always elicits crying). The original data matrix also contained incomplete cases, but we used as a point of departure the $n = 705$ complete cases, collected in the data matrix **X**. We also created six versions of **X** that each contained missing item scores using the following methodology.

Simulation Study Design. For three matrices, fixed proportions ($\overline{q} = .01$, .05, and .10) of ignorable (MCAR) item score missingness were simulated, and for the other three matrices nonignorable item score missingness was simulated. Ignorable missingness was simulated by randomly deleting item scores using a fixed probability for a score being missing. Nonignorable item score missingness was simulated as follows. From the original data it was determined that Australians, Belgians and Indians had missing item scores according to the ratio $m_A : m_B : m_I = 1 : 4 : 8$. Items were weighted by social desirability indices, $s_1, ..., s_{23}$, ranging from 0.4 (most social conventions would require respondents to cry), to 10 (most social conventions would prohibit respondents to cry). Item score missingness was then simulated by using for each entry of **X** the probability $P(M_{ij} = 1) = m_i s_j (1 + x_{ij})c$, where c is a constant chosen such that the desired proportion of item score missingness is obtained. Thus, the probability $P(M_{ij} = 1)$ was highest for Indians and lowest for Australians; higher the more an item's content stimulated a socially desirable answer; and higher when the item score was 1 rather than 0.

Each of the methods PM, TW, RF, and MRF were used to impute scores in each empty cell of each of the six incomplete versions of **X**. For each incomplete version of **X**, this resulted in four imputed data matrices. Then, for each matrix we used Huisman's (1999) global test and we checked $\mathbf{R}_M$ to identify possibly deviant items. These analyses gave evidence whether these methods produced the correct conclusion about the ignorability or the nonignorability of the item score missingness.

Outcome Statistics. For **X** and each of the 24 imputed data matrices based on **X**, we calculated quality indices, well known in classical test theory (Lord & Novick, 1968), Mokken scale analysis (Mokken & Lewis, 1982; Sijtsma & Molenaar, 2002), and the Rasch (1960) model (also see Fischer & Molenaar, 1995), respectively: (a) Cronbach's (1951) alpha, used here as a lower bound to the reliability of the test score, X_+; (b) Mokken's (1971) scalability coefficient, H, which is an index for the precision of person ordering using X_+; and (c) the Rasch model chi-squared goodness-of-fit statistics, R_{1c} (Glas & Verhelst, 1995) and Q_2 (Van den Wollenberg, 1982). Statistic R_{1c} tests whether the response functions of the J items are logistic with the same slope against the alternative that they deviate from these conditions, and statistic Q_2 tests whether the test is unidimensional against the alternative of multidimensionality. These coefficients and statistics were compared among nonignorable and ignorable missingness, percentages of missingness, and imputation methods.

Results

For MCAR, the null hypothesis of random missingness across cells of the data matrix was not rejected for any percentage of item score missingness, using either X^2, G^2, or CR (Table 4). For nonignorable item score missingness, for $\bar{q} = 0.01$ the sample size ($n = 705$) was too small to detect this nonignorability by any of the three statistics. This is consistent with the results of the simulation study on minimally required sample sizes (Table 3). The null hypothesis was rejected correctly for $\bar{q} = 0.05$ and $\bar{q} = 0.10$.

The correlation matrix $\mathbf{R}_M$ contained 253 unique (but mutually dependent) correlations. Because of the skewness of the marginals in the two-by-two frequency tables, Fisher's exact test (e.g., Agresti, 1990, pp. 59-66) was used to test for independence (implying $\rho = 0$). The last row of Table 4 gives the percentage of significant results at the $\alpha = .05$ level. Because tests were dependent, we compared percentages of rejections of the null hypothesis between ignorable and nonignorable item score missingness. The bottom line of Table 4 shows that the percentage of significant Fisher exact test statistics was higher for nonignorable item score missingness than for ignorable item score missingness.

Table 4

Power Divergence Statistics X^2, G^2, and CR ($df = 24$), Type I Error Rate and Percentage of Significant Fisher Exact Tests (Last Row), for Ignorable and Nonignorable Item Score Missingness, for $\bar{q} = 0.01$, 0.05, and 0.10

Statistic		Ignorable (MCAR)			Nonignorable		
	$\bar{q}$:	.01	.05	.10	.01	.05	.10
X^2		7.15	11.36	16.06	21.52	56.73	229.11
		.9999	.9861	.8859	.6080	.0002	.0000
G^2		8.32	10.57	18.35	25.45	62.30	170.18
		.9978	.9918	.7856	.3812	.0000	.0000
CR		7.48	11.70	16.70	22.20	57.90	205.12
		.9995	.9885	.8611	.5673	.0001	.0000
Sign. Fisher test		2.8%	3.2%	4.0%	4.7%	7.9%	18.2%

Table 5

Student's *t*-test and Type I Error Rate for Difference in *PM* Means of Respondents and Nonrespondents ($\bar{q}$ = .05) to Item *j*, for Nonignorable (Nonign.Miss.) and Ignorable Item Score Missingness (Ign.Miss.).

Item	Ign.Miss.		Nonign.Miss.		Item	Ign.Miss.		Nonign.Miss.	
	t	*p*	*t*	*p*		*t*	*p*	*t*	*p*
1	−0.08	.9364	2.52	.0119	13	−2.22	.0265	1.52	.1284
2	−2.14	.0324	1.87	.0614	14	−0.45	.6499	1.32	.1844
3	1.15	.2517	2.82	.0048	15	0.08	.9313	**3.08**	**.0020**
4	−0.67	.5029	2.60	.0093	16	−0.44	.6601	2.91	.0037
5	0.79	.4393	2.77	.0057	17	−0.69	.4922	2.71	.0068
6	0.32	.7560	**3.57**	**.0004**	18	0.58	.5563	**3.85**	**.0001**
7	0.89	.3723	1.48	.1370	19	0.16	.8735	2.73	.0065
8	−1.86	.0627	2.03	.0434	20	0.03	.9758	**3.46**	**.0006**
9	−1.19	.2327	2.94	.0033	21	−0.77	.4427	1.90	.0575
10	−0.26	.7945	2.48	.0132	22	1.47	.1432	**4.70**	**.0000**
11	−0.94	.3447	1.29	.1959	23	2.30	.0421	**4.18**	**.0000**
12	−1.03	.3015	2.99	.0029					

Note. Significant Differences are in bold face; Bonferroni Alpha = 0.0022.

Other local analysis of item score missingness was done by comparing the mean *PM*s of nonrespondents and respondents to item *j*, for all items. To avoid tedious detailed results, the discussion is limited to the data matrices with $\bar{q}$ = 0.05 ignorable missing item scores (MCAR) and $\bar{q}$ = 0.05 nonignorable missing item scores, respectively. Table 5 shows that for nonignorable item score missingness data, for six items the mean *PM*s of both groups differed significantly (two-sided; using Bonferroni correction, α = .05/23 = .0022). Thus, item score missingness was found indeed to be nonignorable. For ignorable item score missingness data there were no significant mean differences between mean *PM*s. This correctly indicated ignorable nonresponse.

Table 6 shows that the bias in Cronbach's alpha ranged from −.024 to .011 (alpha found for **X** was .924; theoretical maximum is 1). Method RF showed almost no bias. In general, imputed data sets showed little variation

Table 6

Bias in Cronbach's Alpha, for Ignorable (MCAR) and Nonignorable Missingness Mechanisms, $\bar{q}$ = .01, .05, and .10, and Imputation Methods PM, TW, RF, and MRF; Cronbach's Alpha = .924 for Complete Data

Method		Missingness Mechanism					
		Ignorable			Nonignorable		
	$\bar{q}$:	.01	.05	.10	.01	.05	.10
PM		.001	.005	.011	.001	.005	.010
TW		.001	.005	.010	.001	.004	.008
RF		.000	.000	−.003	.000	.000	.000
MRF		.000	−.006	−.024	.000	−.002	−.014

Table 7

Bias in coefficient H, for Ignorable (MCAR) and Nonignorable Missingness Mechanisms, $\bar{q}$ = .01, .05, and .10, and Imputation Methods PM, TW, RF, and MRF; H = .448 for Complete Data

Method		Missingness Mechanism					
		Ignorable			Nonignorable		
	$\bar{q}$:	.01	.05	.10	.01	.05	.10
PM		.004	.018	.038	.004	.018	.041
TW		.005	.023	.045	.005	.023	.046
RF		.001	.000	−.014	.002	.007	.005
MRF		.000	−.028	−.091	−.002	−.011	−.056

between ignorable and nonignorable item score missingness and different values of $\bar{q}$. Table 7 shows that the bias in scalability coefficient H ranged from −.091 to .046 (H value found for **X** was .448; theoretical maximum is 1). There was almost no variation in the bias of H for $\bar{q}$ = 0.01, more variation for $\bar{q}$ = 0.05 and the most for $\bar{q}$ = 0.10. Method RF was the least biased.

Methods PM and TW had greater positive bias the higher the percentage of nonresponse, and method MRF had greater negative bias the higher the percentage of nonresponse.

For statistic R_{1c}, the value found (157 with $df = 88$) for data matrix $\mathbf{X}$ means that the 23 response functions are not all logistic with the same slopes, as the Rasch model predicts. In general, method RF was closest to this target value (Table 8). Each of the other methods showed at least one result that was much too low (but also led to the rejection of the null hypothesis). The more interesting result was that for nonignorable item score missingness the imputation methods produced results that are hardly distinguishable from those found for ignorable item score missingness. For statistic Q_2, the value found was 2112 with $df = 1150$, meaning that the 23 items together seem to measure several latent traits instead of one. For methods PM and TW, the Q_2 values were always too high and they were higher the greater the percentage of item score missingness (Table 9). For method RF, a similar pattern of results was found for ignorable item score missingness. For method MRF, in this case an opposite pattern was found with Q_2 values that were too low. This pattern was also found for methods RF and MRF for nonignorable item score missingness. In general, methods PM and TW seem to favor the conclusion that multidimensionality holds (too high Type I error), whereas method MRF seems to favor the conclusion that the test is unidimensional (too low Type I error). The results for method RF are less clear.

Table 8

Rasch Analysis Bias Results for R_{1c}, for Ignorable (MCAR) and Nonignorable Missingness Mechanisms, $\bar{q} = .01, .05,$ and $.10,$ and Imputation Methods PM, TW, RF, and MRF; $R_{1c} = 157$ ($df = 88$) for Complete Data

Method		Missingness Mechanism					
		Ignorable			Nonignorable		
	$\bar{q}:$	.01	.05	.10	.01	.05	.10
PM		−5	−11	−25	−10	−15	6
TW		−6	−18	−37	−9	−12	1
RF		−10	−13	−1	−5	−12	−5
MRF		−8	−12	−25	−10	−16	−21

Note. $J = 23$; due to Rasch model estimation properties n varies from 620 to 643 across cells.

Table 9

Rasch Analysis Bias Results for Q_2, for Ignorable (MCAR) and Nonignorable Missingness Mechanisms, $\bar{q} = .01, .05,$ and $.10,$ and Imputation Methods PM, TW, RF, and MRF; $Q_2 = 2112$ ($df = 1150$) for Complete Data

| | | Missingness Mechanism | | | | | |
| | | Ignorable | | | Nonignorable | | |
Method	$\bar{q}$:	.01	.05	.10	.01	.05	.10
PM		140	387	947	208	544	587
TW		24	239	1053	159	883	2119
RF		122	450	755	271	−216	−279
MRF		114	−353	−427	−122	−349	−448

Note. $J = 23$; due to Rasch model estimation properties n varies from 620 to 643 across cells.

Discussion

In our one-data set example, Huisman's (1999) overall test statistic was effective to detect both simulated ignorable and nonignorable item score missingness correctly, given an appropriate sample size. When ignorable item score missingness is found, we may have confidence that single imputation or another method probably will not greatly invalidate the data. Alternative classifications of missingness patterns than those used for Huisman's method may provide additional ways to test for MCAR or MAR. Under MCAR any classification of the respondents or the items should fit. Possibly useful classifications are those based on meaningful covariates, such as gender, social-economic status and age.

Imputation methods PM and TW are so simple that they can be explained easily to researchers that are not statistically trained. Also, they are easy to compute using major software packages such as SPSS and SAS. Methods RF and MRF use the response function, estimated nonparametrically from the fully observed respondents, thus ignoring the common and more restrictive assumptions typical of IRT models. These methods are also rather easy to explain, but their computation can be cumbersome. This is true especially for method RF when the restscore groups are small and have to

be joined. A simple computer program called impute.exe with the four imputation methods implemented for both dichotomous and polytomous items can be obtained from the authors at *http://www.uvt.nl/faculteiten/fsw/ organisatie/departementen/mto/software2.html*. The software was written in Borland Pascal 7.0. The maximum order of data matrix **X** for which the program works has not yet been explored.

Method RF was superior to methods PM, TW, and MRF in estimating the alpha and *H* coefficients, and the Rasch model statistics R_{1c} and Q_2. Method TW produced higher percentages of hits than the other methods, but this resulted sometimes in estimates of alpha and *H* that were too high. Method RF may produce unstable results for small numbers of fully observed respondents. Consequently, the estimates of the response probabilities may be inaccurate. Method TW may be more stable, and may be preferred for smaller sample sizes. Methods RF and TW may be also be useful when item score missingness is nonignorable. A reviewer suggested that deleting cases from the analysis with more than, say, half of the item scores missing may further improve results. This is a possible topic for future research. Finally, each of the methods probably works best when the data are unidimensional. Multidimensionality is addressed by Van der Ark and Sijtsma (in press).

The error introduced in the data by single imputation may be too small, resulting in standard errors that are too small (Little & Rubin, 1987, p. 256). The analysis of test data usually is more involved, however, calculating large numbers of statistics, testing many hypotheses, and selecting items based on such calculations. Moreover, test construction has a cyclic character, leaving out items in one cycle, re-analyzing the data for remaining items, leaving out another item as well or re-selecting a previously rejected item in another cycle, and so on. It would be interesting to see how multiple imputation (e.g., Rubin, 1991) can help to obtain more stable conclusions for item analysis. This is a topic for future research.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Baker, F. B. (1992). *Item response theory. Parameter estimation techniques*. New York: Marcel Dekker.

Bernaards, C. A. & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research*, *34*, 277-313.

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.

Cressie, N. & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society*, *Series B*, *46*, 440-464.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *Series B*, *39*, 1-38.

Fischer, G. H. & Molenaar, I. W. (1995, Eds.). *Rasch models. Foundations, recent developments, and applications*. New York: Springer.

Glas, C. A. W. & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 69-95). New York: Springer.

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

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

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

Junker, B. W. & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*, 65-81.

Kim, J. O. & Curry, J. (1978). The treatment of missing data in multivariate analysis. In D. F. Alwin (Ed.), *Survey design and analysis* (pp. 91-116). London: Sage.

Koehler, K. & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, *75*, 336-344.

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

Little, R. J. A. & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39-75). New York: Plenum.

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

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

Mokken, R. J. & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6*, 417-430.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.

Rubin, D. B. (1991). EM and beyond. *Psychometrika*, *56*, 241-254.

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

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

Smits, N., Mellenbergh, G. J., & Vorst, H. C. M. (2002). Alternative missing data techniques to grade point average: Imputing unavailable grades. *Journal of Educational Measurement*, *39*, 187-206.

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

Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika*, *47*, 123-140.

Van der Ark, L. A. & Sijtsma, K. (in press). The effect of missing data imputation on Mokken scale analysis. In L. A. Van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences*. Mahwah NJ: Erlbaum.

Vingerhoets, A. J. J. M. & Cornelius, R. R. (Eds.) (2001). *Adult crying. A biopsychosocial approach*. Hove, UK: Brunner-Routledge.

Von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data — Results of a Monte Carlo Study. *Methods of Psychological Research Online*. Retrieved January 3, 2002, from the World Wide Web: *http://www.mpr-online.de*.