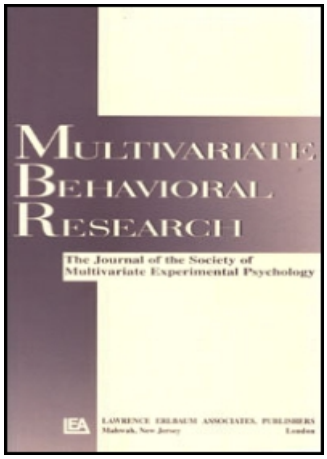
Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results

van Ginkel, J.R.; van der Ark, L.A.; Sijtsma, K.

Published in:
Multivariate Behavioral Research

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record



Multivariate Behavioral Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653673

Multiple Imputation of Item Scores in Test and Questionnaire Data, and Influence on Psychometric Results

Joost R. van Ginkel [a]; L. Andries van der Ark [a]; Klaas Sijtsma [a]
[a] Tilburg University, The Netherlands

Online Publication Date: 29 June 2007
To cite this Article: van Ginkel, Joost R., van der Ark, L. Andries and Sijtsma, Klaas (2007) 'Multiple Imputation of Item Scores in Test and Questionnaire Data, and Influence on Psychometric Results', Multivariate Behavioral Research, 42:2, 387 - 414
To link to this article: DOI: 10.1080/00273170701360803
URL: http://dx.doi.org/10.1080/00273170701360803

MULTIVARIATE BEHAVIORAL RESEARCH, *42*(2), 387–414

Multiple Imputation of Item Scores in Test and Questionnaire Data, and Influence on Psychometric Results

Joost R. van Ginkel, L. Andries van der Ark, and Klaas Sijtsma
Tilburg University, The Netherlands

The performance of five simple multiple imputation methods for dealing with missing data were compared. In addition, random imputation and multivariate normal imputation were used as lower and upper benchmark, respectively. Test data were simulated and item scores were deleted such that they were either missing completely at random, missing at random, or not missing at random. Cronbach's alpha, Loevinger's scalability coefficient H, and the item cluster solution from Mokken scale analysis of the complete data were compared with the corresponding results based on the data including imputed scores. The multiple-imputation methods, two-way with normally distributed errors, corrected item-mean substitution with normally distributed errors, and response function, produced discrepancies in Cronbach's coefficient alpha, Loevinger's coefficient H, and the cluster solution from Mokken scale analysis, that were smaller than the discrepancies in upper benchmark multivariate normal imputation.

Tests and questionnaire data consist of the scores of N subjects on J items. Together these items measure one or more psychological traits. Scores in test and questionnaire data can be missing for several reasons. For example, a respondent accidentally skipped an item or even a whole page of items, he/she found a particular question too personal to answer, or he/she became bored filling out the test or questionnaire and skipped some questions on purpose.

Let $\mathbf{X}$ be an incomplete data matrix of size $N \times J$ with an observed part $\mathbf{X}_{obs}$ and a missing part $\mathbf{X}_{mis}$, so that $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$. Let $\mathbf{R}$ be an $N \times J$

Correspondence concerning this article should be addressed to Joost van Ginkel, Department of Methodology and Statistics, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: j.r.vanginkel@uvt.nl

indicator matrix of which an element equals one if the corresponding score in $\mathbf{X}$ is observed, and zero if the corresponding score in $\mathbf{X}$ is missing. Furthermore, let $\boldsymbol{\xi}$ be an unknown parameter vector that characterizes the missingness mechanism. Missingness mechanisms can be divided into three categories: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Little & Rubin, 2002, p. 12; Rubin, 1976). MCAR is formalized as

$$P(\mathbf{R} \mid \mathbf{X}_{obs}, \mathbf{X}_{mis}, \boldsymbol{\xi}) = P(\mathbf{R} \mid \boldsymbol{\xi}). \tag{1}$$

MCAR means that the missing scores in the data are a random sample of all scores in the data, and that the missingness does not depend on either the observed scores ($\mathbf{X}_{obs}$) or values of the missing scores ($\mathbf{X}_{mis}$).

MAR means that the missing values depend on the observed scores,

$$P(\mathbf{R} \mid \mathbf{X}_{obs}, \mathbf{X}_{mis}, \boldsymbol{\xi}) = P(\mathbf{R} \mid \mathbf{X}_{obs}, \boldsymbol{\xi}). \tag{2}$$

For example, if gender is observed for all subjects it may be found that men find it more difficult or embarrassing to answer a question about depression than women do. Therefore, the probability of not answering such a question is higher for men than for women. If in addition the missing scores within each covariate class are a random sample of all scores, the scores are said to be MAR.

Any missingness mechanism that cannot be formalized as in Equation (1) or Equation (2) is NMAR. NMAR means that the missingness on variable X either depends on variables that are not part of the investigation, or on the missing score on variable X itself, or both. If people, who are depressed, have a higher probability of not responding to a question about depression than people who are not depressed, the missingness is NMAR.

A popular method for dealing with missing data is listwise deletion. This method entails the removal of all cases with at least one missing score from the statistical analysis. Listwise deletion reduces the sample size and therefore results in a loss of power. Moreover, if listwise deletion results in only a few complete cases statistical analyses may be awkward. Additionally, when the missingness mechanism is not MCAR, the resulting sample may be biased.

Another procedure of missing-data handling is imputation of scores to replace missing data. Examples are hot-deck imputation (Rubin, 1987, p. 9) and regression imputation (Rubin, 1987, pp. 166–169). Hot-deck imputation matches to each nonrespondent another respondent who resembles the nonrespondent on variables that are observed for both, and donates the observed scores of this respondent to the missing scores of the nonrespondent (Bernaards et al., 2003; Huisman, 1998). Regression imputation estimates scores under a regression model, using one or more independent variables to predict the most likely scores (Bernaards et al., 2003; Smits, Mellenbergh, & Vorst, 2002).

In multiple imputation (Rubin, 1987, p. 2), an imputation method is applied w times to the same incomplete data set, so as to produce w different plausible versions of the complete data set. Each of these w data sets is analyzed by standard complete-data methods and the results are combined into one overall estimate of the statistics of interest. This way, the uncertainty about the imputed values is taken into account when drawing a final conclusion. Software programs for multiple imputation under the multivariate normal model are, for example, NORM (Schafer, 1998) and the missing-data module of S-plus 6 for Windows (2001). The method used by NORM is also available in SAS 8.1, in the procedure PROC MI (Yuan, 2000). The program AMELIA by King, Honaker, Joseph, and Scheve (2001a,b) imputes scores according to a multivariate normal model, but uses another computational method (Schafer & Graham, 2002). The stand-alone software package SOLAS (2001) performs hot-deck imputation and multiple imputation that relies on regression models (Schafer & Graham, 2002). Multiple imputation under the saturated logistic model and the general location model can be applied by means of the missing-data module of S-plus 6 for Windows (2001) (Schafer & Graham, 2002).

Simulation studies on the performance of multiple-imputation methods have been conducted (Ezzati-Rice et al., 1995; Graham & Schafer, 1999; Schafer, 1997; Schafer et al., 1996). These studies showed that these methods produce small bias in statistical analyses, and are robust against departures of the data from the imputation model. Most of these methods require the use of algorithms like EM (Dempster, Laird, & Rubin, 1977; Rubin, 1991) or data augmentation (Tanner & Wong, 1987), that appear complicated to social scientists who lack enough training in statistics and programming to effectively apply these methods. Instead, these researchers often resort to listwise deletion.

Alternatively, simpler methods have been developed, such as corrected item-mean substitution (CIMS; Huisman, 1998, p. 96), two-way imputation (TW; Bernaards & Sijtsma, 2000), and response-function imputation (RF; Sijtsma & Van der Ark, 2003). Subroutines in SPSS (2004) for methods TW, RF, and CIMS have been made available by van Ginkel and van der Ark (2005a,b). These methods are easy to comprehend and can be useful alternatives to listwise deletion. The question is to what extent the simplicity of these methods goes at the expense of their performance. The aim of this study was to determine the extent to which multiple-imputation versions of simple methods produced discrepancies in results of statistical techniques, and the extent to which they produced stable results over replicated data sets. Moreover, the aim was to compare the results of these methods to those obtained by means of lower and upper benchmark methods.

Bernaards and Sijtsma (1999, 2000) found that factor loadings could be recovered well using simple single-imputation methods. Huisman (1998) used real data to study the effects of nine imputation methods on the discrepancy

in Cronbach's (1951) alpha and Loevinger's (1948) H, and found that method CIMS performed best in recovering these statistics. Smits (2003, chap. 3) investigated the influence of simple and more advanced single-imputation methods on the reliability, the test score, and the external validity of a test. Van der Ark and Sijtsma (2005) used multiple-imputation methods to recover item clusters from Mokken (1971) scale analysis in real data sets.

In the present study, we investigated the influence of six imputation methods on Cronbach's alpha, coefficient H, and the cluster solution from Mokken scale analysis. The results of the analyses of completely observed data sets were compared with the results of analyses of the same data sets but with some scores missing according to some specified research design, and replaced by imputed scores. The data were simulated following methodology used by Bernaards and Sijtsma (1999, 2000). Unlike the studies of Bernaards and Sijtsma (1999, 2000) and Huisman (1998, chap. 5 & chap. 6), multiple-imputation versions of imputation methods were studied.

METHOD

Data sets were simulated according to an item response theory (IRT) model proposed by Kelderman and Rijkes (1994). In these data sets, denoted *original data*, missingness was simulated according to either MCAR, MAR, or NMAR. The resulting data sets were denoted *incomplete data*. Next, the missing scores were estimated according to multiple-imputation versions of six imputation methods, and the resulting data sets were denoted *completed data*. The results of Cronbach's alpha, coefficient H, and the cluster solution from Mokken scale analysis based on the original data were compared with the results based on the completed data. Differences were denoted *discrepancies*.

Imputation Methods

Random imputation (RI). Let the random variable for the score on item j be denoted X_j, with integer values $x_j = 0, \ldots, m$. RI inserts an integer item score for missing item scores. This value is drawn at random from a uniform distribution of integers $0, \ldots, m$. RI was used as a lower benchmark.

Two-way imputation (TW). Method TW (Bernaards & Sijtsma, 2000) corrects both for a person effect and an item effect. Let PM_i be the mean of the observed item scores of person i, IM_j the mean of the observed item scores of item j, and OM the overall mean of all observed item scores; then in cell

(i, j) of the data matrix, define

$$TW_{ij} = PM_i + IM_j - OM \tag{3}$$

A random component is added to the result of Equation (3) as follows: If TW_{ij} is a real number that lies between integers a and b, it is rounded to a with probability $|TW_{ij} - b|$ or to b with probability $|TW_{ij} - a|$ (Sijtsma & Van der Ark, 2003), and the result is imputed in cell (i, j). If TW_{ij} is outside the range of the scores $0, \ldots, m$, it is rounded to the nearest feasible score.

Two-way with normally distributed errors (TW-E). Bernaards and Sijtsma (2000) added a random error to TW_{ij}, denoted ε_{ij}, which was drawn from a normal distribution with zero mean and a variance σ_ε^2. In order to obtain values of ε_{ij}, first the expected item scores are computed for all observed scores by means of Equation (3). Second, let *obs* denote the set of all observed cells in data matrix **X**, and let #*obs* be the size of set *obs*. The sample error variance S_ε^2 is computed as

$$S_\varepsilon^2 = \sum_{i,j \in obs} \sum (X_{ij} - TW_{ij})^2 / (\#obs - 1).$$

Third, ε_{ij} is drawn from $N(0, S_\varepsilon^2)$. The imputed value in cell (i, j) then equals

$$TW_{ij}(E) = TW_{ij} + \varepsilon_{ij}.$$

$TW_{ij}(E)$ is rounded to the nearest integer within the range of the scores $0, \ldots, m$.

Corrected item-mean substitution with normally distributed errors (CIMS-E). Let $obs(i)$ be the set of all observed cells in **X** for person i and let $\#obs(i)$ be the size of set $obs(i)$. Then $CIMS_{ij}$ is defined as

$$CIMS_{ij} = \left(\frac{PM_i}{\frac{1}{\#obs(i)} \sum_{j \in obs(i)} IM_j} \right) \times IM_j$$

(Huisman, 1998, p. 96; also, see Bernaards & Sijtsma, 2000). Thus, the item mean is corrected for person i's score level relative to the mean of the items to which he/she responded. Normally distributed errors are added to $CIMS_{ij}$ using a procedure similar to the procedure used for adding normally distributed errors in method TW-E. The final result is rounded to the nearest integer within the range $0, \ldots, m$.

Response-function imputation (RF). In IRT, the regression of the score on item j on latent variable θ, $P(X_j = x \mid \theta)$, is called the response function. Method RF (Sijtsma & Van der Ark, 2003) uses the estimated response function to impute item scores. Restscore $R_{(-j)}$ (this is the total score on $J - 1$ items without X_j) is used as an estimate of person parameter θ (Junker & Sijtsma, 2000), and the response function is estimated by means of $P[X_j = x \mid R_{(-j)}]$. Method RF has three steps.

1. The restscore of respondent i on item j is estimated by means of

$$\hat{R}_{(-j)i} = PM_i \times [J - 1].$$

   If respondent i has no missing values, $\hat{R}_{(-j)i} = R_{(-j)i} = \sum_{k \neq j}^{J} X_{ik}$ is an integer, but if respondent i has missing values $\hat{R}_{(-j)i}$ need not be an integer.
2. Probability $P[X_j = x \mid R_{(-j)} = r]$ is estimated for $x = 0, \ldots, m$ and $r = 0, \ldots, m(J - 1)$, by dividing the number of respondents with both $X_j = x$ and $\hat{R}_{(-j)} = r$ by the number of respondents with $\hat{R}_{(-j)} = r$. If r is not an integer and the nearest integers are a and b, such that $a < r < b$, then $P[X_j = x \mid R_{(-j)} = r]$ is estimated by linear interpolation of $P[X_j = x \mid R_{(-j)} = a]$ and $P[X_j = x \mid R_{(-j)} = b]$. See Sijtsma and Van der Ark (2003) for details.
3. An integer score is drawn from a multinomial distribution with category probabilities corresponding to the estimated probabilities $P[X_j = x \mid R_{(-j)} = r]$. This integer score is imputed for a missing score of person i on item j, with restscore $\hat{R}_{(-j)i}$.

When restscore groups contain few observations, adjacent restscore groups are joined until resulting groups exceed an acceptable minimum size, denoted *minsize*. In a pilot study, it was found that *minsize* = 10 was the optimal value for estimating the response function that, while adequately balancing bias and accuracy, recovered the estimates of Cronbach's alpha, coefficient H, and the cluster solution from Mokken scale analysis best.

Multivariate normal imputation (MNI). Method MNI assumes that the data are a random sample from a multivariate normal distribution. An iterative procedure is used to obtain the distribution of the missing item scores, given the observed item scores and the model parameters. This procedure is known as data augmentation (Schafer, 1997; Tanner & Wong, 1987). Initial estimates of the model parameters are obtained by means of the EM algorithm. EM posterior modes estimates serve as the starting values for the data augmentation chain. Finally, scores are imputed by randomly drawing values from the conditional

distribution $P(\mathbf{X}_{mis} \mid \mathbf{X}_{obs})$. MNI was implemented using the missing-data library in S-plus 6 for Windows (2001). The imputed scores were rounded to the nearest integer within the range of $0, \ldots, m$. We used method MNI as an upper benchmark because it is a well-known method with readily available software, and simulation studies indicated that the method works well.

Note that a saturated logistic model (Schafer, 1997, chap. 7 & chap. 8) may be a more logical upper benchmark because item scores in test and questionnaire data are discrete. However, estimating the parameters of a logistic model requires the evaluation of a contingency table with $(m+1)^J$ cells, which makes the logistic model inappropriate for test and questionnaire data sets with large numbers of items. Van der Ark and Sijtsma (2005) found that the missing-data procedure in S-plus could not estimate a logistic model for a data set with 17 items. Graham and Schafer (1999) found that method MNI is robust against departures from the multivariate normal model.

Simulating the Original Data

All respondents in the population had scores on a two-dimensional latent variable, $\boldsymbol{\theta}$, driving the item responses, and a binary score on an observed covariate Y. Both covariate scores had equal probability, $P(Y = 1) = P(Y = 2) = .50$. The latent variable had a bivariate normal distribution with mean vectors $\boldsymbol{\mu}_1 = [-0.25, -0.25]$ for $Y = 1$, and mean vector $\boldsymbol{\mu}_2 = [0.25, 0.25]$ for $Y = 2$. The covariance matrix (which is also the correlation matrix) was in both classes

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \\ \rho & 1 \end{bmatrix}.$$

Responses to J items with $m + 1$ ordered answer categories were generated using the multidimensional polytomous latent trait (MPLT) model (Kelderman & Rijkes, 1994).

Let θ_{iq} $(i = 1, \ldots, N;\ q = 1, \ldots, Q)$ be the score of respondent i on latent variable q; let ψ_{jqx} $(j = 1, \ldots, J;\ q = 1, \ldots, Q,\ x = 0, \ldots, m)$ be the separation parameter of item j, latent variable q, and answer category x; and let B_{jqx} $(j = 1, \ldots, J;\ q = 1, \ldots, Q;\ x = 0, \ldots, m)$ be the (nonnegative) discrimination parameter of item j, latent variable q, and answer category x. The MPLT model is defined as

$$P(X_{ij} = x \mid \theta_{i1}, \ldots, \theta_{iQ}) = \frac{\exp\left[\sum_{q=1}^{Q}(\theta_{iq} - \psi_{jqx})B_{jqx}\right]}{\sum_{y=0}^{x}\left\{\exp\left[\sum_{q=1}^{Q}(\theta_{iq} - \psi_{jqy})B_{jqy}\right]\right\}}. \quad (4)$$

Parameters B_{jq0} and ψ_{jq0} must be set to 0 to ensure uniqueness of the parameters.

The following factors were considered for simulating of the original data:

Test length. The test length was fixed at $J = 20$ items.

Number of answer categories. The number of answer categories was either two (dichotomous items) or five (polytomous items).

Sample sizes. The sample size were $N = 200$ and $N = 1000$, representing small and large samples, respectively.

Correlation between latent variables. The correlation ρ was varied to be 0, .24, and .50 (these values were based on Bernaards & Sijtsma, 1999).

Discrimination parameters for polytomous items. In the main design, item sets were either unidimensional (meaning one θ in Equation (4)), or consisted of ten items that were mainly driven by one latent variable (θ_1) and to a lesser degree by another latent variable (θ_2), and ten other items that were mainly driven by θ_2 and to a lesser degree by θ_1. In a specialized design, the first ten items were completely driven by θ_1 and the other ten items were completely driven by θ_2. The degree to which item responses were driven by latent variables was manipulated by means of the discrimination parameters, B_{jqx} (in the simulation study the discrimination parameters were equivalent for categories $1, \ldots, m$; therefore, the subscript x will be dropped.)

For unidimensional tests, for an item j, discrimination parameters B_{j1} and B_{j2} were either both equal to 0.25 or both equal to 1, summing up to 0.5 or 2, respectively (choices loosely based on Thissen & Wainer, 1982). This means that responses to items were driven in the same degree by the two latent variables, either weakly ($B = 0.25$) or strongly ($B = 1$). This is expressed by the ratio of B_{j1} and B_{j2}, which is called a *latent-variable ratio* and denoted Mix 1:1. The responses to all items in a test may be driven in the same degree by two latent variables, such as reading ability and arithmatic ability. Mathematically, this is an instance of unidimensionality because all items measure the two latent variables in the same ratio.

In the second dimensionality configuration, for fixed item j, parameters B_{j1} and B_{j2} were unequal, expressing dependence on the latent variables in different degrees. For the first ten items, B_{j1} was three times B_{j2}. For the last ten items this ratio was reversed. Numerically, for the same item the two B parameters were either 0.125 and 0.375 (summing up to 0.5; this represents weak discrimination) or 0.5 and 1.5 (summing up to 2; this represents strong discrimination). The ratio of the B parameters was 3:1 for the first ten items and 1:3 for the last

TABLE 1
Discrimination Parameters, B_{jq}, of All ISRFs of the Items

	Mix 1:0		*Mix 3:1*		*Mix 1:1*	
Items	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
1, 3, 5, 7, 9	0.5	0	0.375	0.125	0.25	0.25
2, 4, 6, 8, 10	2	0	1.5	0.5	1	1
11, 13, 15, 17, 19	0	2	0.5	1.5	1	1
12, 14, 16, 18, 20	0	0.5	0.125	0.375	0.25	0.25

ten items. This latent-variable ratio is denoted Mix 3:1. For example, the first ten items may be influenced more by reading ability than by arithmetic ability, and for the last ten items this may be reversed.

The third latent-variable ratio (to be treated in a specialized design) had the B parameter of one latent variable set to 0 and of the other set to either 0.5 or 2. For the first ten items $B_{j2} = 0$ and for the last ten items $B_{j1} = 0$. Thus, the ratio of the B parameters was 1:0 for the first ten items and 0:1 for the last ten items. This latent-variable ratio is denoted Mix 1:0. See Bernaards and Sijtsma (1999) for the use of the same three latent-variable ratios. For the first ten items in each data set, items with even numbers had B_{j1} and B_{j2} values adding up to 2, and items with odd numbers had B_{j1} and B_{j2} values adding up to 0.5. For the last ten items, this was reversed. Table 1 shows the discrimination parameters for all items, latent-variable ratios, and latent variables.

Separation parameters for polytomous items. Because the polytomous items had five answer categories, each item had four adjacent response functions defined by Equation (4). The distance between two adjacent separation parameters, $\psi_{jq,x-1}$ and ψ_{jqx}, was 0.5, for all j; $q = 1, 2$; and $x = 1, 2, 3, 4$. These values fell within the interval $(-3, 3)$, which Thissen and Wainer (1982) considered to be realistic, given a standard normal distribution of θ. Because the responses to the items were driven by two latent variables and because there were four adjacent response functions per latent variable, each item had eight ψ parameters. The values of the separation parameters are given in Table 2. The separation parameters of the first ten items for θ_1 were equal to the separation parameters of the last ten items for θ_2. Likewise, the separation parameters of the last ten items for θ_1 were equal to the separation parameters of the first ten items for latent θ_2. This way, within the same test items had varying difficulty. For example, if an item is difficult with respect to θ_1 but easy with respect to θ_2, the four values of the separation parameters for θ_1 were higher on average than the four values of the separation parameters for θ_2.

TABLE 2
Separation Parameters, ψ_{jqx}, of Polytomous Items

Items	ψ_{j11}	ψ_{j12}	ψ_{j13}	ψ_{j14}	ψ_{j21}	ψ_{j22}	ψ_{j23}	ψ_{j24}
1, 2, 19, 20	−2.75	−2.25	−1.75	−1.25	1.25	1.75	2.25	2.75
3, 4, 17, 18	−1.75	−1.25	−0.75	−0.25	0.25	0.75	1.25	1.75
5, 6, 15, 16	−0.75	−0.25	0.25	0.75	−0.75	−0.25	0.25	0.75
7, 8, 13, 14	0.25	0.75	1.25	1.75	−1.75	−1.25	−0.75	−0.25
9, 10, 11, 12	1.25	1.75	2.25	2.75	−2.75	−2.25	−1.75	−1.25

Item parameters for dichotomous items. The discrimination parameters for dichotomous items had the same values as those for polytomous items; see Table 1. For dichotomous item j, the separation parameter ψ_{jqx} was chosen such that it was equal to the mean of the four ψ parameters of polytomous item j. This resulted in integer ψ_{jqx} values ranging from −2 to 2.

Simulating Missing Item Scores: Incomplete Data

After simulating the original data sets, incomplete data sets were created by removing some values from the original data. Two steps were taken to achieve this result:

1. The percentages of missingness that were studied were 5 and 15. For example, for $N = 200$, $J = 20$ and 5% missing scores, 200 item scores were selected to be missing.
2. Missingness was simulated by removing item scores from the data following particular missingness mechanisms. Covariate variable Y was always observed. For MCAR all item scores had equal probability of being missing. For MAR the probability of item scores being missing was twice as high for subjects within covariate class $Y = 1$ as for subjects within covariate class $Y = 2$. Using these relative probabilities, a sample of scores was removed from the complete data. Finally, NMAR was simulated as follows: Let trunc$(m/2)$ be a cut-off value that divides item scores into low scores and high scores (Van der Ark & Sijtsma, 2005). For scores above this cut-off value, the probability of being missing was twice as high as for scores below this cut-off value. Using these relative probabilities, a sample of item scores was removed from the complete data.

Imputing Item Scores: Completed Data

After simulating the incomplete data, completed data sets were created. Two aspects of the imputation process were varied.

Imputation method. Missing data were estimated according to six imputation methods: methods RI, TW, TW-E, RF, CIMS-E, and MNI.

Including or excluding the covariate. In using the imputation methods, the covariate may either be included or excluded. When missingness depends on the covariate and this covariate is used in the imputation procedure, missingness is MAR. When the covariate is excluded, missingness becomes NMAR because it depends on a variable that is not used in the imputation procedure.

Methods RI, TW, TW-E, RF, and CIMS-E, were applied to each covariate class separately. For method MNI, covariate Y was included in the multivariate normal model estimated from the data. When the covariate was excluded, methods RI, TW, TW-E, RF, and CIMS-E were applied to the whole dataset, and for method MNI the covariate was not included in the multivariate normal model. Both options were studied.

Designs

Main Design

The six factors relevant to the main study were: (1) Latent-variable ratio (Mix 1:1 and Mix 3:1); (2) Sample size ($N = 200$ and $N = 1000$); (3) Percentage of missingness (5% and 15%); (4) Missingness mechanism (MCAR, MAR, and NMAR); (5) Imputation method (RI, TW, TW-E, RF, CIMS-E, and MNI), and (6) Covariate treatment (included, excluded). The correlation between the latent variables was .24 throughout. The number of answer categories was 5, the number of items was 20, and the number of imputations in multiple imputation was 5. The design consisted of 2 (latent-variable ratio) $\times$ 2 (sample size) $\times$ 2 (percentage of missingness) $\times$ 3 (missingness mechanism) $\times$ 6 (imputation method) $\times$ 2 (covariate treatment) $=$ 288 cells. In each cell 100 replicated original data sets, indexed by v, were drawn. Table 3 gives an overview of the factors and the fixed design characteristics.

Specialized Designs

The four factors held constant in the specialized designs were sample size ($N = 1000$), percentage of missingness (5%), missingness mechanism (MAR), and covariate treatment (it was included in the imputation procedure). The following factors were varied.

Correlation between latent variables. In practice, latent variables are often correlated. In this specialized design, performance of the imputation methods was studied for different correlations between latent variables. Following Bernaards and Sijtsma (2000), the correlation between latent variables was 0,

TABLE 3
Factors and Fixed Characteristics of the Main Design

Factors	*Levels*
Latent-variable ratio	Mix 1:1, Mix 3:1
Sample size	200, 1000
Missingness percentage	5%, 15%
Missingness mechanism	MCAR, MAR, NMAR
Imputation methods	RI, TW, TW-E, RF, CIMS-E, MNI
Covariate	Included, Excluded
Fixed Design Characteristics	*Value*
Number of latent variables	2; bivariate normal
Correlation between latent variables	.24
Number of items	20
Number of answer categories	5
Number of imputations	5
Separation parameter, ψ_{jqx}	Fixed per item, see Table 2

.24, and .50. Only latent-variable ratio Mix 3:1 was considered. This design had 3 (correlation) $\times$ 6 (imputation method) $=$ 18 cells.

Latent-variable ratios. According to Sijtsma and Van der Ark (2003), imputation methods produce the smallest discrepancies when a test is unidimensional. In the main design, latent-variable ratios Mix 1:1 and Mix 3:1 were studied, representing unidimensional tests and two-dimensional tests, respectively. To study the effects of larger deviations from unidimensionality, Mix 1:0 was investigated in a specialized design. The correlation between latent variables was .24. All imputation methods were studied, resulting in a completely crossed 3 (latent-variable ratio) $\times$ 6 (imputation method) design with 18 cells.

Number of answer categories. In this design, dichotomous items were studied, and the results were compared with the results based on polytomous items. The number of answer categories could either be 2 or 5. Only latent-variable ratio Mix 1:1 was considered, and the correlation between the latent variables was .24. A completely crossed 2 (number of answer categories) $\times$ 6 (imputation method) design (12 cells) was used.

Dependent Variables

The dependent variables were the discrepancy in Cronbach's (1951) alpha, coefficient H, and in the cluster solution from Mokken (1971) scale analysis. Cron-

bach's alpha is reported in almost every study that uses tests or questionnaires; Loevinger's H is an easy-to-use coefficient that is important in nonparametric IRT for evaluating the scalability of a set of items (Sijtsma & Molenaar, 2002, pp. 149–150, provide a list of 22 studies in which H was used, many of which had incomplete data); and Mokken's item selection cluster algorithm is used for investigating the dimensionality of test and questionnaire data (see, e.g., Van Abswoude, Van der Ark, & Sijtsma, 2004). Together these dependent variables provide a good impression of the degree of success of the proposed imputation methods.

Discrepancy in Cronbach's alpha. Within each design cell, Cronbach's alpha was computed for each original data set (indexed $v = 1, \ldots, 100$), and denoted $\alpha_{or,v}$; and for each of the five completed data sets corresponding to original data set v. The mean of these five values was denoted $\alpha_{imp,v}$. The discrepancy in alpha was defined as $\alpha_{imp,v} - \alpha_{or,v}$, and served as dependent variable in an ANOVA. The mean (M) and standard deviation (SD) of the discrepancy were computed within each design cell across 100 replications. The tables show results that have been aggregated across design cells.

Discrepancy in coefficient H. Let $Cov(X_j, X_k)$ be the covariance between items j and k, and $Cov(X_j, X_k)_{\max}$ the maximum covariance given the marginal distributions of the bivariate frequency table for the item scores. The H coefficient, which is a scalability coefficient for all J items together, is defined as

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

(Mokken, 1971, pp. 148–153, 1997; Sijtsma & Molenaar, 2002, pp. 49–64). Similar to discrepancy in Cronbach's alpha, the discrepancy in coefficient H in the vth replication is defined as $H_{imp,v} - H_{or,v}$. This was the dependent variable in an ANOVA. The mean (M) and standard deviation (SD) of the discrepancy were computed within each design cell across 100 replications. The results in the tables have been aggregated across design cells.

Discrepancy in cluster solution from Mokken scale analysis. Mokken (1971) scale analysis is a method for test construction based on nonparametric item response theory (Boomsma, Van Duijn & Snijders, 2001; Sijtsma & Molenaar, 2002; Van der Linden & Hambleton, 1997). It may be used for exploratory

test construction. Exploratory test construction selects one or more scales from the data, and uses the H coefficient as a selection criterion. The algorithm for the selection of items into clusters is contained in the computer program MSP (Molenaar & Sijtsma, 2000). The discrepancy in the cluster solution, to be denoted *cluster discrepancy*, was determined as follows: For each original data matrix, the five replicated completed data matrices yielded five cluster solutions of which one or more could be different from the others. From these five cluster solutions, one modal cluster solution was obtained, which was compared with the cluster solution based on the original data matrix.

A plausible measure for the discrepancy in the modal cluster solution relative to the original-data cluster solution is the minimum number of items that have to be moved from the modal cluster solution in order to reobtain the original-data cluster solution (Van der Ark & Sijtsma, 2005). In doing this, the nominal cluster numbering is ignored. The minimum number of items to be moved was computed for each data set, and these numbers were used as the dependent variable in logistic regression with binomial counts. The mean (M) cluster discrepancy over replications and the standard deviation (SD) of the cluster discrepancy over replications are reported.

Statistical Analyses

Two full-factorial 2 (latent-variable ratio) × 2 (sample size) × 2 (percentage of missingness) × 3 (missingness mechanism) × 5 (imputation method: TW, TW–E, RF, CIMS–E, MNI) × 2 (include/exclude covariate) ANOVAs had the discrepancies in Cronbach's alpha and coefficient H as dependent variables. Sample size was a between-subjects factor. Percentage of missingness and missingness mechanism were within-subjects factors because different kinds of missingness were simulated per replication in the same original data set. Because each of the five imputation methods plus method RI were applied to the same incomplete data set in each replication, imputation method was also treated as a within-subjects factor. Variation of the factors latent-variable ratio, correlation between latent variables, and the number of answer categories resulted in different data sets. These data sets were mutually dependent because the same seeds were used in each cell of the design. Thus, these factors also had to be treated as within-subjects factors.

A logistic regression with binomial counts was used to analyze the cluster discrepancies because this variable was ordinal (implying that it was not normally distributed). Let y_{vt} be the cluster discrepancy of data set v in design cell t, and let e_{vt} be the maximum number of items that can be incorrectly clustered. Theoretically, for a test of 20 items the cluster discrepancy can be 19 at most. This happens if in the original cluster solution all items form one scale, and in the modal cluster solution of five completed data sets all items remain

unselected (Van der Ark & Sijtsma, 2005); thus, $e_{vt} = 19$. Furthermore, let $\boldsymbol{\beta}$ be a column vector with regression coefficients, and for simulated data set v, let $\mathbf{z}_v$ be a row vector with responses to the independent (dummy) variables. The probability of one incorrectly clustered item is

$$\pi_{t,\mathbf{z}_v} = \frac{\exp(\mathbf{z}_v\boldsymbol{\beta})}{1 + \exp(\mathbf{z}_v\boldsymbol{\beta})}.$$

The logistic regression model with binomial counts is

$$P(y_{vt} \mid \mathbf{z}_v, e_{vt}) = \left[\frac{e_{vt}!}{y_{vt}(e_{vt} - y_{vt})!}\right](\pi_{t,\mathbf{z}_v})^{y_{vt}}(1 - \pi_{t,\mathbf{z}_v})^{e_{vt}-y_{vt}}$$

(see Vermunt & Magidson, 2005b, p. 11). To correct for dependency among measures, primary sampling units were used (Vermunt & Magidson, 2005b, p. 97). As in the ANOVAs for the discrepancy in Cronbach's alpha and coefficient H, sample size was the only factor treated as an independent measure.

We excluded method RI from the analyses because it is a lower benchmark not recommended for practical purposes and we expected that this method would have a large effect on the results of the statistical analyses, which would have a disproportional effect on significance tests. For method RI, only the means and standard deviations of the discrepancy are reported. Leaving out method RI reduced the design from 288 to 240 cells. The ANOVAs were conducted in SPSS (2004), the logistic regressions with binomial counts were conducted in Latent Gold 4.0 (Vermunt & Magidson, 2005a).

RESULTS

ANOVA is robust in some degree against violations of normality (e.g., Stevens, 2002, pp. 261–262) and, in balanced designs, equal variances (e.g., Stevens, 2002, p. 268). Histograms of discrepancy in Cronbach's alpha and coefficient H showed approximate normality. The designs in this study were balanced. Based on this information conclusions from ANOVA were considered valid.

Main Design

Discrepancy in Cronbach's Alpha

Thirty-five effects out of 61 from the ANOVA of the discrepancy in Cronbach's alpha were significant. Following Cohen's (1988) guidelines for effect

TABLE 4
ANOVA for Discrepancy in Cronbach's Alpha and Discrepancy in Coefficient *H*. All *p*-Values Were Less Than .001

Effect	*F*	*df1*	*df2*	η^2
Discrepancy in Cronbach's alpha				
Imputation method	24057.81	4	792	.67***
Percentage missingness	458.99	1	198	.02*
Percentage of missingness × method	16947.73	4	792	.17***
Discrepancy in coefficient H				
Imputation method	55778.37	4	792	.67***
Percentage missingness	735.45	1	198	.02*
Percentage of missingness × method	36295.45	4	792	.19***

*Small effect. **Medium effect. ***Large effect.

sizes, only small ($\eta^2 > .01$), medium ($\eta^2 > .06$), and large effects ($\eta^2 > .14$) are reported. Table 4 (upper panel) shows the effects that have a discernable effect size.

Interaction Effects

Effect of percentage of missingness × imputation method. Table 5 shows that in general, mean discrepancy (M) and standard deviation of discrepancy (SD) were small. For all combinations of percentage of missingness and imputation method, mean discrepancy ranged from $M = -.059$ ($SD = .012$; 15% missingness, method RI) to $M = .015$ ($SD = .002$; 15% missingness, method TW).

The discrepancy in Cronbach's alpha was larger for 15% missingness (upper panel, third and fourth column) than for 5% missingness (upper panel, first two columns). This effect was stronger for imputation methods that already produced a relatively large discrepancy for 5% missingness. Upper benchmark method MNI produced a small discrepancy in Cronbach's alpha for 5% missingness and a somewhat larger discrepancy for 15% missingness. With the exception of methods RI and TW, the simple methods produced smaller discrepancies and also smaller increases in discrepancy in going from 5% to 15% missingness. Methods TW-E and CIMS-E in particular produced almost no discrepancy in results for both 5% and 15% missingness. Methods RF and MNI produced small negative discrepancy for 5% missingness and larger negative discrepancy for 15% missingness (Table 5, middle panel, columns 1–4). Method TW produced

TABLE 5
Mean (*M*) and Standard Deviation (*SD*) of the Discrepancy in Cronbach's Alpha and Discrepancy in Coefficient H for All Combinations of Percentage of Missingness and Imputation Method. Totals Represent Results Aggregated Across Either Imputation Method (Rows), Percentage of Missingness (Columns), or Both (Lower Right Corner in Both Panels). Entries in the Table Must Be Multiplied by 10^{-3}

		Percentage of Missingness					
		5%		*15%*		*Total*	
Dependent Variable	*Method*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Discrepancy in alpha	RI	−18	4	−59	12	−38	22
	TW	5	1	15	2	10	6
	TW-E	0	1	1	2	0	2
	RF	−1	2	−3	3	−2	3
	CIMS-E	0	1	0	2	0	2
	MNI	−1	1	−3	3	−2	2
	Total*	1	3	2	7	1	5
Discrepancy in H	RI	−37	7	−100	14	−68	33
	TW	13	3	41	5	27	15
	TW-E	0	3	0	5	0	4
	RF	−1	3	−6	7	−4	6
	CIMS-E	0	3	0	5	0	4
	MNI	−2	3	−6	6	−4	5
	Total*	2	6	6	19	4	4

*Aggregated across all imputation methods, except method RI.

relatively large positive discrepancy for 5% missingness, and discrepancy that was three times larger for 15% missingness.

For most imputation methods the standard deviation of the discrepancy was close to .001 for 5% missingness, and close to .004 for 15% missingness. This means that if mean discrepancy equals .003 for 15% missingness, then assuming normality the 95% confidence interval of the discrepancy ranges from −.005 to .011.

Main Effects

Effect of percentage of missingness. Table 5 (last row of upper panel, first two columns) shows that the discrepancy in Cronbach's alpha was smaller for 5% missingness than for 15% missingness (last row of upper panel, third and fourth column).

Effect of imputation method. Table 5 (last two columns of upper panel) shows the mean discrepancy and the standard deviation of discrepancy in Cronbach's alpha for all imputation methods, aggregated across all other design factors. Method MNI produced small discrepancy in Cronbach's alpha, but the simple methods TW-E and CIMS-E produced even smaller discrepancy. The positive discrepancy produced by method TW and the negative discrepancy produced by method RI were substantially larger.

Discrepancy in Coefficient *H*

Conclusions about discrepancy in H based on effect sizes and F-values (Table 4, lower panel) were similar to those for Cronbach's alpha. All means and standard deviations of discrepancy in H were approximately two times larger than the corresponding statistics for Cronbach's alpha (Table 5, lower panel). For all combinations of percentage of missingness and imputation method, discrepancy in coefficient H ranged from $M = -.100$ ($SD = .014$; 15% missingness, method RI) to $M = .041$ ($SD = .005$; 15% missingness, method TW).

Cluster Discrepancy

Logistic regression with binomial counts produced many small significant effects; only the means and standard deviations of the largest effects are discussed.

Interaction Effects

Effect of percentage of missingness × imputation method. A Wald-test for individual effects revealed a significant interaction of percentage of missingness and imputation method [$\chi^2(4) = 348.66$, $p < .001$]. Table 6 (last two columns) shows that for all methods the minimum number of items to be moved was larger for 15% missingness than for 5% missingness. Method MNI produced small discrepancy for 5% missingness, and a small increase in discrepancy in going to 15% missingness. For methods TW-E and RF similar results were found. Method TW produced the largest increase in discrepancy (not counting method RI) when going from 5% (second row of upper panel) to 15% missingness (second row of middle panel), followed by method CIMS-E (fifth row of upper panel; fifth row of middle panel). Compared to the theoretical maximum cluster discrepancy of 19, the means and standard deviations reported in Table 6 are small.

Effects of sample size × imputation method. The interaction effect of sample size and imputation method was significant [$\chi^2(4) = 120.22$, $p < .001$].

TABLE 6
Mean (*M*) and Standard Deviation (*SD*) of the Cluster Discrepancy for all Combinations of Percentage of Missingness, Imputation Method, and Sample Size. In Each Panel, Totals Represent Results Aggregated Across Either Imputation Method (Rows), Sample Size (Columns), or Both (Lower Right Corner in Each Panel). Bottom Panel Represents All Totals Aggregated Across Percentage of Missingness

		Sample Size					
		200		*1000*		*Total*	
Percentage Missingness	*Method*	*M*	*SD*	*M*	*SD*	*M*	*SD*
5%	RI	2.08	1.14	1.93	.76	2.01	.97
	TW	1.12	1.03	1.18	.94	1.15	.99
	TW-E	1.01	1.04	.79	1.03	.90	1.04
	RF	1.01	1.00	.79	.99	.90	1.00
	CIMS-E	1.02	1.04	.91	1.09	.97	1.07
	MNI	1.05	1.05	.74	.97	.89	1.02
	Total*	1.04	1.03	.88	1.02	.96	1.03
15%	RI	4.16	1.32	2.95	.97	3.55	1.31
	TW	2.70	1.14	3.45	1.04	3.08	1.16
	TW-E	1.67	1.23	1.42	1.19	1.55	1.22
	RF	1.67	1.23	1.38	1.16	1.52	1.20
	CIMS-E	1.81	1.24	1.81	1.32	1.81	1.28
	MNI	1.80	1.28	1.32	1.20	1.56	1.26
	Total*	1.93	1.29	1.87	1.44	1.90	1.36
Total	RI	3.12	1.61	2.44	1.01	2.78	1.39
	TW	1.91	1.34	2.31	1.51	2.11	1.44
	TW-E	1.34	1.19	1.10	1.16	1.22	1.18
	RF	1.34	1.17	1.08	1.12	1.21	1.15
	CIMS-E	1.41	1.21	1.36	1.29	1.39	1.25
	MNI	1.42	1.23	1.03	1.19	1.22	1.19
	Total*	1.49	1.25	1.38	1.34	1.43	1.30

*Aggregated across all imputation methods, except method RI.

Table 6 (lower panel) shows that with the exception of method TW, the other imputation methods produced smaller discrepancy for $N = 1000$ than for $N = 200$. Methods TW and CIMS–E produced larger discrepancy for $N = 1000$ than for $N = 200$. With the exception of method TW, all other methods had a larger standard deviation for $N = 200$ than for $N = 1000$. For methods TW and CIMS–E this was reversed.

Main Effects

Effect of percentage of missingness. Percentage of missingness had a main effect [$\chi^2(1) = 899.08$, $p < .001$]. Table 6 shows that cluster discrepancy was smaller for 5% missingness (last row of upper panel, last two columns) than for 15% missingness (last row of middle panel, last two columns).

Effect of imputation method. Imputation method had a main effect [$\chi^2(4) = 549.82$, $p < .001$]. Table 6 (last two columns, bottom panel) shows that the results of methods TW-E, RF, and CIMS-E differed little from those of method MNI. Of the other methods except method RI, method TW produced the largest discrepancy.

Specialized Designs

Correlation between latent variables. A 3 (correlation) × 6 (imputation method) ANOVA had discrepancy in Cronbach's alpha as a dependent variable. A similar ANOVA was done for discrepancy in coefficient H. For cluster discrepancy, a 3 (correlation) × 6 (imputation method) logistic regression with binomial counts was done. All effects of all analyses were significant.

For Cronbach's alpha, the interaction effect of imputation method and correlation was small [$F(8, 792) = 1068.25$, $p < .001$, $\eta^2 = .02$], the effect of correlation was small [$F(2, 198) = 211.01$, $p < .001$, $\eta^2 = .01$], and the effect of imputation method was large [$F(4, 396) = 6636.21$, $p < .001$, $\eta^2 = .92$]. The effect sizes showed that most variance was explained by differences between imputation methods. The large effect of imputation method was mainly caused by method TW, which produced a larger discrepancy than the other imputation methods. Because of the large contribution of method TW to effect size, we also compared the cell means (multiple t-tests using Bonferroni corrections) of the interaction of imputation method and correlation between latent variables. These tests revealed that as the correlation between latent variables increased, discrepancy decreased for methods TW, TW-E, and CIMS-E, but this decrease was small (Table 7, upper panel). For methods RF and MNI discrepancy was the same for different correlations.

For discrepancy in coefficient H (Table 7, middle panel), only the effect of imputation method was large [$F(4, 396) = 8950.37$, $p < .001$, $\eta^2 = .92$]; the other effects were not discernable. Furthermore, multiple t-tests using Bonferroni correction revealed that methods TW, TW-E, and CIMS-E produced a downward shift of discrepancy in H which was greater as the data came closer to unidimensionality (represented by a correlation of $\rho = .50$).

For cluster discrepancy, the largest effect was the main effect of correlation [$\chi^2(2) = 42.62$, $p < .001$]. As correlation increased, more items had to be moved to reobtain the original cluster solution (Table 7, bottom panel). The

TABLE 7
Mean (*M*) and Standard Deviation (*SD*) of the Discrepancy of Cronbach's Alpha, Coefficient *H*, and Cluster Solution for the Specialized Design With Different Correlations Between Latent Variables. Totals Represent Results Aggregated Across Either Imputation Method (Rows), Correlation (Columns), or Both (Lower Right Corner in Each Panel). Entries of Discrepancy in Alpha and Coefficient H Must Be Multiplied by 10^{-3}

		Correlation							
		0		*.24*		*.50*		*Total*	
Dependent Variable	*Method*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Discrepancy in alpha	RI	−23	2	−20	2	−17	2	−20	3
	TW	7	1	6	1	5	1	6	1
	TW-E	1	1	0	1	0	1	1	1
	RF	0	1	0	1	0	1	0	1
	CIMS-E	1	1	0	1	0	1	0	1
	MNI	0	1	−1	1	−1	1	−1	1
	Total*	2	3	1	3	1	2	1	3
Discrepancy in *H*	RI	−34	3	−38	4	−42	4	−38	5
	TW	13	2	13	2	13	2	13	2
	TW-E	1	2	0	2	0	2	0	2
	RF	0	2	0	2	0	2	0	2
	CIMS-E	1	2	0	2	−1	2	0	2
	MNI	−1	2	−1	2	−1	2	−1	2
	Total*	2	5	2	6	2	6	2	5
Discrepancy in cluster solution	RI	.45	.61	1.88	.73	2.80	.79	1.71	1.20
	TW	.59	.55	1.04	.95	1.53	1.16	1.05	1.00
	TW-E	.29	.56	.54	.83	1.00	1.21	.61	.95
	RF	.27	.57	.74	.93	.96	.99	.66	.90
	CIMS-E	.27	.51	.79	.98	1.04	1.29	.70	1.03
	MNI	.28	.59	.50	.86	.95	1.05	.58	.89
	Total*	.33	.57	.71	.92	1.09	1.14	.71	.96

*Aggregated across all imputation methods, except method RI.

imputation methods had a larger standard deviation of cluster discrepancy as correlation increased.

Latent-variable ratio. For the specialized design with different latent-variable ratios, a 3 (mix) × 7 (method) ANOVA was carried out, with discrepancy in Cronbach's alpha as the dependent variable. All effects were significant.

TABLE 8
Mean (*M*) and Standard Deviation (*SD*) of the Discrepancy of Cronbach's Alpha, Coefficient H, and Cluster Solution for the Specialized Design With Different Latent-Variable Ratios. Totals Represent Results Aggregated Across Either Imputation Method (Rows), Latent-Variable Ratio (Columns), or Both (Lower Right Corner in Each Panel). Entries of Discrepancy in Alpha and Coefficient H Must Be Multiplied by 10^{-3}

		Latent-Variable Ratio							
		Mix 1:0		*Mix 3:1*		*Mix 1:1*		*Total*	
Dependent Variable	*Method*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Discrepancy in alpha	RI	−32	3	−20	2	−16	2	−19	3
	TW	8	1	6	1	5	1	6	2
	TW-E	1	1	0	1	0	1	1	1
	RF	−1	1	0	1	0	1	0	1
	CIMS-E	0	1	0	1	0	1	0	1
	MNI	−1	1	−1	1	0	1	−1	1
	Total*	2	4	1	3	1	2	1	3
Discrepancy in H	RI	−39	4	−38	4	−36	3	−38	4
	TW	12	2	13	2	13	2	13	2
	TW-E	0	2	0	2	1	2	0	2
	RF	−1	2	0	2	0	2	0	2
	CIMS-E	0	2	0	2	1	2	0	2
	MNI	−2	2	−1	2	−1	2	−1	2
	Total*	1	5	2	6	2	5	2	5
Discrepancy in cluster solution	RI	3.27	1.04	1.88	.73	1.98	.82	2.38	1.08
	TW	.57	.71	1.04	.95	1.38	1.04	1.00	.97
	TW-E	.72	.98	.54	.83	1.00	1.20	.75	1.03
	RF	.82	.97	.74	.93	.90	1.02	.82	.97
	CIMS-E	.82	.95	.79	.98	1.04	1.18	.88	1.04
	MNI	.51	.69	.50	.86	.88	1.04	.63	.89
	Total*	.72	.90	.71	.92	1.02	1.10	.82	.99

*Aggregated across all imputation methods, except method RI.

The interaction effect of imputation method and latent-variable ratio was small [$F(8, 792) = 1184.15$, $p < .001$, $\eta^2 = .04$], and the main effect of imputation method was large [$F(4, 396) = 6613.77$, $p < .001$, $\eta^2 = .71$] .

For all imputation methods, discrepancy in Cronbach's alpha decreased as the data approached unidimensionality more closely (from Mix 1:0, via Mix 3:1, to Mix 1:1); this decrease was small for all methods (Table 8, upper panel). Method TW produced a larger (positive) discrepancy than the other methods

(not counting method RI). Differences in discrepancies found between imputation methods were small.

All effects on discrepancy in H were significant, but only the main effect of imputation method was discernable [$F(4, 396) = 8873.46$, $p < .001$, $\eta^2 = .89$]. Table 8 (middle panel) shows that the discrepancy in H varied little across different latent-variable ratios (not counting method RI). Method TW, which showed the largest differences in discrepancy over the three latent-variable ratios, produced discrepancies of .012 ($SD = .002$), .013 ($SD = .002$) and .013 ($SD = .002$) for Mix 1:0, Mix 3:1, and Mix 1:1, respectively.

All effects on cluster discrepancy were significant. Logistic regression yielded the following results: for the interaction of imputation method and latent-variable ratio: $\chi^2(8) = 45.29$, $p < .001$; for the main effect of imputation method: $\chi^2(4) = 44.14$, $p < .001$; and for the main effect of latent-variable ratio: $\chi^2(2) = 11.13$, $p < .001$. Table 8 (bottom panel) shows that for most methods discrepancy decreased in going from Mix 1:0 to Mix 3:1, but increased in going from Mix 3:1 to Mix 1:1. For method TW discrepancy increased as the data came closer to unidimensionality. The standard deviation of discrepancy showed an irregular pattern. Methods TW-E and RF had the smallest standard deviation for Mix 3:1, and the largest standard deviation for Mix 1:1. For methods TW, CIMS-E, and MNI the standard deviation increased as the data came closer to unidimensionality.

Number of answer categories. All effects of the ANOVAs for the specialized design with dichotomous and polytomous items were significant. For discrepancy in Cronbach's alpha, the interaction effect of imputation method and number of answer categories was medium [$F(4, 396) = 797.54$, $p < .001$, $\eta^2 = .07$], and the main effect of imputation method was large [$F(4, 396) = 3524.56$, $p < .001$, $\eta^2 = .66$]. Table 9 (upper panel) shows that method MNI produced larger means and larger standard deviations of discrepancy in Cronbach's alpha for dichotomous items than for polytomous items. For methods TW, TW-E, RF, and CIMS-E only small differences in discrepancy were found between dichotomous and polytomous items. The standard deviations of discrepancy were larger for dichotomous items than for polytomous items.

For discrepancy in coefficient H, the interaction effect of imputation method and number of answer categories was medium [$F(4, 396) = 3932.28$, $p < .001$, $\eta^2 = .11$], the main effect of imputation method was large [$F(4, 396) = 6071.88$, $p < .001$, $\eta^2 = .71$], and the main effect of number of answer categories was small [$F(1, 99) = 243.55$, $p < .001$, $\eta^2 = .05$]. The results for coefficient H (Table 9, middle panel) differed from the results for Cronbach's alpha. Discrepancy in coefficient H was smaller for dichotomous items, than for polytomous items. This was found for five imputation methods but not for method MNI: this method showed larger discrepancy for dichotomous items

TABLE 9
Mean (*M*) and Standard Deviation (*SD*) of the Discrepancy of Cronbach's Alpha, Coefficient *H*, and Cluster Solution for the Specialized Design With Different Number of Answer Categories. Totals Represent Results Aggregated Across Either Imputation Method (Rows), Number of Answer Categories (Columns), or Both (Lower Right Corner in Each Panel). Entries of Discrepancy in Alpha and Coefficient *H* Must Be Multiplied by 10^{-3}

		Number of Answer Categories					
		2		*5*		*Total*	
Dependent Variable	*Method*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Discrepancy in alpha	RI	−21	2	−17	2	−19	3
	TW	7	2	1	1	4	3
	TW-E	0	2	0	1	0	2
	RF	0	2	0	1	0	1
	CIMS-E	0	2	0	1	0	2
	MNI	−4	2	−1	1	−2	2
	Total*	0	4	0	1	0	3
Discrepancy in *H*	RI	−14	2	−36	3	−25	11
	TW	5	1	13	2	9	4
	TW-E	0	2	1	2	0	2
	RF	0	1	0	2	0	2
	CIMS-E	0	2	1	2	0	2
	MNI	−3	1	−1	2	−2	2
	Total*	0	3	2	5	1	4
Discrepancy in cluster solution	RI	2.55	1.77	1.98	.82	2.26	1.41
	TW	.20	.53	1.38	1.04	.79	1.02
	TW-E	.55	.81	1.00	1.20	.78	1.04
	RF	.15	.48	.90	1.02	.53	.88
	CIMS-E	.51	.85	1.04	1.18	.78	1.06
	MNI	.24	.49	.88	1.04	.56	.87
	Total*	.31	.65	1.02	1.10	.66	.97

*Aggregated across all imputation methods, except method RI.

than for polytomous items. Unlike Cronbach's alpha, the standard deviation of the discrepancy in coefficient H was smaller for dichotomous items than for polytomous items.

For cluster discrepancy, all effects were significant: interaction of imputation method and number of answer categories [$\chi^2(4) = 38.91$, $p < .001$]; imputation method [$\chi^2(4) = 54.07$, $p < .001$]; and number of answer categories [$\chi^2(1) = 37.22$, $p < .001$]. In general, cluster discrepancy was larger

for polytomous items than for dichotomous items, but the difference varied across methods. Table 9 (lower panel, first two columns) shows that method MNI produced a small cluster discrepancy for dichotomous items. For dichotomous items, discrepancies produced by TW and RF resembled discrepancy produced by method MNI. Methods TW-E and CIMS-E produced largest cluster discrepancy for dichotomous items (not counting method RI). However, for polytomous items (third and fourth column of lower panel), method TW produced the largest cluster discrepancy (not counting method RI), followed by method CIMS-E. Methods TW-E, RF, and MNI produced smaller cluster discrepancy for polytomous than the other methods. For method RI the standard deviation of the cluster discrepancy was larger for dichotomous items than for polytomous items. For the other imputation methods, the opposite result was found.

DISCUSSION

The aim of this study was to determine the influence of simple multiple-imputation methods on results of psychometric analyses of test and questionnaire data. The statistically more elegant and advanced multiple-imputation method MNI was included as an upper benchmark for these simpler methods.

Surprisingly, in most situations multiple-imputation method TW-E produced the smallest discrepancy, which often was even smaller than that produced by upper benchmark MNI. For MAR and MCAR with 5% missingness, the discrepancy in Cronbach's alpha and the H coefficient produced by method TW-E came close to 0. Method TW-E also produced small cluster discrepancy.

Methods CIMS-E and RF were the next best methods. Method CIMS-E produced discrepancy in Cronbach's alpha and coefficient H similar to that produced by method TW-E, but larger cluster discrepancy. Method RF produced larger discrepancy in Cronbach's alpha and coefficient H than method TW-E, but cluster discrepancy close to that of method TW-E. For dichotomous items, method RF produced the smallest cluster discrepancy of all methods.

Method MNI has been claimed to be robust against departures from multivariate normality (Graham & Schafer, 1999) but the highly discrete item-response data used here nevertheless may have led MNI to produce larger discrepancy relative to statistically simpler methods that are free of these distributional assumptions.

A noticeable result was that, although significant, missingness mechanism did not have much influence on the discrepancy measures. This may be due to the large number of variables (20 items and one covariate) included in the imputation procedures causing even NMAR mechanisms to closely approach MAR (see, e.g., Schafer, 1997, p. 28).

Finally, it may be noted that for data sets other than those obtained from typical 'multiple-items' tests and questionnaires, such as medical data containing variables like age, body mass, and total serum cholesterol, and data sets containing only total scores for various scales (but no underlying item scores), the simple methods investigated in this study cannot be used. For these kinds of data sets method MNI is recommended. For test and questionnaire data, methods TW-E, CIMS-E, and RF may be preferred, but differences relative to MNI with respect to expected discrepancy often are so small that advocates of this method can also use it for analyzing such data sets without running serious risks of obtaining distorted results.

REFERENCES

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.

Bernaards, C. A., Farmer, M. M., Qi, K., Dulai, G. S., Ganz, P. A., & Kahn, K. L. (2003). Comparison of two multiple imputation procedures in a cancer screening survey. *Journal of Data Science, 1*, 293–312.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, J. L. (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 (with discussion). *Journal of the Royal Statistical Society Series B, 39*, 1–38.

Ezzati-Rice, T. M., Johnson, W., Khare, M., Little, R. J. A., Rubin, D. B., & Schafer, J. L. (1995). A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examination surveys. *Proceedings of the Annual Research Conference* (pp. 257–266). Washington, DC: Bureau of the Census.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: Sage.

Huisman, M. (1998). *Item nonresponse: Occurrence, causes, and imputation of missing answers to test items*. Leiden, The Netherlands: DSWO Press.

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

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59*, 149–176.

King, G., Honaker, J., Joseph., A., & Scheve, K. (2001a). Analyzing incomplete political science data. *American Political Science Review, 95*, 49–69.

King, G., Honaker, J., Joseph., A., & Scheve, K. (2001b). AMELIA: A program for missing data Version 2.1. Retrieved May 29, 2006, from http://gking.harvard.edu/stats.shtml

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. *Psychological Bulletin, 45*, 507–530.

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

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

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

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

Schafer, J. L. (1998). NORM: Version 2.02 for Windows 95/98/NT. Retrieved May 29, 2006, from http://www.stat.psu.edu/∼jls/misoftwa.html

Schafer, J. L., Ezzati-Rice, T. M., Johnson, W., Khare, M., Little, R. J. A., & Rubin, D. B. (1996). The NHANES III multiple imputation project. *Proceedings of the survey research methods section of the American Statistical Association* (pp. 28–37). Retrieved May 29, 2006, from http://www.amstat.org/sections/srms/Proceedings/papers/1996_004.pdf

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.

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.

Smits, N. (2003). *Academic specialization choices and academic achievement: Prediction and incomplete data*. Unpublished doctoral dissertation, University of Amsterdam.

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.

SOLAS (2001). *SOLAS for missing data analysis 3.0* [Computer software]. Cork, Ireland: Statistical solutions.

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

SPSS Inc. (2004). SPSS 12.0.1 for Windows [Computer software]. Chicago: Author.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NJ: Erlbaum.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*, 528–540.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 397–412.

Van Abswoude, A. A. H., Van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28*, 3–24.

Van der Ark, L. A., & Sijtsma K. (2005). 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* (pp. 147–166). Mahwah, NJ: Erlbaum.

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. (2005a). TW.ZIP, RF.ZIP, and CIMS.ZIP [Computer code]. Retrieved May 29, 2006, 2005, from http://www.uvt.nl/mto/software2.html

Van Ginkel, J. R., & Van der Ark, L. A. (2005b). SPSS syntax for missing value imputation in test and questionnaire data. *Applied Psychological Measurement, 29*, 152–153.

Vermunt, J. K., & Magidson, J. (2005a). *Latent GOLD 4.0* [Computer software]. Belmont MA: Statistical Innovations.

Vermunt, J. K., & Magidson, J. (2005b). *Technical Guide for Latent GOLD: Basic and Advanced* [Software manual]. Belmont, MA: Statistical Innovations.

Yuan, Y. C. (2000). Multiple imputation for missing data: Concepts and new development. *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference* (Paper, No. 267). Cary, NC: SAS Institute. Retrieved May 29, 2006, from http://www.ats.ucla.edu/stat/sas/library/multipleimputation.pdf