Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 4013–4027
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation

Joost R. Van Ginkel*, L. Andries Van der Ark, Klaas Sijtsma, Jeroen K. Vermunt

Department of Methodology and Statistics, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Received 23 May 2006; received in revised form 11 December 2006; accepted 11 December 2006
Available online 17 December 2006

Abstract

Previous research has shown that method two-way with error for multiple imputation in test and questionnaire data produces small bias in statistical analyses. This method is based on a two-way ANOVA model of persons by items but it is improper from a Bayesian point of view. Proper two-way imputations are generated using data augmentation. Simulation results show that the resulting method two-way with data augmentation produces unbiased results in Cronbach's alpha, the mean of squares in ANOVA, the item means, and small bias in the mean test score and the factor loadings from principal components analysis. The data with imputed scores result in statistics having a slightly larger standard deviation than the original complete data. Method two-way with error produces results that are only slightly more biased, especially for low percentages of missingness. Thus, it may serve as an accurate approximation to the more involved method two-way with data augmentation.

Keywords: Effect of imputation on psychometrically important statistics; Missing item scores; Multiple imputation of item scores; Two-way imputation with data augmentation; Two-way imputation with error

1. Introduction

Tests and questionnaires are used as measurement instruments in psychological, sociological, marketing, and medical research. Data collected by means of tests and questionnaires consist of the scores of N subjects (N is relatively large, e.g., $N = 200$) on J items (J is relatively small, e.g., $J = 20$). Together the items may measure one attribute, such as introversion (psychology), religiosity (sociology), service quality (marketing), and health-related quality of life (medicine). Occasionally, subsets of items measure different attributes, such as different aspects of introversion (e.g., fear, depression, and shame). Typically, an attribute is measured by multiple items. Often, several respondents do not answer all the questions, which results in item nonresponse. Reasons may be sloppiness, tiredness, lack of motivation, or the personal nature of the questions causing people to experience feelings of irritation or uneasiness, threat, or invasion of privacy. The result is an incomplete data matrix.

Multiple imputation (Rubin, 1987) may be used for handling missing item scores by estimating the missing scores M times according to a statistical model. The resulting M different complete data sets are analyzed by means of

* Corresponding author. Tel.: +31 13 4668046; fax: +31 13 4663002.
E-mail address: j.r.vanginkel@uvt.nl (J.R. Van Ginkel).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.12.022

standard statistical procedures. Results are combined into overall estimates of the statistics. Multiple imputation corrects estimates and their standard errors for the uncertainty caused by the missing data, using rules proposed by Rubin (1987).

Some statistical techniques that rely on, for example, full information maximum likelihood or procedures using multilevel modeling as described by Maas and Snijders (2003) do not require the use of multiple imputation. The advantage of multiple imputation, however, is that it produces complete data sets that may be used for just *any* further statistical analysis. In this study, we investigate two multiple-imputation methods that both yield complete data sets.

Multiple imputation for tests and questionnaires may be done by means of statistically involved and often superior methods, or simpler methods that require little statistical knowledge of the substantive researcher. Involved methods often use data augmentation (Tanner and Wong, 1987) for estimation of the imputation model. Examples are multiple imputation under the multivariate normal model, the saturated multinomial or loglinear models, or the general location model (Schafer, 1997). Examples of simple methods are two-way imputation with normally distributed errors (TW-E; Bernaards and Sijtsma, 2000), corrected item-mean substitution (Huisman, 1998), and response-function imputation (Sijtsma and Van der Ark, 2003). Several studies (Bernaards and Sijtsma, 2000; Huisman, 1998; Sijtsma and Van der Ark, 2003; Smits et al., 2002; Van der Ark and Sijtsma, 2005; Van Ginkel et al., in press(a)) have produced evidence that these simple methods perform rather well in recovering results of factor analysis, classical test theory, and item response theory.

This study combines features of a statistically sound approach to multiple imputation with the simplicity often appreciated by substantive researchers. Method TW-E (Bernaards and Sijtsma, 2000), which is the most promising among the simple methods, produces little bias and relatively accurate standard errors in several statistical computations (e.g., Van Ginkel et al., in press (a, b)) but the method is statistically improper (Schafer, 1997, p. 105), and also has some other statistical flaws. These problems still may produce some bias in results of statistical analysis. We propose a multiple-imputation version of method TW-E that generates proper multiple imputations under a two-way ANOVA model.

This study has three goals. First, we propose a proper multiple-imputation method (TW-DA; DA stands for data augmentation). Second, we investigate how much bias of method TW-E can be attributed to its improperness and its statistical problems. Also, we study whether method TW-DA can eliminate this bias. Third, we study how much bias the methods produce in practically useful statistics. The first two goals are pursued by studying the bias produced by the methods in several two-way ANOVA-based statistics. The third goal is pursued by studying the bias in the mean test score, Cronbach's (1951) alpha, and in factor loadings.

First, method TW-E is discussed. Second, the novel proper method TW-DA is explained. Third, the results of two simulation studies on the performance of methods TW-DA and TW-E are discussed. Finally, recommendations on the practical use of both methods are given.

2. Two-way imputation

Notation. Let $\mathbf{X}$ be an N (persons) $\times J$ (items) data matrix with an observed part, $\mathbf{X}_{obs}$, and a missing part, $\mathbf{X}_{mis}$, so that $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$. The set of observed scores is denoted by *obs*, and the set of missing scores is denoted by *mis*. The total number of observed scores in the set *obs* is denoted by #*obs*; likewise, #*mis* is defined. The set of observed scores on item j is denoted by $obs(j)$, and the set of missing scores on item j is denoted by $mis(j)$; counts in these sets are denoted by $\#obs(j)$ and $\#mis(j)$, respectively. The set of observed scores of person i is $obs(i)$, and the set of his/her missing scores is $mis(i)$; counts in these sets are denoted by $\#obs(i)$ and $\#mis(i)$, respectively.

Definition of properness. Schafer (1997, p. 105) defined a multiple-imputation method to be Bayesianly proper if the imputed values are independent realizations of $P(\mathbf{X}_{mis}|\mathbf{X}_{obs})$, given some complete-data model and a prior distribution of a set of model parameters, denoted by λ. If this condition is met, the distribution of $P(\mathbf{X}_{mis}|\mathbf{X}_{obs})$ equals

$$P(\mathbf{X}_{mis}|\mathbf{X}_{obs}) = \int P(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \lambda) P(\lambda|\mathbf{X}_{obs}) \, \mathrm{d}\lambda, \tag{1}$$

and the imputed values reflect both uncertainty about $\mathbf{X}_{mis}$, given λ, and the unknown model parameters λ.

2.1. Two-way with normally distributed errors (TW-E)

Method TW-E is based on a two-way ANOVA model of persons by items (Bernaards and Sijtsma, 2000, p. 333). Define $\mu_{i\bullet}$ as the population mean of person i, $\mu_{\bullet j}$ as the mean score on item j in the population of persons, and μ as the overall mean across both $\mu_{i\bullet}$ and $\mu_{\bullet j}$. The error term is denoted by ε_{ij} with $\varepsilon \sim N(0, \sigma^2)$. The two-way ANOVA model is defined as

$$X_{ij} = \mu_{i\bullet} + \mu_{\bullet j} - \mu + \varepsilon_{ij} \quad \text{with } \varepsilon \sim N(0, \sigma^2). \tag{2}$$

Parameter $\mu_{i\bullet}$ is estimated by the mean of all observed scores for person i, denoted by PM_i, $\mu_{\bullet j}$ is estimated by the mean of all observed scores on item j, denoted by IM_j, and μ is estimated by the overall mean of all observed item scores, denoted by OM. The estimators are defined as

$$PM_i = \sum_{j \in obs(i)} X_{ij}/\#obs(i), \quad IM_j = \sum_{i \in obs(j)} X_{ij}/\#obs(j), \quad OM = \sum_{i,j \in obs} \sum X_{ij}/\#obs.$$

For a missing score in cell (i, j), we define a preliminary estimate of the item score as

$$\hat{X}_{ij} = PM_i + IM_j - OM. \tag{3}$$

The error variance σ^2 is estimated by means of

$$S^2 = \sum_{i,j \in obs} \sum (X_{ij} - \hat{X}_{ij})^2/(\#obs - 1). \tag{4}$$

Following the two-way ANOVA model, error score ε_{ij} is drawn from $N(0, S^2)$ and added to $\hat{X}_{ij}$ to obtain the final estimated item score,

$$\tilde{X}_{ij} = \hat{X}_{ij} + \varepsilon_{ij}, \tag{5}$$

and this item score $\tilde{X}_{ij}$ is imputed in cell (i, j). Before imputing, scores may (Van Ginkel et al., in press (a, b)) or may not (Bernaards and Sijtsma, 2000) be rounded to the nearest feasible integer. Both options were studied.

Besides being improper, this method has two potential problems. One pertains to the preliminary estimate of the missing score (Eq. (3)), and the other to the magnitude of the error variance (Eq. (4)). For complete data matrix $\mathbf{X}$, in a balanced design variations in the item scores due to overall main effects are additive (Winer, 1971, pp. 402–404). Then, the two-way ANOVA model may be formalized as in Eq. (3). The parameter in cell (i, j) equals $\mu_{ij} = \mu_{i\bullet} + \mu_{\bullet j} - \mu$.

Given normality of errors, the sample means $\bar{X}_{i\bullet}$, $\bar{X}_{\bullet j}$, and $\bar{X}$ are maximum likelihood estimates (MLEs) of their corresponding population means. The mean of squares (MS) of the error, $MS(E) = SS(E)/[(N-1)(J-1)]$, is an unbiased estimate of the error variance (Brennan, 2001, p. 27). Then, given that the design is balanced, the estimate $\hat{X}^*_{ij} = \bar{X}_{i\bullet} + \bar{X}_{\bullet j} - \bar{X}$ is an MLE of μ_{ij}. However, due to missing item scores the design is unbalanced and additivity of main effects is lost; thus, Eq. (3) does not provide an MLE of μ_{ij}.

Due to this problem, the numerator in Eq. (4) is biased in an unknown direction. Also, because method TW-E uses $(\#obs - 1)$ instead of $(\#obs - N - J + 1)$ the number of degrees of freedom is too large and the error variance in Eq. (4) may be biased. This bias may produce biased standard errors and confidence intervals in the results from statistical analysis. A Bayesianly proper method TW-E version could resolve such problems.

2.2. Two-way imputation with data augmentation (TW-DA)

Because it uses data augmentation, method TW-DA yields Bayesianly proper multiple imputations (Schafer, 1997, p. 106). First, we reparameterize the two-way ANOVA model (Eq. (2)) such that $\alpha_i = \mu_{i\bullet}$ and $\beta_j = \mu_{\bullet j} - \mu$; this is a necessary step for the data augmentation sampling scheme, which is explained later on. Next, we assume that α_i is a random person effect with normal distribution $N(\mu, \tau^2)$, and that β_j a fixed item effect, which is restricted such that $\sum_{j=1}^{J} \beta_j = 0$. The two-way ANOVA model can now be considered to be a random intercept model:

$$X_{ij} = \alpha_i + \beta_j + \varepsilon_{ij} \quad \text{with } \alpha \sim N(\mu, \tau^2) \text{ and } \varepsilon \sim N(0, \sigma^2). \tag{6}$$

The parameters are β_j, μ, σ^2, and τ^2. Data augmentation (Tanner and Wong, 1987) is used to obtain values of α_i, and to generate proper multiple imputations according to the two-way ANOVA model.

Following a sampling scheme proposed by Hoijtink (2000) and using noninformative priors for each parameter, the following steps are taken:

1. Starting values are assigned to μ, β_j, α_i, σ^2, and τ^2 which are denoted by $\mu^{(0)}$, $\beta_j^{(0)}$, $\alpha_i^{(0)}$, $\sigma^{2(0)}$, and $\tau^{2(0)}$, respectively. Starting values are obtained using the TW-E estimates (Eq. (5)); that is,

$$\mu^{(0)} = \frac{\sum\sum_{i,j\in obs} X_{ij} + \sum\sum_{i,j\in mis} \tilde{X}_{ij}}{NJ},$$

$$\beta_j^{(0)} = \frac{\sum_{i\in obs(j)} X_{ij} + \sum_{i\in mis(j)} \tilde{X}_{ij}}{N} - \mu^{(0)},$$

$$\alpha_i^{(0)} = \frac{\sum_{j\in obs(i)} X_{ij} + \sum_{j\in mis(i)} \tilde{X}_{ij}}{J},$$

$$\sigma^{2(0)} = \frac{\sum\sum_{i,j\in obs} (X_{ij} - \alpha_i^{(0)} - \beta_j^{(0)})^2 + \sum\sum_{i,j\in mis} [\tilde{X}_{ij} - \alpha_i^{(0)} - \beta_j^{(0)}]^2}{(N-1)(J-1)},$$

$$\tau^{2(0)} = \sum_{i=1}^{N} [\alpha_i^{(0)} - \mu^{(0)}]^2/(N-1).$$

   Note that the random error component of $\tilde{X}_{ij}$ (Eq. (5)) produces different starting values for each chain.
2. At iteration t, person effect α_i is sampled from a normal posterior distribution, conditional on the other current parameter estimates. Specifically, $\alpha_i^{(t)}|\beta_j^{(t-1)}, \mu^{(t-1)}, \sigma^{2(t-1)}, \tau^{2(t-1)}, \mathbf{X}_{obs}$ is sampled with mean

$$\frac{\mu^{(t-1)}/\tau^{2(t-1)} + \sum_{j\in obs(i)} [X_{ij} - \beta_j^{(t-1)}]/\sigma^{2(t-1)}}{1/\tau^{2(t-1)} + \#obs(i)/\sigma^{2(t-1)}}$$

   and variance

$$\frac{1}{1/\tau^{2(t-1)} + \#obs(i)/\sigma^{2(t-1)}}.$$

3. At iteration t, item effect β_j is sampled from a normal posterior distribution, given the other current parameter draws. Specifically, $\beta_j^{(t)}|\alpha_1^{(t)}, \ldots, \alpha_N^{(t)}, \sigma^{2(t-1)}, \mathbf{X}_{obs}$ is sampled with mean $\sum_{i\in obs(j)}[X_{ij} - \alpha_i^t]/\#obs(j)$ and variance $\sigma^{2(t-1)}/\#obs(j)$.
4. At iteration t, the error variance is sampled from a posterior distribution, conditional on the current parameter draws. That is, $\sigma^{2(t)}|\alpha_1^{(t)}, \ldots, \alpha_N^{(t)}, \beta_1^{(t)}, \ldots, \beta_J^{(t)}, \mathbf{X}_{obs}$ is sampled from a scaled-inverse chi-square distribution with degrees of freedom (v) equal to $\#obs$ and scale (S^2) equal to $\sum\sum_{i,j\in obs} [X_{ij} - \alpha_i^{(t)} - \beta_j^{(t)}]^2/\#obs$. This is achieved by drawing a random variable from a chi-square distribution with v degrees of freedom, and letting $\sigma^2 = vS^2/\chi_v^2$ (Gelman et al., 2003, p. 580).
5. The overall mean μ is sampled from a normal posterior distribution, conditional on the other current parameter estimates. Specifically, $\mu^{(t)}|\alpha_1^{(t)}, \ldots, \alpha_N^{(t)}, \tau^{2(t-1)}, \mathbf{X}_{obs}$ is drawn with mean $\sum_{i=1}^{N}\alpha_i^{(t)}/N$ and variance $\tau^{2(t-1)}/N$.
6. The variance of the person effect is sampled from a posterior distribution, conditional on the current parameter draws. Specifically, $\tau^{2(t)}|\alpha_1^{(t)}, \ldots, \alpha_N^{(t)}, \mu^{(t)}, \mathbf{X}_{obs}$ is sampled from its posterior distribution, which is a scaled-inverse chi-square distribution with N degrees of freedom and scale $\sum_{i=1}^{N}[\alpha_i^{(t)} - \mu^{(t)}]^2/N$.
7. Steps 2–6 are repeated $2T$ times. The first T iterations of this chain are used as burn in, and the last T for assessment of the convergence of the algorithm.
8. Steps 2–7 are repeated M times, creating M chains used for generating M multiple imputations and checking convergence. For checking convergence a measure for multiple chains is used (Gelman et al., 2003, p. 461). Let $\phi_m^{(t)}$

be a parameter within chain m at iteration t, $\bar{\phi}_m$ the mean parameter of chain m, and $\bar{\phi}$ the mean parameter across all chains and iterations. Further, let S_m^2 be the variance of parameter $\phi_m^{(t)}$ within chain m. The within-chains variance is computed as $W = \sum_{m=1}^{M} S_m^2/M$, the between-chains variance as $B = T\sum_{m=1}^{M}(\bar{\phi}_m - \bar{\phi})^2/(M-1)$, and the total variance as $V = (1 - T^{-1})W + T^{-1}B$. The convergence criterion is defined as $\sqrt{R} = \sqrt{V/W}$. As the variances between chains decreases, $\sqrt{R}$ approaches 1, and convergence is more plausible. After doing some preliminary simulations, we found that $\sqrt{R} \leqslant 1.001$ produced good results for all parameters.

9. For each chain, a completed data set is created by simulating draws of the missing data according to the two-way model, using the parameter draws from the last iteration. Thus, $X_{ij,mis}|\alpha_i^{(2T)}, \beta_j^{(2T)}, \sigma^{2(2T)}, \tau^{2(2T)}, \mathbf{X}_{obs}$ is drawn from a normal distribution with mean $\alpha_i^{(2T)} + \beta_j^{(2T)}$ and variance $\sigma^{2(2T)}$. The resulting imputed values are proper because each chain has different random values of $\alpha_i^{(2T)}$,$\beta_j^{(2T)}$, and $\sigma^{2(2T)}$, which is equivalent to integrating out the parameters, as in Eq. (1).

In data augmentation it is common to let the imputation of the missing data be part of the sampling steps at each iteration t (e.g., Schafer, 1997, p. 72). This is useful when the sampling of the unknown model parameters is simpler for complete than for incomplete data, such as in a multivariate normal model with an unrestricted covariance matrix. However, this does not apply to the two-way ANOVA model; thus, we do not impute missing values in $\mathbf{X}_{mis}$ during the estimation of the ANOVA model but only the values of the random effects α_i. Imputation of the missing values in $\mathbf{X}_{mis}$ was done after the last iteration of the sampling scheme.

3. Two simulation studies

Simulation Study 1 was done to find out how much bias of method TW-E could be attributed to its improperness and its statistical problems, and to study whether method TW-DA could eliminate this bias. Data were generated under the two-way ANOVA model and the multidimensional polytomous latent trait (MPLT) model (Kelderman and Rijkes, 1994). The two-way ANOVA model is the basis of both methods but does not describe test and questionnaire data well. The MPLT model gives a more accurate description of such data (e.g., Van der Linden and Hambleton, 1997) but is the wrong model for both methods. For data generated under the ANOVA model, bias produced by method TW-E in parameter estimates of the ANOVA model must be due to its improperness and its statistical problems whereas the proper method TW-DA is expected to produce unbiased estimates. For data generated using the MPLT model, ANOVA parameter estimates may be biased for both the original data and the completed data. Due to its improperness and its statistical problems, method TW-E is expected to produce bias that deviates from the bias of the original data, whereas method TW-DA is expected to produce bias of similar magnitude as the bias of the original data.

Simulation Study 2 studied the influence of methods TW-E and TW-DA on practically useful statistics in realistic data sets. Only the MPLT model was used because this is a realistic model for test and questionnaire data (e.g., Van der Linden and Hambleton, 1997). Next, the simulation models and dependent variables are discussed in more detail.

The two-way ANOVA used was a random intercept model of random persons and fixed items (Eq. (6)). The MPLT model version used is a constrained version of the original MPLT model and expresses the probability of giving a response $X_{ij} = x$ to item j, given person i's values on two latent variables (this choice was based on Bernaards and Sijtsma, 2000; Van Ginkel et al., in press (a)). The latent variables are denoted by $\theta_g (g = 1, 2)$, ψ_{jx} is the separation parameter of item j for answer category x, and B_{jg} $(B_{jg} \geqslant 0)$ is the discrimination parameter of item j with respect to latent variable g. The constrained MPLT model is defined as

$$P(X_{ij} = x|\theta_{i1}, \theta_{i2}) = \frac{\exp[\sum_{g=1}^{2}(\theta_{ig} - \psi_{jx})B_{jg}]}{\sum_{y=0}^{x}\{\exp[\sum_{g=1}^{2}(\theta_{ig} - \psi_{jy})B_{jg}]\}}. \tag{7}$$

The item parameters with respect to $x = 0$ are set to 0 to ensure uniqueness of the parameters.

The study was programmed in Borland Delphi 6.0 (2001). The MPLT model was used to generate an artificial population of 1,000,000 simulees based on the following choices. The latent traits were drawn from a bivariate standard normal distribution with correlation $\rho = 0.24$ (based on Van Ginkel et al., in press (a)). The tests contained 20 items with five ordered answer categories. Items 1–10 were driven by θ_1, and items 11–20 by θ_2. The item parameters are in Table 1 (based on Van Ginkel et al., in press (a)).

Table 1
Location parameters ψ_{jx} of item j and answer category x, discrimination parameters B_{jg} of item j and latent variable θ_g, and item mean $\mu_{\bullet j}$ of item j in the artificial population

Items	ψ_{j1}	ψ_{j2}	ψ_{j3}	ψ_{j4}	B_{j1}	B_{j2}	$\mu_{\bullet j}$
1	−2.75	−2.25	−1.75	−1.25	0.5	0	2.72
2	−2.75	−2.25	−1.75	−1.25	2	0	3.05
3	−1.75	−1.25	−0.75	−0.25	0.5	0	2.15
4	−1.75	−1.25	−0.75	−0.25	2	0	2.22
5	−0.75	−0.25	0.25	0.75	0.5	0	1.55
6	−0.75	−0.25	0.25	0.75	2	0	1.34
7	0.25	0.75	1.25	1.75	0.5	0	1.03
8	0.25	0.75	1.25	1.75	2	0	0.63
9	1.25	1.75	2.25	2.75	0.5	0	0.64
10	1.25	1.75	2.25	2.75	2	0	0.22
11	1.25	1.75	2.25	2.75	0	0.5	0.05
12	1.25	1.75	2.25	2.75	0	2	0.39
13	0.25	0.75	1.25	1.75	0	0.5	0.22
14	0.25	0.75	1.25	1.75	0	2	0.64
15	−0.75	−0.25	0.25	0.75	0	0.5	0.63
16	−0.75	−0.25	0.25	0.75	0	2	1.03
17	−1.75	−1.25	−0.75	−0.25	0	0.5	1.34
18	−1.75	−1.25	−0.75	−0.25	0	2	1.55
19	−2.75	−2.25	−1.75	−1.25	0	0.5	2.22
20	−2.75	−2.25	−1.75	−1.25	0	2	2.15

Items have a discrimination parameter for one latent trait; thus, the MPLT model can be conceptualized as two separate generalized partial credit models (Muraki, 1992) with correlated latent variables.

We computed the item means $\mu_{\bullet j}$ (Table 1, last column), the variance of the person means ($\tau^2 = 0.21$), the error variance (σ^2=0.75), and Cronbach's alpha (α=0.81). The values for $\mu_{\bullet j}$, τ^2, and σ^2 were also used to define a population for simulating data sets under a two-way ANOVA model. Under this model, Cronbach's alpha was computed by means of $\alpha = \tau^2/[(\tau^2 + \sigma^2)/J]$ (e.g., McGraw and Wong, 1996, p. 36); this resulted in $\alpha = 0.85$. Notice that under the ANOVA model items are parallel; thus, Cronbach's alpha equals the test-score reliability.

3.1. Simulation Study 1: studying the effect of the problems of method TW-E

3.1.1. Fixed factors

Within each design cell, 10,000 ($D = 10,000$) samples ($N = 200$) were drawn, with replacement after each computation round. Twenty items were used ($J = 20$), each with ordered scores 0, 1, 2, 3, 4.

3.1.2. Independent variables

Simulation model: The two-way ANOVA model and the MPLT model.

Percentage of missingness: In each of the samples, 5%, 10%, or 20% of the item scores were randomly removed. The number of completed data matrices in multiple imputation depends to a great extent on the fraction of missing information (Schafer, 1997, pp. 106–107). Thus, for higher percentages of missingness, a larger number of completed data matrices may be needed. The number of completed data sets used here was proportional to the percentage of missingness, yielding $M = 5$, 10, and 20 completed data matrices.

Missingness mechanism: Missingness was ignorable (Little and Rubin, 2002, pp. 199–120): missing completely at random (MCAR) or missing at random (MAR). For MCAR, scores were drawn at random with equal probability from the data and removed. For MAR, missingness depended on one completely observed item: for subjects with $X_{ij} > 2$, the probability of scores on the other items being missing was twice as high as for subjects with $X_{ij} \leqslant 2$ (item 4 was chosen because $P(X_{i4} > 2) \approx P(X_{i4} \leqslant 2)$). These probabilities were used to remove a random sample of cells from the data.

Imputation methods: Methods TW-E and TW-DA were used. Imputed scores were not rounded because this study was based on a theoretical two-way ANOVA framework, which assumes continuous data.

3.1.3. Dependent variables

For several ANOVA-related statistics, the bias, the standard deviation (denoted by *SD*), and the coverage percentage were studied. Let $\hat{Q}_{or}$, d estimate parameter Q in original data set d ($d = 1, \ldots, D$), and let $\hat{Q}_{imp,dm}$ estimate Q in the mth completed (denoted by *imp*) data set ($m = 1, \ldots, M$) based on sample d. For the original data, the bias in $\hat{Q}_{or}$ is computed as

$$b(\hat{Q}_{or}) = \frac{\sum_{d=1}^{D}(\hat{Q}_{or}, d - Q)}{D},$$

and for the completed data, bias is computed as

$$b(\hat{Q}_{imp}) = \frac{\sum_{d=1}^{D}[\sum_{m=1}^{M}(\hat{Q}_{imp,dm})/M - Q]}{D}.$$

As a measure of efficiency, the *SD* of $\hat{Q}_{or}$ and $\hat{Q}_{imp}$ was used. The coverage percentage of estimate $\hat{Q}_{or}$ is the percentage of coverage intervals based on the original data that include the true Q. The coverage percentage of $\hat{Q}_{imp}$ is the percentage of coverage intervals based on the completed data that include the true Q. The standard error (SE) of $\hat{Q}_{imp,dm}$ is adjusted for extra uncertainty caused by the missing data (Rubin, 1987).

Bias, *SD*, and coverage percentage were computed for (1) the mean of item 1, denoted by $\bar{X}_{\bullet 1}$. Results were expected to be the same for the other item means. For the coverage percentage of $\bar{X}_{\bullet 1,imp,dm}$, the *SE* was adjusted using a correction of the degrees of freedom (Barnard and Rubin, 1999); and for (2) Cronbach's alpha, because it is used in almost every study that uses test and questionnaire data. Moreover, Kristof (1963) derived the sampling distribution of Cronbach's alpha under the assumptions of the two-way ANOVA model. The 95% confidence intervals of Cronbach's alpha were obtained by transformation of the alpha value of each data set to a Fisher z score; for more details see McGraw and Wong (1996, p. 46) Only bias and *SD* were studied for (1) the mean of squares of the person effect, denoted by $MS(A)$ and (2) the mean of squares of the error, denoted by $MS(E)$.

Because of the large number of replications, the bias, the standard deviation, and the coverage percentage of the original data and the completed data were compared by means of inspection of the differences without statistical testing.

3.2. Simulation Study 2: influence of imputation methods on practical statistics

3.2.1. Fixed factors

Data were generated using the MPLT model (Eq. (7). Sample size was $N = 200$. The number of items was $J = 20$. Each item had ordered scores 0, 1, 2, 3, 4.

3.2.2. Independent variables

Percentage of missingness: The percentages of missingness were 5%, 10%, and 20%, and the corresponding numbers of completed data sets were $M = 5$, 10, and 20.

Missingness mechanism: Missingness mechanisms were MCAR and MAR.

Imputation methods: Methods TW-E and TW-DA were used. Because researchers prefer to have complete data sets with imputed integer scores that can be used for any statistical analysis and because of the practical context of this study, imputed scores were rounded to the nearest integer in the 0–4 interval.

3.2.3. Dependent variables

Test-score distribution: The test score of person i is defined as $X_{i+} = \sum_{j=1}^{J} X_{ij}$. The test score estimates a psychological property of interest, such as posttraumatic stress disorder. A test score may be computed across both unidimensional and multidimensional item sets. The latter possibility applies to this study. A practical example is a questionnaire that consists of different subscales, each of which measures a different symptom of posttraumatic stress disorder (e.g., Simms et al., 2005). Then, the test score is a summary of different posttraumatic stress symptoms. Let $J\mu$ be the population mean of the test score; for our constructed population, $J\mu = 25.79$. Bias, *SD*, and coverage percentage of

Table 2
Population factor loadings of the artificial population

Item	Factor 1	Factor 2	Item	Factor 1	Factor 2
1	0.51	0.04	11	0.03	0.47
2	0.74	0.09	12	0.03	0.38
3	0.55	0.05	13	0.06	0.67
4	0.80	0.09	14	0.04	0.46
5	0.55	0.04	15	0.08	0.77
6	0.81	0.09	16	0.04	0.51
7	0.51	0.04	17	0.10	0.80
8	0.75	0.08	18	0.05	0.54
9	0.44	0.03	19	0.11	0.77
10	0.62	0.06	20	0.06	0.53

the mean test score were studied. To study the coverage percentage of mean test score $\bar{X}_+$ based on the completed data, the *SE* was corrected using an adjusted number of degrees of freedom (Barnard and Rubin, 1999).

Cronbach's alpha: Bias, *SD*, and coverage percentage of Cronbach's alpha were studied. Unlike Simulation Study 1, this study computed Cronbach's alpha for completed data sets with rounded imputed scores.

Factor loadings from PCA and Varimax rotation: The bias in the factor loadings of the completed data was computed as follows. First, the correlation matrices of M completed data sets were added and then each element was divided by M so as to obtain one overall correlation matrix. A principal components analysis (PCA) followed by Varimax rotation was done on this correlation matrix. Suppose $\hat{a}_{jk,imp,d}$ is the estimated factor loading of item j on factor k, based on data set d. The bias in the factor loadings of the completed data was computed as

$$b(\hat{a}_{jk,imp}) = \frac{\sum_{d=1}^{D}(\hat{a}_{jk,imp,d} - a_{jk})}{D}.$$

The computation of the *SD* was straightforward. Coverage percentages were not determined. The population factor loadings are given in Table 2.

4. Results: Simulation study 1

4.1. Results for the item mean

Bias: Table 3 shows the bias in $\bar{X}_{\bullet 1}$ for the original data, and for all combinations of simulation model, imputation method, missingness mechanism, and percentage of missingness (first column). The largest bias in $\bar{X}_{\bullet 1}$ (equal to -0.022) was found for data simulated under the MPLT model, for method TW-E, 20% missingness, and missingness mechanism MAR. Thus, in this worst case a population item mean of $\mu_{\bullet 1} = 2.720$ on average was underestimated as 2.698.

For method TW-E, negative bias in $\bar{X}_{\bullet 1}$ increased as percentage of missingness increased. For MAR, this bias increase was larger for the MPLT model than for the two-way ANOVA model. Method TW-DA produced unbiased results for $\bar{X}_{\bullet 1}$ in almost all situations. For the MPLT model and missingness mechanism MAR, method TW-DA produced a small negative bias in $\bar{X}_{\bullet 1}$ which increased as the percentage of missingness increased. This increase was smaller than that for method TW-E under the same circumstances.

Standard deviations: Table 3 (second column) shows that the *SD* of $\bar{X}_{\bullet 1}$ increased for methods TW-E and TW-DA as percentage of missingness increased. This increase of *SD* is due to the increased uncertainty caused by the missing data. In general, the increase was small. For example, for the two-way ANOVA model $\bar{X}_{\bullet 1}$ of the original data had $SD = 0.070$ but for MAR and 20% missingness method TW-DA produced $\bar{X}_{\bullet 1}$'s with $SD = 0.077$. This means that when $\mu_{\bullet 1} = 2.720$, for the original data the 95% coverage interval ranged from 2.580 to 2.860, whereas for 20% missingness, MAR, and method TW-DA, the interval ranged from 2.566 to 2.874.

Table 3
Bias in $\bar{X}_{\bullet 1}$, standard deviation of $\bar{X}_{\bullet 1}$, and coverage percentage of $\bar{X}_{\bullet 1}$, for methods TW-E and TW-DA, compared to the results of the original data

Simulation model	Data sets	Missingness mechanism	Percentage missingness	Bias	*SD*	Coverage percentage
Two-way	Original			0	70	94.7
ANOVA model	TW-E	MCAR	5	−3	72	94.7
			10	−7	73	94.8
			20	−15	77	94.5
		MAR	5	−3	72	94.8
			10	−6	73	95.3
			20	−12	76	94.8
	TW-DA	MCAR	5	0	72	94.8
			10	0	73	94.8
			20	0	77	94.9
		MAR	5	0	72	94.9
			10	0	73	95.0
			20	0	77	95.1
MPLT	Original			1	81	95.0
model	TW-E	MCAR	5	−3	84	94.5
			10	−7	86	94.2
			20	−15	90	93.5
		MAR	5	−5	84	94.8
			10	−11	86	94.2
			20	−22	91	93.3
	TW-DA	MCAR	5	0	84	94.4
			10	0	86	94.2
			20	0	91	93.7
		MAR	5	−2	84	94.7
			10	−5	86	94.0
			20	−11	91	93.4

Entries for bias and *SD* must be multiplied by 10^{-3}.

Coverage percentage: For the two-way ANOVA model, the percentage of $\mu_{\bullet 1}$'s covered by the coverage intervals was close to 95% for both imputation methods (Table 3, last column, upper half). For both methods, the coverage percentage was nearly constant for different missingness mechanisms and percentages of missingness. For the MPLT model, the coverage percentage showed a different pattern (Table 3, lower half). For 5% missingness, the coverage percentage was close to 95%, but the coverage percentage was smaller for both imputation methods as percentage of missingness increased.

4.2. Results for Cronbach's alpha

Bias: For the two-way ANOVA model (Table 4, first column, upper half), bias in Cronbach's alpha was zero or nearly zero. For example, for the two-way ANOVA model, 20% missingness, and MAR, method TW-E produced a positive bias of 0.008. Thus, a population alpha of 0.85 is on average overestimated as 0.858. For the MPLT model, in general bias in Cronbach's alpha was somewhat larger. For example, for 20% missingness and MAR, method TW-E produced a bias of 0.018. Thus, a population Cronbach's alpha of 0.850 is on average estimated as 0.868.

Method TW-E produced small positive bias in Cronbach's alpha, which increased as percentage of missingness increased. Compared to the two-way ANOVA model, for the MPLT model, bias was larger and a little increased faster. Method TW-DA produced almost unbiased results in almost all situations.

Standard deviation: Table 4 (second column) shows that for method TW-E the *SD* of Cronbach's alpha decreased a little as percentage of missingness increased. This decrease is unexpected because more missingness causes more uncertainty in an estimate. This result was only found for Cronbach's alpha. For method TW-DA, the *SD* increased as expected as percentage of missingness increased.

Table 4
Bias in Cronbach's alpha, standard deviation of Cronbach's alpha, and coverage percentage of Cronbach's alpha, for methods TW-E and TW-DA, compared to the results of the original data

Simulation model	Data sets	Missingness mechanism	Percentage missingness	Bias	*SD*	Coverage percentage
Two-way	Original			−2	16	95.5
ANOVA model	TW-E	MCAR	5	1	15	96.0
			10	3	15	95.6
			20	8	14	93.3
		MAR	5	1	15	96.0
			10	3	15	95.5
			20	8	14	93.2
	TW-DA	MCAR	5	−2	16	95.5
			10	−2	16	95.4
			20	−2	17	95.5
		MAR	5	−2	16	95.6
			10	−2	16	95.5
			20	−2	17	95.4
MPLT model	Original			−2	20	95.2
	TW-E	MCAR	5	2	19	95.6
			10	5	19	95.1
			20	13	18	91.1
		MAR	5	3	19	95.4
			10	7	19	93.9
			20	18	17	85.2
	TW-DA	MCAR	5	−2	20	95.1
			10	−2	21	95.1
			20	−3	21	95.0
		MAR	5	−2	20	95.3
			10	−1	20	95.3
			20	1	21	95.0

Entries for bias and *SD* must be multiplied by 10^{-3}.

Coverage percentage: For method TW-E, for 5% and 10% missingness the percentage of intervals that included the true Cronbach's alpha was close to 95%. This percentage was smaller as the percentage of missingness increased to 20 (Table 4, last column). The worst decrease was for 20% missingness and MAR, when the true alpha was covered by only 85.2% of the intervals. For method TW-DA, the percentage of intervals that covered the true alpha was close to 95%, and was constant as percentage of missingness increased.

4.3. Results for the mean of squares of the person effect

Bias: Table 5 (first column) shows that for the two-way ANOVA model method TW-DA produced almost unbiased results in $MS(A)$ for all missingness mechanisms and all percentages of missingness. Also, the small bias in $MS(A)$ was always equal to the bias in $MS(A)$ in the original data. Method TW-E produced relatively large positive bias in $MS(A)$ for data under the two-way ANOVA model. This bias increased as percentage of missingness increased. Part of the bias may be attributed to the random sampling of persons in the two-way ANOVA model, whereas method TW-E treats the persons as fixed.

For the MPLT model, both methods TW-E and TW-DA produced large negative bias in $MS(A)$. For MCAR, bias produced by method TW-DA was almost equal to the bias in the original data; bias remained constant as percentage of missingness increased. Method TW-E produced bias in $MS(A)$ that differed from bias in the original data. Bias was smaller as percentage of missingness increased.

Standard deviation: Table 5 shows that the *SD* of $MS(A)$ increased a little both for method TW-E and method TW-DA as percentage of missingness increased.

Table 5
Bias in $MS(A)$ and standard deviation of $MS(A)$ for methods TW-E and TW-DA, compared to the results of the original data

Simulation model	Data sets	Missingness mechanism	Percentage missingness	Bias	*SD*
Two-way	Original			−3	481
ANOVA model	TW-E	MCAR	5	90	488
			10	165	492
			20	370	503
		MAR	5	89	489
			10	166	494
			20	378	506
	TW-DA	MCAR	5	−3	487
			10	−2	492
			20	−3	502
		MAR	5	−3	488
			10	−3	494
			20	−5	503
MPLT model	Original			−759	410
	TW-E	MCAR	5	−676	415
			10	−585	420
			20	−369	433
		MAR	5	−656	417
			10	−540	424
			20	−253	443
	TW-DA	MCAR	5	−761	415
			10	−761	421
			20	−761	433
		MAR	5	−747	418
			10	−735	423
			20	−705	438

Entries for bias and *SD* must be multiplied by 10^{-3}.

4.4. Results for the mean of squares of the error

The results for the $MS(E)$ (Table 6) were comparable to the results for the $MS(A)$ (Table 5) but the absolute numbers are much different. Most important is that under the ANOVA model method TW-DA produced unbiased $MS(E)$ and method TW-E nearly unbiased $MS(E)$ (Table 6). Under the MPLT model, both methods produced the similar bias.

5. Results: Simulation study 2

5.1. Results for the mean test score

Bias: Table 7 (first column, upper half) shows for both methods TW-E and TW-DA that the positive bias in $\bar{X}_+$ increased as percentage of missingness increased. Method TW-E produced more bias in $\bar{X}_+$ than method TW-DA. However, bias in $\bar{X}_+$ was small. The largest bias in $\bar{X}_+$ (method TW-E, 20% missingness, MAR) was 0.431.

Standard deviation: Table 7 (second column, upper half) shows that the *SD* of $\bar{X}_+$ increased for both methods TW-E and TW-DA as percentage of missingness increased. Methods TW-E and TW-DA showed similar results with respect to *SD*.

Coverage percentage: For both methods TW-E and TW-DA, Table 7 (last column, upper half) shows that the coverage percentage of $\bar{X}_+$ was smaller as percentage of missingness increased. This effect was equal for both methods.

Table 6
Bias in $MS(E)$ and standard deviation of $MS(E)$ for methods TW-E and TW-DA, compared to the results of the original data

Simulation model	Data sets	Missingness mechanism	Percentage missingness	Bias	*SD*
Two-way ANOVA model	Original			0	17
	TW-E	MCAR	5	1	18
			10	2	19
			20	7	20
		MAR	5	1	18
			10	2	18
			20	7	20
	TW-DA	MCAR	5	0	18
			10	0	18
			20	0	20
		MAR	5	0	18
			10	0	18
			20	0	19
MPLT model	Original			40	22
	TW-E	MCAR	5	40	22
			10	41	23
			20	47	24
		MAR	5	40	22
			10	40	23
			20	44	24
	TW-DA	MCAR	5	40	22
			10	40	23
			20	40	24
		MAR	5	40	22
			10	38	23
			20	37	24

Entries for bias and *SD* must be multiplied by 10^{-3}.

5.2. Results for Cronbach's alpha

Bias: Table 7 (first column, lower half) shows that for method TW-E the positive bias in Cronbach's alpha increased as percentage of missingness increased. For method TW-DA, the negative bias in Cronbach's alpha increased as percentage of missingness increased. Rounding the imputed scores has the effect of inducing almost no extra bias in Cronbach's alpha (compare Table 4, lower half, with the results in Table 7, lower half).

Standard deviation: Table 7 (second column, lower half) shows that under all conditions both methods TW-E and TW-DA produce an *SD* in Cronbach's alpha of approximately 0.02.

Coverage percentage: The results with respect to coverage percentage of Cronbach's alpha (Table 7, last column, lower half) are difficult to interpret. Under MCAR, method TW-E had coverage percentages that are larger than 95% but remained nearly stable as percentage of missingness increased. Under MAR, method TW-E had a smaller coverage percentage as percentage of missingness increased. For method TW-DA, an opposite result was found: under MAR, method TW-DA had a coverage percentage that was stable as percentage of missingness increased; and under MCAR, the coverage percentage was smaller as percentage of missingness increased. In general, both methods produced coverage percentages that were rather closer to the theoretical 95% coverage interval.

5.3. Results for factor loadings

Bias: For both factors 1 and 2, similar results were found for bias in loadings; thus, only bias results for the first factor are discussed. Methods TW-E and TW-DA produced relatively large bias in the factor loadings (Table 8). The high loadings of items 1–10 are biased downwards and the low loadings of items 11–20 are biased upwards. This result is probably due to the use of a unidimensional model for imputing scores in two-dimensional test data, which biases the

Table 7
Bias in $\bar{X}_+$ and Cronbach's alpha, and standard deviation and coverage percentage of $\bar{X}_+$ and Cronbach's alpha, for methods TW-E and TW-DA, compared to the results of the original data

Statistic	Data sets	Missingness mechanism	Percentage missingness	Bias	*SD*	Coverage percentage
$\bar{X}_+$	Original			4	646	94.9
	TW-E	MCAR	5	94	647	94.9
			10	186	650	94.1
			20	375	653	91.8
		MAR	5	101	647	95.0
			10	204	646	94.6
			20	431	649	91.2
	TW-DA	MCAR	5	91	647	94.6
			10	179	650	93.8
			20	355	653	90.9
		MAR	5	79	647	95.0
			10	155	646	94.4
			20	307	650	92.2
Cronbach's	Original			−2	20	95.2
Alpha	TW-E	MCAR	5	1	20	95.6
			10	1	19	95.9
			20	5	19	95.5
		MAR	5	1	19	95.6
			10	4	19	95.1
			20	13	18	91.4
	TW-DA	MCAR	5	−4	20	94.9
			10	−6	21	94.5
			20	−10	22	93.5
		MAR	5	−3	20	95.2
			10	−3	20	95.2
			20	−4	21	95.1

Entries for bias and *SD* must be multiplied by 10^{-3}.

data towards unidimensionality. Method TW-E produced the smallest bias in the loadings of items 1–10, and method TW-DA produced the smallest bias in the loadings of items 11–20. Thus, the performance of these methods seems to be similar.

Even though bias is large, conclusions based on imputed data may not differ dramatically from those based on the original data. The largest bias found was 0.16 (item 11, method TW-E, 20% missingness, MAR); thus, the population loading ($a_{11,1} = 0.03$) on average is overestimated to be 0.19. Rules of thumb claim that loadings below 0.32 should not be interpreted (Comrey and Lee, 1992). Thus, this bias seems to have little consequence.

Standard deviations: The *SD* of the loadings is nearly stable across imputation methods, missingness mechanisms, and percentages of missingness (Table 8).

6. Discussion

Two multiple-imputation methods were compared that both use a two-way ANOVA model: the Bayesianly proper method TW-DA, and the simpler, statistically suboptimal but practically attractive method TW-E. Simulation Study 1 studied the degree to which bias produced by method TW-E could be attributed to this method's improperness and statistical problems, and whether method TW-DA could eliminate this bias. The influence of both methods on ANOVA statistics was studied. Method TW-DA produced unbiased results under the two-way ANOVA model. Moreover, for data simulated under the two-way ANOVA model it was insensitive to different missingness mechanisms and increased percentages of missingness, and for data simulated under the MPLT model it was almost insensitive to these factors. Method TW-E always produced biased results. Bias was not stable as percentage of missingness increased: for the two-way ANOVA model, bias often increased as percentage of missingness increased, and for the MPLT model, as

Table 8
Mean bias in factor loadings from PCA, *SD*s in parentheses, for methods TW-DA and TW-E, compared to the results of the original data

Bias in: Or. data	Method																							
	TW-E									TW-DA														
	Missingness mechanism																							
	MCAR				MAR						MCAR						MAR							
	Percentage missingness																							
	5		10		20		5		10		20		5		10		20		5		10		20	
$a_{1,1}$	−2	(61)	−5	(61)	−7	(60)	−11	(60)	−6	(61)	−9	(61)	−15	(61)	−7	(61)	−12	(61)	−21	(61)	−8	(61)	−13	(61)	−24	(61)
$a_{2,1}$	−2	(29)	−18	(31)	−35	(32)	−67	(34)	−20	(33)	−37	(35)	−71	(35)	−19	(31)	−37	(32)	−72	(34)	−20	(31)	−38	(33)	−73	(35)
$a_{3,1}$	−3	(58)	−7	(59)	−10	(58)	−16	(57)	−8	(58)	−12	(58)	−21	(58)	−9	(59)	−14	(59)	−25	(58)	−10	(59)	−17	(58)	−30	(59)
$a_{4,1}$	−2	(25)	−21	(27)	−40	(28)	−77	(30)	−5	(25)	−8	(26)	−15	(26)	−22	(27)	−42	(28)	−81	(31)	−5	(25)	−8	(25)	−15	(25)
$a_{5,1}$	−3	(60)	−7	(60)	−11	(59)	−17	(59)	−8	(59)	−13	(59)	−21	(59)	−9	(60)	−15	(60)	−26	(60)	−10	(60)	−18	(60)	−32	(60)
$a_{6,1}$	−2	(25)	−22	(27)	−41	(28)	−80	(31)	−23	(29)	−44	(31)	−84	(31)	−23	(27)	−43	(28)	−84	(32)	−24	(27)	−46	(29)	−90	(32)
$a_{7,1}$	−4	(66)	−6	(66)	−9	(66)	−13	(64)	−7	(65)	−10	(64)	−15	(64)	−9	(67)	−14	(66)	−23	(65)	−10	(66)	−16	(66)	−28	(65)
$a_{8,1}$	−2	(33)	−22	(35)	−42	(36)	−80	(38)	−24	(35)	−44	(37)	−84	(37)	−23	(35)	−44	(36)	−84	(39)	−25	(35)	−47	(36)	−90	(38)
$a_{9,1}$	−4	(76)	−4	(75)	−5	(74)	−6	(72)	−3	(73)	−2	(70)	0	(70)	−7	(76)	−10	(75)	−16	(73)	−6	(75)	−9	(74)	−15	(72)
$a_{10,1}$	−3	(51)	−27	(54)	−48	(55)	−85	(56)	−26	(52)	−47	(51)	−79	(51)	−27	(54)	−49	(55)	−89	(57)	−27	(53)	−49	(53)	−87	(53)
$a_{11,1}$	0	(84)	28	(81)	51	(77)	91	(74)	51	(78)	92	(74)	160	(74)	21	(81)	38	(76)	66	(71)	41	(81)	74	(77)	126	(72)
$a_{12,1}$	2	(87)	19	(87)	35	(86)	67	(84)	25	(86)	48	(84)	95	(84)	15	(87)	27	(86)	51	(85)	19	(87)	37	(86)	71	(84)
$a_{13,1}$	1	(71)	13	(71)	26	(70)	50	(69)	23	(71)	45	(70)	89	(70)	9	(71)	18	(70)	34	(68)	18	(72)	33	(71)	64	(69)
$a_{14,1}$	1	(83)	14	(83)	26	(83)	53	(81)	16	(82)	33	(81)	68	(81)	11	(83)	20	(82)	40	(81)	12	(83)	24	(82)	49	(81)
$a_{15,1}$	0	(62)	8	(62)	15	(62)	32	(62)	8	(63)	17	(62)	39	(62)	5	(62)	10	(62)	20	(62)	5	(62)	10	(63)	24	(63)
$a_{16,1}$	0	(80)	11	(80)	22	(79)	46	(79)	11	(79)	22	(79)	49	(79)	9	(80)	17	(80)	35	(79)	8	(80)	16	(79)	35	(78)
$a_{17,1}$	0	(59)	5	(60)	11	(60)	26	(60)	2	(59)	6	(59)	19	(59)	4	(60)	7	(60)	16	(60)	0	(60)	2	(60)	8	(60)
$a_{18,1}$	1	(77)	10	(77)	19	(77)	41	(77)	8	(76)	16	(75)	37	(75)	8	(77)	15	(77)	31	(77)	6	(77)	12	(76)	26	(75)
$a_{19,1}$	0	(61)	5	(62)	11	(62)	25	(62)	0	(61)	2	(60)	12	(60)	3	(62)	7	(62)	15	(62)	−1	(61)	0	(62)	5	(61)
$a_{20,1}$	0	(78)	9	(78)	18	(78)	40	(78)	5	(77)	12	(76)	30	(76)	7	(78)	14	(78)	30	(78)	4	(78)	9	(77)	22	(76)

Entries must be multiplied by 10^{-3}.

percentage of missingness increased bias deviated more from bias found in results from the original data. To summarize, we found that the problems of method TW-E produced only small bias and that method TW-DA successfully eliminated the bias resulting from the statistical problems of method TW-E.

Simulation Study 2 investigated the influence of methods TW-E and TW-DA on practically useful statistics in realistic data sets. Differences between method TW-E and TW-DA were small and sometimes unclear. Method TW-DA performed better with respect to $\bar{X}_+$ than method TW-E, but equally well with respect to Cronbach's alpha. The differences between these methods were less obvious than in Simulation Study 1. Also, both methods showed similar performance in recovering the factor loadings.

Other noteworthy results are the following. Cronbach's alpha was estimated with little bias both when the imputed scores were not rounded (Simulation Study 1) and when they were rounded (Simulation Study 2). This limited result suggests that rounding only has little effect on bias results.

Despite large bias in estimated factor loadings, this bias did not have consequences for the final item clustering based on rules of thumb (Comrey and Lee, 1992). Similar results have been used (Van Ginkel et al., in press (b)) to adapt method TW-E to be applicable to multidimensional data, and similar adaptations may be pursued for method TW-DA. From Study 2 it can be concluded that for practical purposes both methods perform equally well.

Researchers may obtain proper multiple imputations by means of method TW-DA (programming code for method TW-DA is available on request). Researchers in substantive areas such as psychology, sociology, marketing, and quality-of-life research, who are unfamiliar with advanced Bayesian statistics, may safely use method TW-E, especially for percentages of missingness no larger than 5%. Moreover, this method is available as an SPSS macro (Van Ginkel and Van der Ark, 2005). Method TW-E offers a simple and often accurate approximation to method TW-DA, and produces only slightly more biased results.

References

Barnard, J., Rubin, D.B., 1999. Small-sample degrees of freedom with multiple imputation. Biometrika 86, 949–955.

Bernaards, C.A., Sijtsma, K., 2000. Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. Multivariate Behav. Res. 35, 321–364.

Borland Delphi 6.0, 2001. Computer software. Borland Software Corporation, Scotts Valley, CA.

Brennan, R.L., 2001. Generalizability Theory. Springer, New York.

Comrey, A.L., Lee, H.B., 1992. A First Course in Factor Analysis. second ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Cronbach, J.L., 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis. second ed. Chapman & Hall, London.

Hoijtink, H., 2000. Posterior inferences in the random intercept model based on samples obtained with Markov chain Monte Carlo methods. Comput. Statist. 3, 315–336.

Huisman, M., 1998. Item Nonresponse: Occurrence, Causes, and Imputation of Missing Answers to Test Items. DSWO Press, Leiden, The Netherlands.

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

Kristof, W., 1963. The statistical theory of stepped-up reliability when a test has been divided into several equivalent parts. Psychometrika 28, 221–238.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data. second ed. Wiley, New York.

Maas, C.J.M., Snijders, T.A.B., 2003. The multilevel approach to repeated measures for complete and incomplete data. Qual. Quant. 37, 71–89.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psych. Methods 1, 30–46.

Muraki, E., 1992. A generalized partial credit model: application of an EM algorithm. Appl. Psych. Meas. 16, 159–176.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

Sijtsma, K., Van der Ark, L.A., 2003. Investigation and treatment of missing item scores in test and questionnaire data. Multivariate Behav. Res. 38, 505–528.

Simms, L.J., Casillas, A., Clark, L.A., Watson, D., Doebbeling, B.N., 2005. Psychometric evaluation of the restructured clinical scales of the MMPI-2. Psych. Assessment 17, 345–358.

Smits, N., Mellenbergh, G.J., Vorst, H.C.M., 2002. Alternative missing data techniques to grade point average: imputing unavailable grades. J. Ed. Meas. 39, 187–206.

Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. J. Am. Statist. Assoc. 82, 528–540.

Van der Ark, L.A., Sijtsma, K., 2005. The effect of missing data imputation on Mokken scale analysis. In: Van der Ark, L.A., Croon, M.A., Sijtsma, K. (Eds.), New Developments in Categorical Data Analysis for the Social and Behavioural Sciences. Erlbaum, Mahwah, NJ, pp. 147–166.

Van der Linden, W.J., Hambleton, R.K. (Eds.), 1997. Handbook of Modern Item Response Theory. Springer, New York.

Van Ginkel, J.R., Van der Ark, L.A., 2005. SPSS syntax for missing value imputation in test and questionnaire data. Appl. Psych. Meas. 29, 152–153.

Van Ginkel, J.R., Van der Ark, L.A., Sijtsma, K., a. Multiple imputation of test and questionnaire data and influence on psychometric results. Multivariate Behav. Res, in press.

Van Ginkel, J.R., Van der Ark, L.A., Sijtsma, K., b. Multiple imputation of item scores when test data are factorially complex. British J. Math. Statist. Psych, in press.

Winer, B.J., 1971. Statistical Principles in Experimental Designs. second ed. McGraw-Hill, New York.