Original Article

Item Imputation Without Specifying Scale Structure

Stef van Buuren

TNO Quality of Life, Leiden, The Netherlands
University of Utrecht, The Netherlands

Abstract. Imputation of incomplete questionnaire items should preserve the structure among items and the correlations between scales. This paper explores the use of fully conditional specification (FCS) to impute missing data in questionnaire items. FCS is particularly attractive for items because it does not require (1) a specification of the number of factors or classes, (2) a specification of which item belongs to which scale, and (3) assumptions about conditional independence among items. Imputation models can be specified using standard features of the R package MICE 1.16. A limited simulation shows that MICE outperforms two-way imputation with respect to Cronbach's α and the correlations between scales. We conclude that FCS is a promising alternative for imputing incomplete questionnaire items.

Keywords: missing data, MICE, fully conditional specification, two-way imputation, item response theory, MISTRESS

Questionnaires are widely used to measure attitudes, abilities, traits, and behaviors. The occurrence of missing values on one or more of the items presents important methodological and statistical challenges. An obvious consequence of missing data is that we will be unable to calculate the total score. Reasons why missing data occur are varied: the respondent skipped the item, the item was only posed to a subgroup, the respondent failed to reach the item, respondents that fill in different forms are analyzed together, and so on.

Several approaches for dealing with item nonresponse exist. A simple fix is to fill in the person mean, fill in the item mean, or create an additional category for the missing responses. Huisman (1998) presents an overview of these methods. Such ad hoc methods make strong and often unrealistic assumptions about the missing data mechanism. Moreover, they could alter the internal scale structure and the correlation between scales. As these are all single imputation methods, they induce a systematic underestimate of the sampling variation. In general, ad hoc fixes do more harm than good (Little & Rubin, 2002).

A better approach is multiple imputation (Rubin, 1987). An early technique for multiple imputation of items is MISTRESS, which represents persons and items in the same multidimensional space (Van Buuren & Van Rijckevorsel, 1992). Van Ginkel, Van der Ark, and Sijtsma (2007) recently compared six methods for multiple imputation of items: unconditional random imputation, two-way imputation (Bernaards & Sijtsma, 2000), two-way imputation with a random error (Bernaards & Sijtsma, 2000), corrected item-mean imputation with random error (Huisman, 1998), response-function imputation (Sijtsma & Van der Ark, 2003), and rounded multivariate normal imputation (Schafer, 1997). They showed that the two-way imputation was best in the sense that it most closely reproduced Cronbach's α, Loevinger's H, and the item clustering. Van Ginkel, Van der Ark, Sijtsma, and Vermunt (2007) proposed a Bayesianly proper version of two-way imputation. That model was found to be slightly superior, but they conclude that the simpler two-way imputation "offers a simple and often accurate approximation" to the Bayesian method (p. 4026).

This paper suggests an alternative method for imputing items. The idea is to impute the data on a variable-by-variable basis, where the specification of each conditional density is under the control of the user. Such methods are known as fully conditional specification (FCS) (Van Buuren, 2007; Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). Starting from simple random values, imputation proceeds by iterating over the conditionally specified models. FCS has been implemented in R (R Development Core Team, 2008) and S-Plus (Insightful Corporation, 2007) as the multivariate imputation by chained equation (MICE) package (Van Buuren & Oudshoorn, 2000), in SPSS V17.0 (SPSS, 2008), as well as in other environments (cf. Horton & Kleinman, 2007). FCS allows tremendous flexibility in creating multivariate models. It is easy to incorporate specialized imputation methods that preserve unique features in the data, for example, bounds, skip patterns, interactions, bracketed responses, and so on. It is relatively easy to maintain constraints between different variables.

Several unique features make FCS attractive for imputing items. First, FCS does not require a specification of the number of factors or classes. Second, FCS does not require the user to specify which item belongs to which scale. Third, FCS does not assume conditional independence among items within scales or classes. All three features are

DOI: 10.1027/1614-2241/a000004

advantageous because they free the imputer from making assumptions that could ruin the complete-data analysis. In general, the imputer should make fewer assumptions than the analyst (Meng, 1995; Schafer, 1997, p. 140). FCS can create imputations of items without specifying scale structure.

Method

Notation

Suppose that l items are collected in $Y = (Y_1, \ldots, Y_l)$, a vector of l random variables with l-variate distribution $P(Y|\theta)$, where θ is an unknown parameter vector. Data are discrete with k_j response categories per item ($j = 1, \ldots, l$). Let $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_l)$ indicate the set of all $l - 1$ items except Y_j. The matrix $y = (y_1, \ldots, y_n)$ with $y_i = (y_{i1}, y_{i2}, \ldots, y_{ik})$, $i = 1, \ldots, n$ is an i.i.d. sample of Y. Let $y_{-j} = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_l)$ indicate the data of the $l - 1$ items except y_j. We denote y_{obs} to be the observed data and y_{mis} to be the missing data, so that the matrix $y = (y_{\text{obs}}, y_{\text{mis}})$ is partially observed. The random variable R_j is the response indicator of Y_j, with $R_j = 1$ if Y_j is observed, and $R_j = 0$ if Y_j is missing. Let $R = (R_1, \ldots, R_l)$ stand for the response profile. Furthermore, Y_{mis} indicates the parts of Y that are missing, that is, the parts for which $R_{ij} = 0$, while Y_{obs} is the complement with $R_{ij} = 1$ so that $Y = (Y_{\text{obs}}, Y_{\text{mis}})$.

Imputation Model

Without loss of generality, let us assume that all items contain missing data, and that the missing data pattern is general. Rubin (1987, chap. 5) defined three tasks to generate imputations: the *modeling* task, the *estimation* task, and the *imputation* task. The modeling task consists of specifying the full joint distribution $P(Y_{\text{obs}}, Y_{\text{mis}}, R|\theta)$ of the hypothetically complete data. The estimating task is to derive the posterior distribution of $P(\theta|Y_{\text{obs}}, Y_{\text{mis}}, R)$ so that random draws can be made from it. The imputation task consists of drawing imputations y^* from the posterior predictive distribution $P(Y_{\text{mis}}|Y_{\text{obs}}, R, \theta)$ to replace the missing data y_{mis}.

In this paper, we assume that the data are missing at random (MAR) and that the nonresponse mechanism is ignorable. The MAR assumption stipulates that the probability of nonresponse is unrelated to the hypothetically true (but unobserved) value of the missing data after the observed data have been taken into account. If the MAR assumption is not valid, then the data are missing not at random (MNAR). Missing completely at random (MCAR) is the special case of MAR where the probability of nonresponse is entirely random. If the data are MAR and if the parameters of the complete-data model and the process that causes the missingness are a priori independent, then the nonresponse mechanism is said to be ignorable. We refer to Rubin (1976, 1987) for a precise account of these concepts and their relations. The MCAR assumption is very restrictive. The MAR assumption is plausible in many practical applications, especially if covariates are available that are informative about the missing data process. In some cases however, for example, in tests for speediness, the MAR assumption needs careful consideration (Goegebeur, De Boeck, & Molenberghs, 2010). Under MAR, the joint distribution to be specified simplifies to $P(Y_{\text{obs}}|\theta)$. The estimation task becomes equivalent to finding $P(\theta|Y_{\text{obs}})$, and the posterior predictive distribution from which we draw imputation reduces to $P(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$.

Several routes are available to specify a joint model $P(Y_{\text{obs}}|\theta)$ for categorical responses. MISTRESS specifies category probabilities as inversely proportional to object-category distances as calculated by nonlinear principal components (Van Buuren & Van Rijckevorsel, 1992). Schafer (1997) developed a joint model for categorical data based on the log-linear model.

Instead of specifying a joint multivariate model for $P(Y_{\text{obs}}|\theta)$, FCS requires only a specification of the l conditional distributions $P(Y_j|Y_{-j}, \theta_{-j})$. For any item j, we may write $P(Y|\theta) = P(Y_j, Y_{-j}|\theta) = P(Y_j|Y_{-j}, \theta_j)P(Y_{-j}|\theta_{-j})$, where θ_j and θ_{-j} are functions of θ. The parameters θ_j are particular to $P(Y_j|Y_{-j}, \theta_j)$ and are chosen so that they model the distribution of Y_j conditional on Y_{-j}. The parameters θ_{-j} serve to define the joint distribution of the remaining $l - 1$ items Y_{-j}. Both θ_j and θ_{-j} are generally of limited scientific interest. Rubin (1987, p. 161) shows that if Y_{-j} is complete and if θ_j and θ_{-j} are modeled as a priori independent, then no specification is needed for $P(Y_{-j}|\theta_{-j})$. Thus in that case, specifying l models $P(Y_j|Y_{-j}, \theta_j)$ would be sufficient to define $P(Y|\theta)$. However because $Y_{-j} = (Y_{-j}^{\text{obs}}, Y_{-j}^{\text{mis}})$ is incomplete, the result may not apply generally. A remedy is to temporarily replace the incomplete data Y_{-j}^{mis} by imputed values y_{-j}^*. This will render Y_{-j} complete, allowing us to draw imputations y_j^* for Y_j from $P(Y_j|Y_{-j}^{\text{obs}}, Y_{-j}^{\text{mis}} = y_{-j}^*, \theta_j = \theta_j^*)$. The process is then applied to item $j + 1$ and so on. This suggests an iterative algorithm for multivariate imputation by repeated sampling from the chosen conditional distributions. Starting imputations can be drawn from the observed marginal distributions. The performance of this "chained equations" algorithm was found to be quite good under a variety of missing data scenarios (Van Buuren et al., 2006).

If the jth item is binary ($k_j = 2$), incomplete entries are imputed by logistic regression. This involves several steps. First, fit the logistic regression model $\ln(P(Y_j = 1)/P(Y_j = 0)) = \alpha + Y_{-j}\beta$ to the cases with observed Y_j, that is, probability $P(Y_j = 1)$ is predicted from the responses on the other items. Next, take a random draw (α^*, β^*) from $P(\alpha, \beta|Y_j, Y_{-j})$ under the standard noninformative prior and calculate the predicted probability $w_i = \exp(\alpha^* + y_{-ij}\beta^*)/(1 + \exp(\alpha^* + y_{-ij}\beta^*))$ for each incomplete case $i = 1, \ldots, n_j^{\text{mis}}$, where n_j^{mis} is the number of missing observations in Y_j and where y_{-ij} are the values on the "other items."

Draw $u_i \sim \text{unif}(0, 1)$. If $u_i > w_i$, impute $y_{ij}^* = 0$, otherwise impute $y_{ij}^* = 1$. For polytomous items with $k_j > 2$, we fit the multinomial logistic regression model

$$\ln \frac{P(Y_j = c)}{P(Y_j = 0)} = \alpha_c + Y_{-j}\beta_c \quad \text{for } c = 1, \ldots, k_j - 1$$

and perform the analogous draws for α_c^*, β_c^*, and y_{ij}^*. Details can be found in Appendix A of Van Buuren et al. (2006). In MICE V1.16, these methods are the default imputation functions for binary and polytomous data: `mice.impute.logreg` and `mice.impute.polyreg` (Van Buuren & Oudshoorn, 2000).

In general, it is highly recommended to include important covariates like age and sex into the imputation model. MICE implements the steps as m parallel chains that create m imputed data sets. After each pass through the data, we can plot out any function of the imputed data to study convergence of the chains. In general, convergence is fast, say 10–20 iterations, somewhat depending on the pattern and the amount of the missing data.

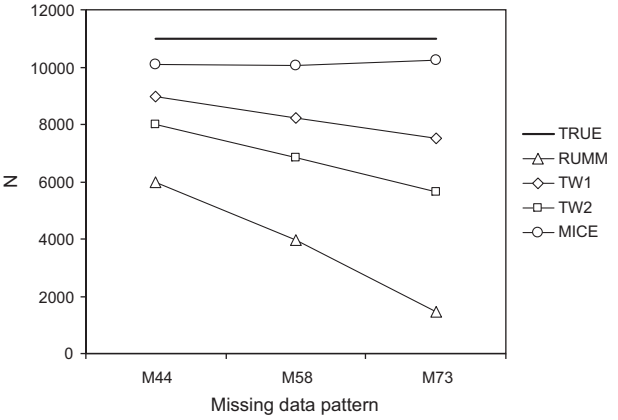

Figure 2. Number of records available for estimating item parameters by the Rasch model under three missing data patterns (binary data) after imputation by MICE, TW1, and TW2.

Simulation Setup

We study the behavior of MICE for imputing items by a limited simulation. Two complete-data sets were generated, one set of binary items and one set of items with five response categories. Both data sets contain 11,000 persons that respond to 10 items. The items belong to two different scales (A and B), each comprising of five items. The item locations were set to $\{-2, -1, 0, 1, 2\}$. The two latent traits have a bivariate normal distribution with a correlation of .24.

We defined three nonresponse patterns to create missing values. Figure 1 provides a graphic illustration of each pattern. The high percentages of missing data (44%, 58%, and 73%, in the remainder referred to as M44, M58, and M73) are not unrealistic in studies where data are combined from different sources, and aid in studying the asymptotic behavior of imputation methods. Note that listwise deletion would result in an empty data set. Six incomplete-data sets (three binary and three categorical) were constructed by randomly deleting data according to each pattern. This mechanism assumes MCAR.

We compared three methods for imputing missing data. The method of primary interest is imputation by chained equations using MICE. The method does not use any knowledge of the structural relationships between items and scales, and just imputes each incomplete item conditional on the values – observed or imputed – on the other nine items. There is no need to specify the dimensionality of the data or the number of latent classes. For comparison, we applied two versions of two-way imputation. Version 1 (TW1) uses the incorrect assumption that all 10 items belong to the same scale. Version 2 (TW2) assumes the correct item-scale structure by applying TW1 separately to the items of scale A and scale B. Imputations for TW1 and TW2 were created using the SPSS implementation of Van Ginkel and Van der Ark (2005).

The outcomes of interest are: the number of records used by each method (quantified by the number of useable cases),

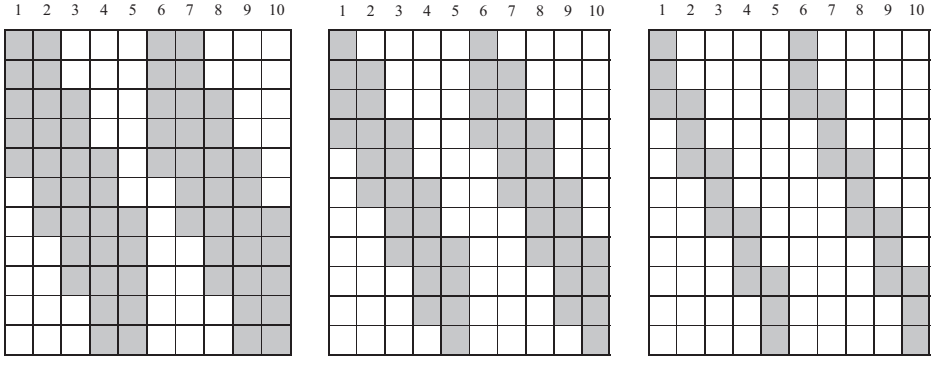

Figure 1. Three missing data patterns for 11,000 respondents and 10 items. Each row represents 1,000 persons. Observed data are dark, missing data are white. The amounts of missing data are equal to 44%, 58%, and 73% of the total data.

the reliability of scale A (quantified by Cronbach's α calculated from the items in scale A), and the structure between both scales (quantified by the correlation between the sum scores of scales A and B).

Results

Figure 2 plots the number of cases used by RUMM 2020 (Andrich, Sheridan, & Luo, 2003) for patterns M44, M58, and M73 when fitting a binary Rasch model to the items of scale A. The total number of cases is equal to 11,000. Note that the estimation after TW1 and TW2 uses far fewer cases than after using MICE. The explanation is that TW1 and TW2 impute many perfect cases, which are discarded by RUMM as "extreme." For completeness, Figure 2 also plots the useable sample size for the incomplete data ("RUMM").

We used MICE 1.16 to draw five multiple imputations per missing value after 30 iterations. Figure 3 plots the course of the correlation between scale A and scale B during the 30 iterations for the case of 73% missing binary data. There is no trend in this plot, and the variability between the different streams is quite stable. We also created similar plots of other parameters (Cronbach's α and proportion correct) and found similar patterns. From these, we concluded that 30 iterations were enough for convergence to the appropriate posterior distribution.

Figure 4a plots Cronbach's α of scale A for the imputed data. The true α is equal to .44. The result from MICE is slightly lower in all patterns. Both TW1 and TW2 however severely overestimate Cronbach's α. For TW1 this is understandable, because application TW1 incorrectly assumes that all 10 items form a scale. It is curious that TW2, which uses the correct model, is more biased than TW1. Figure 4b plots the correlations between both sum scores. The correlation is equal to .10 in the complete data. Imputation by MICE produces estimates around .10. Both TW1 and TW2 lead to inflated interscale correlations.

Figure 5 is equivalent to Figure 4, but now each item has five categories. Though Cronbach's α and the correlation are now higher, the behavior of the methods is similar to the binary case. For MICE, the direction of the bias is toward zero, whereas for TW1 and TW2 the direction of the bias is toward one. MICE is conservative, two-way imputation is optimistic.

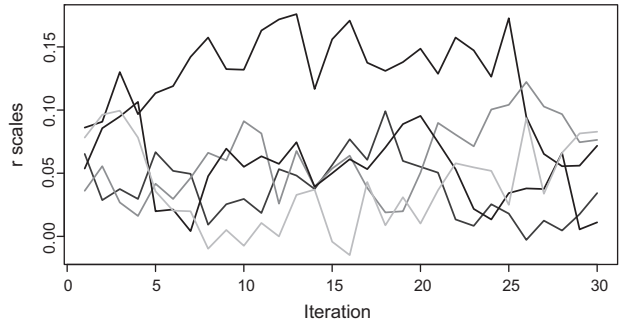

Figure 3. Convergence monitoring plot. Behavior of interscale correlation in the imputed data for five imputations during 30 iterations (73% missing data and binary data).

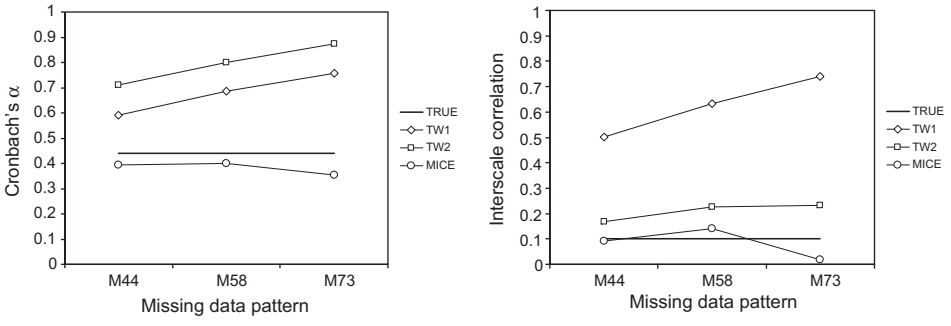

Figure 4. Binary data: Cronbach's α for scale A and the correlation between the total scores of scale A and scale B after imputation by MICE, TW1, and TW2.

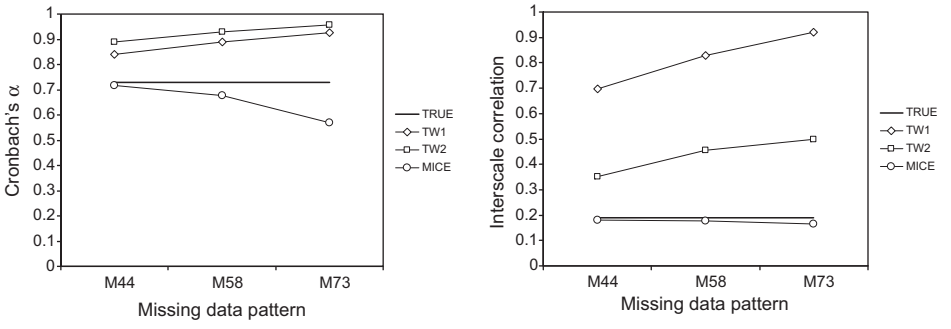

Figure 5. Polytomous data: Cronbach's α for scale A and the correlation between the total scores of scale A and scale B after imputation by MICE, TW1, and TW2.

Conclusion

This paper demonstrates that a simple regression-based algorithm using MICE outperforms two-way imputation on several criteria. Moreover, MICE was able do so without specifying any scale structure. FCS, as implemented in MICE, is a promising option for imputing incomplete questionnaire items.

The suggested modeling strategy bypasses the need to specify a joint model $P(Y_{obs}|\theta)$. On the other hand, some careful thinking is needed. Imputation is a separate modeling activity that comes with its own goals and rules. Imputation should result in valid inferences for statistical estimates from incomplete data. In order to achieve this, imputations should (1) preserve the structure in the data, (2) preserve the uncertainty about this structure, and (3) include any knowledge about the process that caused the missing data. The simulations performed in this paper only addressed point 1.

One might argue that the amounts of missing data used in the simulation are too high to be realistic. If in real life half of the data would be missing during a single administration of the test, then there might be something terribly wrong with it, and we might wish to refrain from doing anything with the data. On the other hand, using a high proportion of missing data in simulations allows us to see whether a method is systematically biased as uncertainty grows. The differences observed between TW1, TW2, and MICE, and other missing data methods will be less dramatic at lower levels of missing data, and at some point these differences will be immaterial. However, we should preferably be using methods with appropriate asymptotic properties, even for low percentages of missing data.

Note that we created missing data under the simple MCAR assumption. MICE can perform well under MAR, where the missingness probability depends on observed data. We expect therefore that imputing items by MICE will also be valid under the more general and often more realistic MAR assumption. More care is needed for data that are MNAR. In general, it is beneficial to include as many predictors as possible of the nonresponse probability, thus making the problem more MAR like. It is not known how the method will behave under severe MNAR mechanisms. More work is needed to establish the robustness of the approach against violations of the MAR assumptions.

A special feature of FCS is that it treats all items on equal footing. For a given item, the category probabilities used for imputation depend on the other items through a regression model with main effects. If we know in advance that the items form a truly unidimensional scale we can formulate the imputation model in a more parsimonious way, for example, as a one-factor model. Doing so would limit the class of models that we may reasonably apply to the imputed data. If the imputation model is unidimensional while the items are multidimensional, the interscale correlations in the imputed data get inflated. This was observed by Song and Belin (2004) who investigated factor models and found that models with too few factors may cause bias. One could try to circumvent this by partitioning the set of multidimensional items into several unidimensional subsets and create imputations separately within each set. Note that this may bias the interscale correlation downward. Using all items on equal footing evades these issues.

If the number of items is large, it is beneficial to constrain the number of predictors used in the regressions, for example, by selecting the 15 most predictive items (Van Buuren, Boshuizen, & Knook, 1999). An alternative strategy is to replace sets of items by their sum score. This requires a priori knowledge about the scale structure, but will substantially reduce the number of items needed for imputation. Items within the same set are entered as items, whereas items belonging to different sets are entered through their item set sum score. The objective is to preserve both within- and between-scale relations. It is straightforward to implement such models in MICE using so-called passive imputation. More work is needed to establish the validity of this approach. Finally, some experience with the practical application of this method to real life cases would be desirable.

References

Andrich, D., Sheridan, B. S., & Luo, G. (2003). *RUMM2020: Rasch unidimensional models for measurement*. Perth: RUMM Laboratory.

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.

Goegebeur, Y., De Boeck, P., & Molenberghs, G. (2010). Person fit for test speededness: Normal curvatures, likelihood ratio tests and empirical Bayes estimates. *Methodology, 6*, 3–16.

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *American Statistician, 61*(1), 79–90.

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

Insightful Corporation. (2007). *S-Plus 8 user's guide*. Seattle, WA: Insightful Corporation.

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

Meng, X. L. (1995). Multiple imputation with uncongenial sources of input (with discussion). *Statistical Science, 10*, 538–573.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.

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

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

Song, J., & Belin, T. R. (2004). Imputation for incomplete high-dimensional multivariate normal data using a common factor model. *Statistics in Medicine, 23*, 2827–2843.

SPSS Statistics, Rel. 17.0.0. (2008). Chicago: SPSS.

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.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*(3), 219–242.

Van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine, 18*(6), 681–694.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049–1064.

Van Buuren, S., & Oudshoorn, C. G. M. (2000). *Multivariate imputation by chained equations: MICE V1.0 user's manual*. Leiden: TNO Preventie en Gezondheid (TNO report PG/VGZ/00.038). Available from www.stefvanbuuren.nl on December 29, 2007.

Van Buuren, S., & Van Rijckevorsel, J. L. A. (1992). Imputation of missing categorical data by maximizing internal consistency. *Psychometrika, 57*(4), 567–580.

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

Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research, 42*(2), 387–414.

Van Ginkel, J. R., Van der Ark, L. A., Sijtsma, K., & Vermunt, J. K. (2007). Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Computational Statistics & Data Analysis, 51*, 4013–4027.

Stef van Buuren

TNO Quality of Life
P.O. Box 2215
2301 CE Leiden
The Netherlands
E-mail stef.vanbuuren@tno.nl

