New Developments in Missing Data Analysis

L. Andries van der Ark & Jeroen K. Vermunt

In this special issue you will find four papers on handling missing data. All papers have been presented at the 2007 Fall Meeting of Social Science Division of the Dutch Statistical Society (VVS-OR) in Tilburg, The Netherlands. Together, these four papers give an excellent overview of state of the art in missing data analysis.

To date, in virtually all fields of the social sciences, researchers are required to deal sophistically with missing data. Ignoring the problem, for example, by simply removing all observations that contain missing data or thoughtlessly applying software that makes the problem go away may lead to seriously biased statistical results and wrong conclusions, and is no longer an option. Instead the researcher must consider the reasons why some of the data are missing and act accordingly. Given that in the social sciences most data are obtained from respondents who responded to tests, questionnaires, surveys, or stimuli in an experimental setting, the first option that comes to mind is approaching those respondents with missing scores again, ask them the reason for their nonresponse, and ask them to respond yet. Unfortunately, this is usually not a realistic option and the researcher must rely on statistical solutions.

One way of dealing with missing data is to incorporate the mechanism that caused the missingness into the statistical modelling of the data. In the context of educational measurement, Goegebeur, De Boeck, and Molenberghs discuss *test speededness*, which refers to the phenomenon that respondents do not respond to certain items in the test or examination due to a lack of time. They clearly explain how speededness can be incorporated into the statistical model. Using this model-based approach, they show how to identify respondents whose scores were affected by speededness. Advantage of this approach is that it allows the researcher to deal with data that are not missing at random.

In some situations, it will not be possible to translate the researcher's theories on the missingness mechanism into a statistical model because such theories are too complex or not available. Probably the best known strategy to deal with missing data is to assume that the missing scores are missing at random and conduct (multiple) *imputation*: Replacing the missing scores in the data by plausible values. Two papers discuss imputation methods. First, Van Ginkel, Sijtsma, Van der Ark, and Vermunt investigated the occurrence of missing data and current practices of handling nonresponse in test and questionnaire data in personality psychology. They found that in the large majority of published research reporting missing data, either the handling of missing data was not discussed, cases with missing values were deleted, or ad hoc procedures were used. In order to improve the use of appropriate methods they proposed using Method Two Way for handling missing data in test and questionnaire data. Method Two Way is a multiple imputation that easy to understand and to use. Simulation studies showed that, with respect to statistics often used in the analysis of test and questionnaire data, Method Two Way yields results comparable to the results obtained with technically more advances methods.

In the second paper on multiple imputation, Van Buuren discusses Fully Conditional Specification to impute scores for missing values. Fully Conditional Specification can be regarded as a technically more advanced method, which is available in software packages

such as R and SPSS. In a simulation study Van Buuren shows that Fully Conditional Specification outperforms Method Two-Way in the computation of Cronbach's alpha. Because the papers by Van Ginkel et al. and Van Buuren reach different conclusions with respect to Method Two-Way, we believe some editorial comments are in order to explain the different results.

We believe that both papers are of high quality but have a different focus. First, the percentages of missing data differ in the study by Van Buuren and the studies referred to by Van Ginkel et al. On the one hand, Van Buuren compared Method Two-Way and Fully Conditional Specification using large percentages of missingness (44% - 78%), showing a superior performance of the technically more advanced method over the simple method, under extreme circumstances. On the other hand, Van Ginkel et al. showed that in practice the percentage of missingness is much lower (on average 9% of the response vectors had at least one missing observation), and referred to studies in which the percentage of missingness ranged from 1% to 20% missingness, showing a similar performance of simple and involved methods under typical circumstances. Moreover, with high percentages of missingness a more sophisticated Bayesian version of Method Two-Way (Van Ginkel, Van der Ark, Sijtsma, & Vermunt, 2007) may be used, which is unlikely to break down in such cases.

In the fourth paper, VanSteelandt, Carpenter, and Kenward discuss *inverse probability weighting* methods and *double robust estimation* methods, which may be plausible alternatives to multiple imputation. The literature on these topics has been quite technical, but we believe the introductory paper provided by VanSteelandt et al. make the methods accessible to researchers in the social sciences. They describe the methods illustrated by small examples and compare their pros and cons with the pros and cons of multiple imputation.

References

Goegebeur,Y., De Boeck, P, Molenberghs, G. (2009). Person fit for test speededness: Normal curvatures, likelihood ratio tests and empirical Bayes estimates. *Methodology, x*, xx-xx

Van Buuren, S. (2009). Item imputation without specifying scale structure. *Methodology, x*, xx-xx

Van Ginkel, J. R., Sijtsma, K., Van der Ark, L. A., & Vermunt, J. K. (2009). Incidence of missing item scores in personality measurement, and simple item-score imputation. *Methodology, x*, xx-xx

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.

Vansteelandt, S., Carpenter, J., & Kenward, M. G. (2009). Analysis of incomplete data using inverse probability weighting and doubly robust Estimators. *Methodology, x*, xx-xx