Incidence of Missing Item Scores in Personality Measurement, and Simple Item-Score Imputation

Joost R. van Ginkel, Leiden University
Klaas Sijtsma, Tilburg University
L. Andries van der Ark, Tilburg University
Jeroen K. Vermunt, Tilburg University

January 14, 2009

First Author's address:
Joost R. Van Ginkel
Leiden University
Faculty of Social and Behavioural Sciences
Data Theory Group
PO Box 9555
2300 RB Leiden
The Netherlands
Tel: +31-(0)71-527 3620
Email: jginkel@fsw.leidenuniv.nl

Abstract

The focus of this study was the incidence of different kinds of missing data problems in personality research and the handling of these problems. Missing-data problems were reported in approximately half of more than 800 articles published in three leading personality journals. In these articles, unit-nonresponse, attrition, and planned missingness were distinguished but missing item scores in trait measurement were reported most frequently. Listwise deletion was the most frequently used method for handling all missing-data problems. Listwise deletion is known to reduce the accuracy of parameter estimates and the power of statistical tests and often to produce biased statistical analysis results. This study proposes a simple alternative method for handling missing item scores, known as two-way imputation, which leaves the sample size intact and has been shown to produce almost unbiased results based on multi-item questionnaire data.

Keywords: *incidence of missing data*, *missing item scores, Two-way imputation, questionnaire data, multiple imputation of item scores.*

Introduction

Multi-item questionnaires, inventories, and checklists—henceforth, generically called questionnaires—are widely used for measuring personality traits. Multiple items are used to cover all relevant aspects of a trait in an effort to measure the trait validly, and to control measurement error to a degree that the total score on the questionnaire is reliable. Examples of traits measured by means of multi-item questionnaires are obsessive-compulsive disorder, depression, and anxiety. The Obsessive-compulsive inventory (Foa, Kozak, Salkovskis, & Amir, 1998) is a well-known questionnaire for measuring obsessive-compulsive disorder, the Beck depression inventory II (e.g., Segal, Coolidge, Cahill, & O'Riley, 2008) measures depression, and the Beck anxiety inventory (e.g., Morin et al., 1999) measures anxiety.

Even when respondents have been instructed explicitly to respond to all items and not leave any responses open, data collection by means of multi-item questionnaires regularly suffers from missing item scores. Often the researcher is in the dark with respect to the reasons of this item nonresponse. In many cases, re-approaching respondents is an unrealistic option because of anonymity guarantee or financial or other restraints. Thus, the researcher often has to accept the incidence of the missing item scores and make a decision how to handle this problem in the statistical analysis of the data. One popular strategy is to leave out the cases that have at least one missing score and analyze only the complete cases. This strategy is called listwise deletion.

Our experience is that listwise deletion is an immensely popular method for handling missing item scores but is has a few serious drawbacks. By definition, it always reduces the sample size, which has the effect of reducing the accuracy of estimation and the power of statistical testing. In addition, under many circumstances listwise deletion may even cause more harm by producing biased statistical results (Little & Rubin, 2002; Schafer, 1997). For example, means and correlations may be distorted, which may affect the outcomes of methods such as the Student's *t* test and factor analysis. Also, see Burton and Altman (2004), who corroborated the dominance of listwise deletion in the context of cancer research.

The large-scale application of listwise deletion suggests that researchers may not always realize the potentially damaging effects of listwise deletion on their research outcomes and also may not be aware of the availability of simple and

statistically superior methods for handling missing data that keep these damaging effects to a minimum. Thus, this study has two purposes. First, by means of a literature search we focus on the incidence of several kinds of missing-data problems that are reported in the literature on personality research. These missing-data problems also include missing item scores in multiple-item questionnaires, which constitute a large portion of the general missing-data problem. Also, we record the methods used in practice to handle missing-data problems. Second, we suggest a simple and statistically superior alternative to listwise deletion, that does not have the damaging effect of listwise deletion in multi-item trait measurement. We illustrate the method by solving the missing item-score problems in a real data set.

Missingness Mechanisms and Real-Data Analysis

An example using a real data set (Vorst, 1992; also, see Van der Ark, 2007) collected by means of a Dutch translation of the Adjective Checklist (ACL; Gough & Heilbrun, 1980) may illustrate the problem of item nonresponse, which leads to missing item scores. The 218 items of the ACL are divided across 22 subscales; see Table 1. A sample of $N = 433$ students from the University of Amsterdam provided ordered scores on a five-point rating scale, scored 0 (completely disagree) to 4 (completely agree). The data were completely observed; thus, there were no missing item scores. The completeness of the real data enabled us to manipulate mechanisms that created item nonresponse so as to illustrate what listwise deletion can do to the statistical results, but first we consider the complete data results.

INSERT TABLE 1 ABOUT HERE

Suppose a researcher uses the total score on the ACL Aggression subscale (items 101-110) and the ACL Dominance subscale (items 21-30) to test the hypothesis that aggressive people tend to be more dominant than non-aggressive people. To this end, (s)he uses a median split of the total scores on Aggression to divide the respondents into 'aggressive' respondents and 'non-aggressive' respondents. The researcher is interested in the mean difference in the total Dominance score between aggressive and non-aggressive people. To test whether this difference is significant, (s)he performs a two-sample *t*-test with the dichotomized

aggression score as the independent variable, and the total Dominance score as dependent variable. The researcher is also interested in the range, the mean, and the reliability of the Dominance subscale in the total sample. Table 2 (first row) shows that Cronbach's (1951) alpha equaled 0.807, and that the relationship between aggression and dominance was significant (p = .024).

INSERT TABLE 2 ABOUT HERE

The statistical literature (Little & Rubin, 2002, p. 12; Schafer, 1997) distinguishes three mechanisms that may produce missing scores on variables. Listwise deletion always leads to a reduced sample size irrespective of which mechanism caused the missing item scores, but it leads to biased results under two of the mechanisms. Unfortunately, these are the mechanisms that are the most likely to cause missing-data problems in practical research. Thus, for a better understanding of the problems involved in using listwise deletion and the solutions of these problems, it is necessary to understand these three mechanisms. Each is explained next, and their effects on data analysis after the application of listwise deletion are illustrated using the ACL data.

The Missing Completely at Random Mechanism

The first mechanism produces missing item scores as if they constituted a simple random sample from all scores in the data. There is no relation to the value of the item score that is missing, nor to any other variable. In this case, the missing item scores are *missing completely at random* (MCAR; Little & Rubin, 2002, p. 12). This is the only situation in which listwise deletion is guaranteed not to result in biased outcomes. However, reduction of the sample size and its effects on accuracy and power are unavoidable.

The MCAR mechanism in the Dominance data was simulated by randomly drawing entries from the data matrix, which consisted of 433 rows (respondents) and 10 columns (Dominance items), removing the item scores corresponding to these entries, and considering the resulting data matrix as suffering from item nonresponse. For this example, entries were drawn with a probability equal to .05 and without replacement; this produced a sample of 217 entries [433 (respondents) × 10 (items) ×

0.05 (probability) = 216.5], and the corresponding item scores were removed. Listwise deletion resulted in a 40% reduction of the sample; that is, $N = 258$ complete cases were left for statistical analysis.

Because the reduced sample was a simple random sample drawn from the complete sample, we did not expect biased results. Table 2 (second row) shows that Cronbach's alpha dropped from 0.807 to 0.802, which reflects sampling error. The mean and the range of the test score were also similar to those found in the complete sample. However, a smaller sample size leads to a loss of power, which was apparent from a nonsignificant *t*-test compared to a significant result in the complete sample. Also, the mean difference has become smaller, which also reflects sampling error. Thus, listwise deletion may have important consequences for the outcomes of research.

The Missing at Random Mechanism

The second mechanism also produces missing item scores as if they constituted a random sample from the data, but the missingness is related to one or more observed variables in the data; hence, the missing item scores do not constitute a simple random sample. Missing scores are now said to be *missing at random* (MAR; Rubin, 1976; Little & Rubin, 2002, p. 12). The next example may further clarify the MAR mechanism.

Suppose we distinguish decent citizens from indecent citizens (e.g., due to hazardous traffic behavior, littering the street, not waiting in line at the bakery). A median split of the ACL Communality subscale total score produced groups of decent people and indecent people. Suppose that indecent people have a probability of not responding to items in the Dominance subscale that is three times as high as the corresponding probability for decent people. Thus, whether scores on dominance items are missing depends on the total score on Communality, which is an observed variable in the data. As this variable explains the missingness, it may be used to fix the missing data problem. Because listwise deletion ignores such explanatory variables, it now produces biased statistical results**.**

The MAR mechanism was simulated by randomly drawing 217 entries from the data (i.e., 5% missingness), such that respondents low on Communality had a probability of missing a Dominance-item score that was three times higher than

respondents high on Communality. After the corresponding item scores were removed, listwise deletion resulted in a 39% reduction of the sample, leaving $N = 265$ cases for statistical analysis. Table 2 (third row) shows that Cronbach's alpha increased by 0.003**,** and that the *t*-test was not significant. The mean test score was similar to the mean test score in the complete-data example and the MCAR example. However, the maximally observed test score decreased from 40 to 38. Hence, the MAR mechanism produced results that are slightly worse than the MCAR mechanism.

The Miscellaneous Category: Not Missing at Random Mechanisms

The third category contains all the mechanisms that produce missingness that is related to the value that is missing or to one or more variables that are not in the data of the study under consideration. These mechanisms produce missingness such that item scores are *not missing at random* (NMAR; Little & Rubin, 2002, p. 12). The problem here is that the researcher has no knowledge of the causes of the missingness, and thus is not in a position to solve the problem adequately. Because of the solution of NMAR problems requires knowledge that is inaccessible, one may resort to solutions assuming MAR in an effort to fix the problem as much as possible.

NMAR was simulated by removing 217 item scores (i.e., 5% missingness), such that for scores of 3 and higher, the probability of being missing was three times as high as for scores lower than 3. Table 2 (fourth row) shows that, compared to the original data, Cronbach's alpha increased by 0.011. The mean test score was underestimated. The maximum test score decreased from 40 to 38. The *t*-test is not significant.

Study 1: Incidence of Missing Data in Personality Measurement

In Study 1, we investigated the frequency with which particular types of missing data were reported in articles discussing personality-trait measurement. Prior to discussing the results from the first study, we discuss the four types of missing data that were frequently reported: item nonresponse, unit nonresponse, attrition, and planned missingness. Because we already discussed item nonresponse, we now limit attention to *unit nonresponse*, *attrition*, and *planned missingness*.

Unit nonresponse occurs when a participant drawn into the sample refuses to take part in the investigation, so that for this person no observed data exist. De Leeuw and Hox (1988), Dillman (1991), and Groves and Couper (1998) have extensively studied the statistical handling of unit nonresponse.

Attrition occurs when participants drop out of a longitudinal study in which they are objected to repeated observation. Dropout may be due to loss of interest or motivation to proceed, having moved to another city, and in medical and health studies due to complete recovery, becoming too ill to further participate, or passing away as a result of the illness. Fleming and Harington (1991) and Andersen, Borgan, Gill, and Kleiding (1993) discuss methods for statistically dealing with attrition.

Planned missingness results from the researcher's intentional planning. For example, in a medical screening using multiple tests, for reasons of efficiency the researcher may not administer all tests to all participants. Eggen and Verhelst (1992) and Mislevy and Wu (1988) discuss statistical methods for handling planned missingness in the context of educational measurement.

Method

We used the following strategy for studying the incidence of missing-data problems in personality measurement. A total of 832 articles from six recent volumes (1995, 1997, 2000, 2002, 2005, and 2007), four issues per volume, of three personality journals (*Psychological Assessment*, *Personality and Individual Differences*, and *Journal of Personality Assessment*) were screened for report of missing-data problems. The four issues per volume were selected as follows: *Psychological Assessment* is issued four times per year, *Personality and Individual Differences* is issued monthly (arbitrarily, the January, April, August, and December issues were selected), and *Journal of Personality Assessment* is issued six times per year (arbitrarily, the February, July, August, and December issues were selected). When multiple types of missingness were reported within the same article, the article was counted multiply. This yielded a total count of 927 cases whitin 832 articles.

Results

Table 3 shows that 30% of the 927 cases pertained to item nonresponse (third column). Unit nonresponse and attrition are typical of survey studies and longitudinal studies, which are types of research that are not published as regularly in the three journals as personality measurement studies. Several articles specified the number of participants who provided incomplete score patterns but did not mention the type of missing data, and a few articles reported the removal of participants but not whether removal was due to missing scores or other reasons (e.g., random responding). Articles that mentioned nonresponse but did not mention the type of nonresponse were classified as 'Not Clear' (Table 3).

INSERT TABLE 3 ABOUT HERE

Table 4 shows descriptive statistics (mean, standard deviation, minimum, and maximum) of the proportion of incomplete score patterns computed across the 369 cases where the proportion of incomplete cases was reported. The distribution of the proportion of incomplete score patterns is positively skewed**,** which means that most articles reported little missing data, and a small number of articles (6%) reported a large proportion of incomplete score patterns (30% or more). For item nonresponse, the percentage of incomplete item-score patterns on average equaled 9%. Thus, on average listwise deletion would result in a sample reduction of approximately 9%. Some articles reported the presence of missing item scores, but not the percentage of incomplete score patterns.

INSERT TABLE 4 ABOUT HERE

Discussion

Almost half of the articles reported missing-data problems. Assuming that some articles failed to report such problems, the incidence of missing-data problems in personality measurement may even be greater. Item nonresponse was reported more often than other types of missing data. Item nonresponse occurs frequently in personality trait measurement using multi-item questionnaires. Item nonresponse is a serious problem in data analysis that calls for effective solutions that are easy to understand and implement.

Study 2: Handling Missing Data in Personality Measurement

In Study 2, we investigated the methods researchers in personality measurement typically use for handling missing-data problems.

Method

The observations were the 927 missing-data problems used in Study 1. The independent variable was missing-data type, which had six levels: unit nonresponse, attrition, item nonresponse, planned missingness, not clear, and none reported (Table 3). The dependent variable was the method researchers in personality measurement use to handle missing-data problems. Seven principal methods for missing-data handling were found to be used in the 832 articles: follow-up, listwise deletion, available-case analysis, single imputation, direct maximum likelihood, variable deletion, and prorating. In addition, four variations or combinations of principal methods were identified: listwise deletion with a check for MCAR and MCAR not rejected; listwise deletion with a check for MCAR but MCAR rejected; Available case analysis with a check for MCAR and MCAR not rejected; and a combination of follow-up and listwise deletion with a check for MCAR). Also, two rest categories were identified and categorized as ‘other’ and ‘none reported’. Addition of these missing-data handling methods led to a dependent variable having 7 + 4 + 2 = 13 levels. The seven principal methods were also used to handle item nonresponse. These methods and another method known as *multiple imputation* are discussed below. Some of the methods are illustrated using an incomplete-data example (see, Sijtsma &

Van der Ark, 2003), which is shown in Table 5. This data set contains the scores of 8 fictitious respondents on 5 items.

INSERT TABLE 5 ABOUT HERE

Follow-up. Perhaps the best way to deal with missing data is re-approaching respondents with incomplete score patterns in an effort to obtain the scores that are missing. When successful, data that were initially missing become observed, and statistical analyses may be carried out without any problems, and without running the risk of obtaining biased results. For an example, see Huisman, Krol, and Van Sonderen (1998) who re-approached patients in a study with respect to the waiting list problem in orthopedic practices. Unfortunately, however, due to many different restraints, in many studies follow-up is not feasible.

Listwise deletion. Consider the data in Table 5. Suppose a researcher plans computing Cronbach's alpha for the total score on the items X_1, X_2, and X_3, and the correlation between the items X_4 and X_5. Listwise deletion uses cases 2, 4, and 7 for both computing Cronbach's alpha and the correlation. Advantages of listwise deletion are that statistical analyses can be done without any modifications on the data and that all statistical analyses are done on the same subsample. Disadvantages are that the reduction of the sample size results in a loss of estimation precision and a reduced power in hypothesis testing. Furthermore, unless the missing scores are MCAR statistics may be biased. Listwise deletion may be preceded by a check whether MCAR is a reasonable assumption. This check may entail testing whether respondents with completely observed item-score patterns and respondents with incomplete or blank item-score patterns differ significantly with respect to demographic variables such as gender or ethnicity. For example, when the background variable 'age' is observed for all respondents, a two-sample *t*-test may be used to test whether respondents with complete score patterns differ systematically with respect to age from respondents with incomplete score patterns. For categorical background variables, such as gender, chi-square tests may be used. See, for example, Hishinuma et al. (2000), and Cole, Hoffman, Tram, and Maxwell (2000) who used this strategy for checking the MCAR assumption.

Available-case analysis. Loss of power may be reduced when all cases are used in the statistical analysis, which have observed values on the variables that are

effective in the analyses. This option is called available-case analysis. When applied to the data from Table 5, available-case analysis uses cases 1, 2, 4, 6, 7, and 8 for computing Cronbach's alpha for the total score on the items X_1, X_2, and X_3. For computing the correlation between the items X_4, and X_5, available-case analysis uses cases 2, 3, 4, and 7. Available-case analysis (Little & Rubin, 2002, pp. 53-54) is the default option for missing-data handling in SPSS (2008).

Compared to listwise deletion, a disadvantage of available-case analysis is that different statistical analyses that use different variables may be based on (partly) different sub-samples with different sample sizes. A disadvantage shared with listwise deletion is that statistics may be biased unless the missingness mechanism is MCAR. Kim and Curry (1977) showed that available-case analysis is superior to listwise deletion when correlations among variables are modest. Haitovsky (1968) and Azen and Van Guilder (1981) showed that listwise deletion is superior to available-case analysis when correlations among variables are large. Little and Rubin (2002, p. 55) argued that both options are generally unsatisfactory.

Because listwise deletion and available-case analysis result in a loss of power and possibly biased results, researchers should be cautious using these methods. It may be recommended to use these methods only when the reduced sample is large and when it has been checked whether there are systematic differences on the background variables between the completely observed cases and the incomplete cases, so that the MCAR assumption at least is plausible.

Single imputation. Single imputation replaces the missing scores by plausible scores, so that cases that have missing scores can be included in the statistical analyses. We discuss two possibilities.

Deterministic imputation replaces the empty cells in the data matrix by estimates of the item scores. For example, Saggino and Kline (1995) replaced each missing score on variable X by the sample mean of X based on the available scores, and Sheviin and Adamson (2005) replaced each missing score by the expected value from a regression model. Table 6 (upper left panel) shows how variable-mean imputation is done in the incomplete-data example in Table 5. The imputed scores are derived readily by computing the means for each variable (last row). For example, the imputed score on variable X_1 is computed as $(2 + 3 + 4 + 1 + 5 + 1 + 3)/7 = 2.71$. Note that the resulting imputed scores are not necessarily integer scores. Depending on the application, imputed scores may be analyzed as real numbers (e.g., as in factor

analysis, which treats rating-scale scores as continuous) or they may be rounded to the nearest feasible integer (e.g., as in item analysis using item response models, which treat rating-scale scores as discrete).

INSERT TABLE 6 ABOUT HERE

Table 6 (upper right panel) also shows the completed data set that results from deterministic regression imputation. Imputations were done using SPSS 16.0 (Analyze, Missing Value Analysis). The imputed scores are less easily derived because the computation procedure that SPSS uses is rather complicated.

The advantage of deterministic imputation is that it provides the researcher with a complete data set, which may be used for further statistical analysis. A disadvantage is that variances and covariances are biased downward (Schafer, 1997, p. 2).

Stochastic imputation improves upon deterministic imputation by imputing a value that includes a random error; For example, in regression imputation the imputed value includes a normally distributed random error with variance equal to the error variance of the regression model. Thus, the imputed values have the same variance as the observed scores. Stochastic imputation keeps the covariance structure intact but in subsequent statistical analyses the imputed scores are treated as if they were observed without taking the uncertainty about these imputed values into account. As a result, the standard errors of the statistics are too small.

Table 6 (lower left panel) shows how stochastic variable mean imputation is done. Here, the imputed values are random draws from from a normal distribution rather than a mean substitution. For example, the imputed score on variable X_1 is a random draw from a normal distribution with a mean of 2.71 and a standard deviation of 1.50 (last row).

Because the detailed explanation of how the computations for both deterministic and stochastic regression imputation are carried out would be too involved, we only show the syntax that performs the imputations in SPSS. Here, it is assumed that the incomplete data set is named example.sav and located in the directory C:\imputation\, and that the completed data files are called deterministic.sav and stochastic.sav. The resulting syntax file is shown in Figure 1. Note that the 12th line (SET SEED = 2 .) is only added to reproduce the results from the example (Table 6) for

stochastic regression imputation. To obtain imputed values that differ from the example, this line may be removed.

INSERT FIGURE 1 ABOUT HERE

Multiple imputation. Multiple imputation improves upon stochastic imputation by substituting multiple random values (i.e., not necessarily integer scores) for each missing score, resulting in several plausible complete versions of the data. These completed data sets are then analyzed by standard statistical procedures, and the results are combined into one overall result, using rules proposed by Rubin (1987, chap. 3). Schafer (1997, p. 106) recommends doing the statistical analyses on three, four, or five completed data sets.

An advantage of multiple imputation compared to single imputation is that statistical analysis takes the uncertainty about the missing data into account, so that standard errors of statistics are not biased downwards. Moreover, whereas listwise deletion and available-case analysis only lead to valid inferences when scores are MCAR, multiple imputation also leads to valid inferences when scores are MAR. A disadvantage of multiple imputation is that the method is rather involved and only available in software packages that are not frequently used among personality researchers. Examples of software are SAS 8.1, in the procedure PROC MI (Yuan, 2000), S-plus 8 for Windows (2007), AMOS 6.0 (Arbuckle & Wothke, 2006), the stand-alone program NORM (Schafer, 1998), ICE in Stata 10.0 (StataCorp, 2007), the MICE library in S-plus, and the stand-alone program WinMICE V1.0 (Jacobusse, 2005).

Table 7 shows 5 completed versions of the incomplete data set in Table 5. Multiple imputation was done using the program NORM (Schafer, 1998). Cronbach's alpha for the total score on the items X_1, X_2, and X_3 may be obtained as the mean of the five alpha values obtained from the five imputed data sets. The same goes for the correlation between the variables X_4 and X_5. To test the significance of the correlation, an overall standard error has to be computed across the five imputed data sets using Rubin's (1987) rules. See Rubin (1987, Chap. 3) for an extensive discussion of these rules.

INSERT TABLE 7 ABOUT HERE

Direct maximum likelihood estimation. Direct maximum likelihood estimation (e.g., Allison, 2002) entails estimating the parameters from a statistical model while ignoring the unobserved scores but without deleting cases. Thus, unlike listwise deletion and available-case analysis, direct maximum likelihood estimation uses all observed item scores instead of using only the scores of respondents with complete item-score patterns. The method is used for the estimation of, for example, item response theory models, latent class models, and structural equation models. An advantage of direct maximum likelihood estimation is that all cases are used to estimate the model. A disadvantage of the method is that, like most multiple imputation methods, it is relatively complex and can only be used in nonstandard statistical procedures and nonstandard statistical software packages. The method cannot be used in popular procedures like principal components analysis and analysis of variance (ANOVA). Moreover, SPSS (2008) does not allow using the method even for procedures that are suited for it, such as factor models or loglinear models.

Prorating test scores. Prorating test scores entails computing a respondent's test score across his/her observed scores and then rescaling the resulting score. Together with the total scores for respondents with complete data, these resulting scores are used as dependent variable in statistical analyses. In Table 5, the test score of person 2 is computed as $3 + 5 + 4 + 5 + 5 = 22$, and the *prorated* test score of person 1 is computed as $[(1 + 1 + 2) / 3] \times 5 = 6.67$.

This method does not explicitly impute scores but is equivalent to substituting for each missing value the person mean across a respondent's available scores. This procedure is common practice, and is even recommended in manuals of many personality-trait questionnaires (e.g., Bracken & Howell, 2004; Hare, 2003). However, from a statistical point of view, prorating test scores is a suboptimal method. First, it does not take the differences between item means into account. Second, because the mean test score across the remaining items does not have an error component, the variance of the test score is biased downwards.

Variable deletion. Variable deletion leaves out variables with missing scores from the statistical analysis. Thus, for items it is the counterpart of listwise deletion. The missing-data literature does not explicitly mention this procedure as a useful method but researchers often use it. For example, when information on gender is missing for some respondents a researcher may decide not to use gender as an

independent variable in statistical tests but to use it only for describing the demographic characteristics of the sample. See, for example, Watson, et al. (2007), who reported that "The sample consisted of 376 women and 121 men (2 participants did not specify their sex)". Another example of variable deletion may concern a particular item, which has so many missing values that the researcher may decide to leave it out of the reliability analysis and compute test scores across the remaining items. In the data example of Table 5, a researcher may decide that item X_4 has too many missing values to be useful for any statistical analysis. Thus, (s)he may decide not to compute the correlation between items X_4 and X_5. Because variable deletion does not result in a selective dropout of respondents, it gives valid results in statistical analyses but limits the substantive meaning of the research.

Results

Table 8 shows that listwise deletion is by far the most frequently used missing-data method, followed by available-case analysis. Single imputation was used 19 times, and multiple imputation was not used at all. Some studies used several methods of handling nonresponse. Each method was counted separately, leading to a total of 1025 cases of missing-data handling rather than 927 as in Table 3. Only few studies checked whether MCAR was plausible prior to deleting the cases from the analyses. All of these studies, regardless of the outcome of this check, conducted the statistical analyses based on the complete cases, and only in the discussion section they mentioned that the sample was probably not completely representative, thus resulting in limited generalizability.

Two articles reported a combination of follow-up and listwise deletion preceded by a check for MCAR (row 12). Specifically, Iversen and Rundmo (2002) reported that "A control study was conducted to find out if the group of respondents who had replied to the questionnaire differed significantly from those who did not. Fifty subjects were contacted by phone and interviewed using the same questionnaire as in the survey. Results from this study showed that the final sample was representative of the population of Norwegian drivers with regard to age, gender and education."

INSERT TABLE 8 ABOUT HERE

Discussion

Personality-trait measurement using multiple-item questionnaires predominantly uses listwise deletion for handling missing data problems. The popularity of listwise deletion probably resides in its simplicity but researchers seem to be unaware of its potential problems. We give two possible explanations. First, it may be incorrectly assumed that missing scores make a score pattern useless so that the pattern better be discarded from the data analysis. Second, it may be incorrectly assumed that deleting cases only reduces power, whereas the bias resulting from nonresponse may not be appreciated. We noted that missing data were often discussed as if they were nothing more than a nuisance in the data-collection process, which could simply be remedied by collecting enough data so that after listwise deletion enough cases were left for analysis.

Sometimes, listwise deletion is a good solution for missing item-score problems. For example, respondents who have almost no observed data may be discarded from the data analyses. Also, when only a few respondents out of a relatively large sample have incomplete item-score records leaving them out of the analysis has little effect on the outcomes of statistical analysis. For example, Boyd-Wilson, Walkey, McClure, and Green (2000) deleted two incomplete cases from a total sample of $N = 205$. However, listwise deletion was used so frequently that it seems safe to conclude that it is often used inappropriately.

The popularity and dominance of listwise deletion seems to have the effect of hiding simple, user-friendly and statistically superior alternatives for the handling of item nonresponse from the researchers' statistical toolbox. Given the availability of such alternatives and the established inferiority of listwise deletion in many research situations, next we discuss an attractive method for handling item nonresponse in multi-item questionnaires for personality-trait measurement.

A Simple Method to Handle Item Nonresponse in Multi-Item Questionnaire Data

For multiple-item questionnaire data, the most promising simple imputation method is *two-way multiple imputation with error* (abbreviated Method TW; Little & Su, 1987, discussed the core of Method TW in the context of incomplete longitudinal data, and

Bernaards & Sijtsma, 2000, proposed using the method for questionnaire data; also see Van Ginkel et al., 2007a; 2007b, and Van Ginkel, Van der Ark, Sijtsma, & Vermunt, 2007). In the Appendix we show how Method TW can be used by means of SPSS (2008).

Method TW is based on a typical ANOVA layout. We assume that the scores of N persons to J items measuring a single personality trait are incomplete. Let PM_i denote the mean item score of person i based on his/her available item scores, let IM_j denote the mean score of item j based on all scores available for this item, and let OM be the overall mean of all available item scores in the $N \times J$ data matrix. A deterministic imputation method may use $TW_{ij} = PM_i + IM_j - OM$ to impute a score for a missing value in cell (i, j) of the data matrix, and a probabilistic imputation method adds an error term ε_{ij} and then imputes $TW_{ij}^* = TW_{ij} + \varepsilon_{ij}$. Depending on the application, imputed TW_{ij}^* scores are analyzed as real numbers (e.g., as in factor analysis) or rounded to the nearest feasible integer (e.g., as in item analysis using item response models).

The computation of TW_{ij}^* is illustrated next using the data example in Table 5 for person 5 and variable X_1. It may be verified that $PM_5 = (3 + 3 + 4) / 3 = 3.33$, $IM_1 = (2 + 3 + 4 + 1 + 5 + 1 + 3) / 7 = 2.71$, and $OM = 95 / 33 = 2.88$; hence, $TW_{51} = 3.33 + 2.71 - 2.88 = 3.16$. The other values of TW_{ij} from the example in Table 5 are shown in Table 9.

INSERT TABLE 9 ABOUT HERE

Next, the error ε_{ij} is drawn from a normal distribution with mean 0 and variance S_ε^2; S_ε^2 is the error variance in the observed data, which is computed as follows. First, for each observed item score X_{ij} the corresponding TW_{ij} score is computed. The TW_{ij} scores are considered to be the expected scores of the two-way model, had the X_{ij} scores been missing. Second, the sum of the squared differences, $(X_{ij} - TW_{ij})^2$, is computed across all observed cells, and this sum is divided by the number of observed scores minus 1 (denoted by M; in Table 5, $M = 33 - 1 = 32$). Thus, we find that $S_\varepsilon^2 = \sum\sum (X_{ij} - TW_{ij})^2 / M$.

Multiple imputation based on five independent draws of the error is done as follows. For the data in Table 5 the error variance equals 0.901 (it may be noted that for computing a TW_{ij} score, the corresponding observed X_{ij} score is treated as missing; as a result, the person and item means vary with each cell (i, j), and the person and item means in Table 9 cannot be used throughout the computation of the error variance. These details are ignored here). Assume that five randomly drawn error terms are: $\varepsilon_{51}^{(1)}$ = -0.1601879, $\varepsilon_{51}^{(2)}$ = -1.0220348, $\varepsilon_{51}^{(3)}$ = 0.4451876, $\varepsilon_{51}^{(4)}$ = 2.5191623, and $\varepsilon_{51}^{(5)}$ = -0.6389984. For producing consecutive data matrices, each of these values is added to $TW_{51} = 3.17$, which yields five different values (rounded to two decimals): TW_{51}^{*} = 3.01, 2.15, 3.61, 5.69, and 2.53, respectively. Each of these values is imputed in the data matrix in Table 5 (thus treating scores as continuous). The same procedure is followed for the other missing values (not shown here), which yields five different completed data sets. Statistical analyses are done on all five data sets separately, and the results are combined using Rubin's (1987, chap. 3) rules.

Simulation results (Van Ginkel et al., 2007a; 2007b, Van Ginkel, Van der Ark, Sijtsma, & Vermunt, 2007) have shown that Method TW produces statistical results with very little or no bias at all, even when missing item scores are NMAR and the percentage of missing item scores increases up to 15% (in these studies, this corresponded to only 4% completely observed cases on average). A plausible explanation why Method Two-Way works so well in case of NMAR is because multiple items are used to measure the same construct. Even if some extreme NMAR missingness results in many missing item scores for certain respondents, these respondents will usually have responded to some items measuring the same construct. The observed item scores contain enough information to predict the missing item scores reasonably well. Only in case of extremely high percentages of missingness, Method Two-Way will result in biased estimates (see, Van Buuren, 2009)**.** This is an important finding implying that a researcher may safely use Method TW to impute item scores in multiple-item questionnaires for measuring personality traits.

To illustrate the usefulness of Method TW, we simulated item nonresponse in the multiple-item ACL Dominance subscale (Table 1) for item scores that were either

MCAR, MAR[1], or NMAR, thus producing three different incomplete data sets. We used Method TW to impute scores in each of the three data sets, and computed the values of Cronbach's alpha, the mean test score, the minimum and maximum observed test scores, and the *t*-test, with Aggression as the independent variable and the Dominance test score as the dependent variable (Table 10).

INSERT TABLE 10 ABOUT HERE

Almost all results produced by multiple imputation using Method TW were closer to the results produced by the complete data than the results produced by listwise deletion (cf. Table 1). For the MAR data set, the maximum test score was underestimated, but less than for listwise deletion (cf. Table 1, fourth column). For the three completed data sets, the *t*-test (last three columns) was significant, as in the original data.

First, it may be noted that when a test contains more than one subscale, Method TW may be applied to each subscale separately. Two other versions of Method TW, not discussed here, use the multidimensionality of the data for imputing scores; see Van Ginkel et al. (2007b) for more details. Second, Method TW should be applied only if PM_i can be interpreted as an indicator of the trait level of person i (method TW capitalizes on each of the J items holding information on the other items). PM_i cannot be interpreted as an indicator of the trait level of person i if items are included that do not measure the intended trait, such as gender or social economic status, or if a respondent has excessively many missing values. In the former case, other methods such as multiple imputation under the latent class model may be used (Vermunt, Van Ginkel, Van der Ark, & Sijtsma, in press), and in the latter case such exceptional cases may be removed before method TW is used.

General Discussion

Item nonresponse occurs frequently in personality measurement. Even though multiple imputation is a highly recommended procedure in the statistical literature for

[1] It may be noted that even though the missingness only depends on the fully observed variable 'Communality group', the default application of Method Two-Way does not impute scores separately for 'Communality group = 1' and 'Communality group = 2'. Therefore, technically, Method Two-Way treats this condition as NMAR.

dealing with item nonresponse, this method appears to be used rarely if ever in personality measurement. Instead, the inferior listwise deletion method is by far the most popular method for handling missing item scores.

The screening of three leading personality journals underlined the need for simple, user-friendly and statistically correct methods to deal with item nonresponse in questionnaire data. Method TW has these properties and may be used for the imputation of item scores. SPSS macros for multiple item-score imputation are available as freeware from the Internet (http://www.uvt.nl/mto/software2.html; Van Ginkel & Van der Ark, 2005a; 2005b). In an empirical-data example, it was shown that method TW accurately recovered several statistics typical of the psychometric analysis of questionnaire data. Thus, method TW may be a good alternative for listwise deletion and other missing-data handling methods for handling missing item scores in personality measurement. Method TW is appropriate for multi-item questionnaire data, in which the items all measure aspects of one underlying personality trait and a total score is typically used for measuring individuals but the method may also be extended to multidimensional questionnaire data.

APPENDIX

SPSS syntax is available to conduct the following types of statistical analyses on test data with missing item scores using Method TW.

1. *Computation of a statistic without standard error* (e.g., reliability statistics such as Cronbach's alpha and corrected item-total correlations; descriptive statistics such as the mean, standard deviation, median, maximum, and minimum; correlation coefficients, loadings from factor analysis). As an example we show how to compute Cronbach's alpha for a dominance test containing 10 items.
2. *Computation of a statistic with standard error*. Note that in several cases SPSS does not provide standard errors and they have to be computed by the researcher. As an example we show how to compute the mean score on a dominance test containing 10 items, its standard error, and 95% confidence interval.
3. *All t-tests and univariate regression analyses* can be computed in a straightforward way. As an example, we show how to compare the mean scores on a dominance test of a group of non-aggressive and a group of aggressive respondents using a two-sample *t*-test.
4. For *other analyses* (multivariate regression, multi-level analysis, ANOVA, significance tests for correlations, mixed models) the procedures are more involved and we refer to Van Ginkel (2006) for detailed information.

Statistical analyses that cannot be performed include MANOVA and structural equation models.

The necessary files for the exemplary statistical analyses can be obtained from http://www.uvt.nl/mto/software2.html in the zip file imputation.zip, which contains four files:

- ACL.sav: An SPSS data file containing the item scores of 433 persons to 10 dominance items (V021 to V030), 5% of the scores are missing (MCAR); and their scores on variable Naggress (score 1 indicates non-aggressive behavior, score 2 indicates aggressive behavior).

- imputation.sps: An SPSS syntax file performing statistical analyses on the incomplete data file ACL.sav, using Method TW[2].
- tw.sps: An SPSS syntax file containing preprogrammed macro tw.
- mi.sps: An SPSS syntax file containing preprogrammed macro mi.

These four files should be unpacked and moved to the same directory. Without loss of generality we assume that this directory is called C:/imputation/. The analyses are performed by running imputation.sps, which is discussed next.

The file imputation.sps contains four steps.

- *Step 1*: *Preliminary commands* (lines 1-7). Determining the working directory (lines 4-5). If the unzipped files are not in C:/imputation/ the FILE HANDLE command (line 5) should be modified before use. Line 7 ensures that the results in the Appendix are reproduced exactly; this line should be removed if imputation.sps is modified for other data sets. Line 7 supresses the printing of syntax commands in the output. The command prevents that the many syntax commands from mi.sps and tw.sps are printed in the output.
- *Step 2: Creating five completed data sets* (lines 9-16). Line 13 reads the preprogrammed macro tw.sps. Five completed versions of acl.sav are created by the command TWOWAY. Subcommand /SELECT specifies the items to which Method TW is applied and subcommand /M specifies the number of required completed data sets; here $M = 5$. Running TWOWAY results in a single SPSS data file containing five completed versions of ACL.sav. This file, which is automatically called ACL_imp.sav, contains all five completed datasets appended one after another. An additional variable called imputation_# has been added, which indicates the data set number.
- *Step 3: Conducting statistical analysis* (lines 18-56). First, data file ACL_imp.sav is read and split into five separate data sets (lines 20-22). In SPSS, the split file option may be found under task bar: Data, Split File. Second, five Cronbach's alphas are computed using the command RELIABILITY (line 31). RELIABILITY is preceded by the command OMS and followed by the

[2] This file is based on the package tw.zip (Van Ginkel & Van der Ark, 2005a, 2005b; Van Ginkel, 2006). This package is more general than the syntax presented here and has an extensive manual. To allow a brief yet concise explanation of Method TW, we have modified these general files and collected them in a single syntax file.

command OMSEND. These commands direct SPSS output into an SPSS data file[3]. The resulting file reliability.sav contains the five values of Cronbach's alpha. Similarly, the mean test score and the standard deviation are computed using DESCRIPTIVES and the output is directed to descriptives.sav (lines 42-44), and the *t*-test is performed and the output is directed to ttest.sav (lines 46-56).

- *Step 4: Combining the results of the five statistical analyses* (lines 58-86). First, the five Cronbach's alphas, collected in reliability.sav, are combined (lines 60-62). The Cronbach's alpha that should be reported is obtained by simply taking the mean of the Cronbach's alphas of the five data sets. The output shows that Cronbach's alpha equals .8105. Second, the mean test scores (Mean) and standard deviations (Std.Deviation), collected in descriptives.sav, are combined (lines 64-74). This is a little bit more involved. The standard error of the mean is not provided by SPSS and must be computed separately as S.E.Mean = Std.Deviation / $\sqrt{N}$ (line 66). Furthermore, the even lines in descriptives.sav contain no information and they are removed (line 65). The command RULESMI gives the correct combination of the statistic and standard error. The output shows that the mean test score equals 24.398, its standard error equals 0.292, and the 95% confidence interval is [23.825; 24.972]; the remaining statistics (*t* statistic, df, and *p*-value) can be ignored here. Third, in a similar way the results of the *t*-test are combined (lines 76-87). Note that ttest.sav contains the results for both 'equal variances assumed' and for 'equal variances not assumed' whereas we are only interested in *t*-tests where equal variances are assumed. The other results are deleted in line 77. For the command RULESMI the difference in mean test scores (MeanDifference; line 84) and its standard error (Std.ErrorDifference; line 85) are provided. The number of degrees of freedom in a two-sample *t*-test equals *N*-2 = 433-2 = 431 (line 86). The output shows that the difference in mean test scores equals -1.201 with standard error 0.589. The corresponding *T* statistic equals *T* = -2.039, *df* = 390.652, *p* = .042, indicating a significant difference between aggressive and non-aggressive respondents.

[3] For other analyses, other OMS options may have to be specified, which can be found under task bar: Utilities, OMS Control Panel.

REFERENCES

Allison, P. D. (2002). *Missing data.* Thousand Oaks, CA: Sage.

Andersen, P. K., Borgan, Ø., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes.* New York: Springer.

Arbuckle J. L., & Wothke, W. (2006). AMOS 6.0 [Computer software]. Chicago: Smallwaters.

Azen, S., & Van Guilder, M. (1981). Conclusions regarding algorithms for handling incomplete data. *1981 Proceedings of the Statistical Computing Section. American Statistical Association,* 53-56.

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.

Boyd-Wilson, B. M., Walkey, F. H., McClure, J., Green, D. E. (2000). Do we need positive illusions to carry out plans? Illusion - and instrumental coping. *Personality and Individual Differences, 29,* 1141-1152.

Bracken, B. A., & Howell, K. (2004). *Clinical Assessment of Depression: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Burton, A., & Altman, D.G. (2004). Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *British Journal of Cancer, 91,* 4-8.

Cole, D. A., Hoffman, K., Tram, J. M., & Marwell, S. E. (2000). Structural differences in parent and child reports of children's symptoms of depression and anxiety. *Psychological Assessment, 12,* 174-184.

Cronbach, J. L. (1951). Coefficient alpha and the internal structure of tests.

Psychometrika, 16, 297-334.

De Leeuw, E. D., & Hox, J. J. (1988). Response stimulating factors in mail surveys. *Journal of Official Statistics, 4,* 241-249.

Dillman, D. A. (1991). The design and administration of mail surveys. *Annual Review of Sociology, 17,* 225-249.

Eggen, T. J. H. M., & Verhelst, N. D. (1992). *Item calibration in incomplete testing designs.* (Measurements and Research Department Reports 92-3). Arnhem, The Netherlands: Cito.

Fleming, T. R., & Harrington, D. P. (1991). *Counting processes and survival analysis.* New York: Wiley.

Foa, E. B., Kozak, M. J., Salkovskis, P. M., Coles, M. E., & Amir, N. (1998). The validation of a new Obsessive-Compulsive Disorder scale: The Obsessive Compulsive Inventory. *Psychological Assessment*, *10*, 206–214.

Gough, H. G., Heilbrun, A. B. (1980). *The Adjective Check List, manual 1980 edition.* Consulting Psychologists Press, Palo Alto, CA: Consulting Psychologists Press.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys.* New York: Wiley.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society, Series B,* 67-81.

Hare, R. D. (2003). *Manual for the Revised Psychopathy Checklist* (2nd ed.). Toronto, Ontario, Canada: Multi-Health Systems.

Hishinuma, E. S., Andrade, N. N., Johnson, R. C., McArdle, J. J., Miyamoto, R. H., Nahulu, L. B., Makini Jr, G. K., Yuen, N. Y. C., Nishimura, S. T., McDermott

Jr, J. F., Waldron, J. A., Luke, K. N., & Yates, A. (2000). Psychometric properties of the Hawaiian culture scale - adolescent version. *Psychological Assessment, 12,* 140-157.

Huisman, M., Krol, B., & Van Sonderen, F.L.P. (1998). Handling missing data by re-approaching nonrespondents. *Quality & Quantity*, *32*, 77-91.

Iversen, H., & Rundmo, T. (2002). Personality, risky driving and accident involvement among Norwegian drivers. *Personality and Individual Differences, 33,* 1251-1263.

Jacobusse G. W. (2005). WinMICE V1.0 The WinMICE application, a standalone software tool for multiple imputation when data have a multilevel structure [Computer software]. Retrieved September 3, 2008 from http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm

Kim, J. O., & Curry, J. (1977). The treatment of missing data in multivariate analysis, *Sociological Methods and Research, 6,* 215-240.

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

Little, R. J. A., & Su, H. L. (1989). Item nonresponse in panel surveys, In D. Kasprzyk, G. Duncan, & M. P. Singh (Eds.), *Panel surveys* (pp. 400-425). New York: Wiley.

Mislevy, R. J., & Wu, P. K. (1988). *Inferring examinee ability when some item responses are missing.* (Research Report RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

Morin, C. M., Landreville, P., Colecchi, C., McDonald, K., Stone, J. & Ling, W. (1999). The Beck anxiety inventory: psychometric properties with older adults. *Journal of Clinical Geropsychology, 5,* 19-29.

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.

Saggino, A., Kline, P. (1995). Item factor analysis of the Italian version of the Myers-Briggs Type Indicator. *Personality and Individual Differences, 19,* 243-249.

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, September 2, 2008, from http://www.stat.psu.edu/~jls/misoftwa.html.

Segal, D. L., Coolidge, F. L., Cahill, B. S., O'Riley, A. A. (2008). Psychometric properties of the Beck Depression Inventory—II (BDI-II) among community-dwelling older adults. *Behavior Modification, 32,* 3-20.

Sheviin, M., & Adamson, G. (2005). Alternative factor models and factorial invariance of the GHQ-12: A large sample analysis using confirmatory factor analysis. *Psychological Assessment, 17,* 231-236.

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.

S-Plus 8 for Windows [Computer software]. (2007). Seattle, WA: Insightful Corporation.

SPSS Inc. (2008). SPSS 16.0 for Windows [Computer software]. Chicago: author.

StataCorp. (2007). Stata Statistical Software: Release 10 [Computer software]. College Station, TX: StataCorp LP.

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

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*, 1-19.

Van Ginkel, J. R. (2006). MI.sps and MI-mul.sps [Computer code]. Retrieved September 3, 2008, from http://www.uvt.nl/mto/software2.html

Van Ginkel, J. R., & Van der Ark, L. A. (2005a). 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. (2005b). TW.SPS and RUNTW.SPS. [Computer code]. Retrieved September 3, 2008, from http://www.uvt.nl/mto/software2.html

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

Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2007b). Multiple imputation for item scores when test data are factorially complex. *British Journal of Mathematical and Statistical Psychology, 60,* 315-337.

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.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (in press). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology.*

Vorst, H.C.M. (1992). [Responses to the Adjective Checklist] Unpublished raw data.

Watson, D., O'Hara, M. W., Simms, L. J., Kotov, R., Chmielewski, M., McDade-Montez, E. A., Gamez, W., Stuart, S. (2007). Development and Validation of the Inventory of Depression and Anxiety Symptoms (IDAS). *Psychological Assessment, 19,* 253-268.

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 September 3, 2007, from http://www.ats.ucla.edu/stat/sas/library/multipleimputation.pdf

Table 1. *Overview of the 22 Subscales in the Adjective Checklist Data (Vorst, 1992) and Corresponding Item Numbers.*

Scale	Item No.	Scale	Item No.
Communality	1-10	Change	111-119
Achievement	11-20	Succorance	120-129
Dominance	21-30	Abasement	130-139
Endurance	31-40	Deference	140-149
Order	41-50	Personal Adjustment	151-159
Intelligence	51-60	Ideal Self	160-169
Nurturance	61-70	Critical Parent	170-179
Affiliation	71-80	Nurturant Parent	180-189
Exhibition	81-90	Adult	190-199
Autonomy	91-100	Free Child	200-209
Agression	101-110	Adapted Child	210-218

Table 2. *Listwise Deletion Results of Statistical Analyses of the ACL data (Vorst, 1992) (First Row), and With 5% of the Item Scores Removed According to Either MCAR (Second Row), MAR (Third Row), or NMAR (Fourth Row).*

Data	Alpha	Mean test score	Min. test score	Max. test score	**Mean diff.**	*t*	df	*p*
original	0.807	24.3764	5	40	**-1.298**	-2.261	431	.024
MCAR	0.802	24.5271	5	40	**-0.740**	-0.994	256	.321
MAR	0.810	24.2943	5	38	**-1.180**	-0.768	263	.114
NMAR	0.818	23.4841	5	38	**-0.972**	-1.254	250	.211

Table 3. *Frequency of Occurence of Missing Data in 24 Issues of Psychological Assessment, Personality and Individual Differences, and Journal of Personality Assessment.*

Journal	Vol.	Type of Nonresponse						
		UN	AT	IN	PL	Not clear	None reported	Total
Psychol. Assessment	1995	2	5	14	1	1	21	44
	1997	8	7	17	3	2	25	62
	2000	3	4	17	1	1	12	38
	2002	12	3	18	0	0	9	42
	2005	1	8	13	0	1	18	41
	2007	11	8	22	0	1	9	51
	Total	37	35	101	5	6	94	278
Pers. Indiv. Differ.	1995	3	4	14	0	0	41	62
	1997	9	3	15	0	1	45	73
	2000	4	1	13	0	2	41	61
	2002	10	6	16	0	1	27	60
	2005	7	3	17	0	3	52	82
	2007	10	2	26	0	1	51	90
	Total	43	19	101	0	8	257	428
J. Pers. Assess.	1995	5	3	19	0	1	25	53
	1997	0	1	8	0	2	27	38
	2000	6	2	7	0	2	14	31
	2002	5	1	14	0	0	15	35
	2005	4	5	13	1	1	11	35
	2007	2	6	11	0	0	10	29
	Total	22	18	72	1	6	102	221
Total	1995	10	12	47	1	2	87	159
	1997	17	11	40	3	5	97	173
	2000	13	7	37	1	5	67	130
	2002	27	10	48	0	1	51	137
	2005	12	16	43	1	5	81	158
	2007	23	16	59	0	2	70	170
	Total	102	72	274	6	20	453	927

Note. UN = unit nonresponse, AT = attrition, IN = item nonreponse, PL = planned missingness.

Table 4. *Statistics of the Types of Nonresponse Encountered in 24 Issues of Psychological Assessment, Personality and Individual Differences, and Journal of Personality Assessment. For the Studies That Reported Missing Values the Mean (M), Standard Deviation (SD), Minimum, and Maximum Number of Incomplete Response Patterns Are Reported.*

Type of nonresponse	*N*	*M*	*SD*	Skewness	Minimum	Maximum
UN	99	0.302	0.219	0.599	0.005	0.856
AT	74	0.186	0.136	1.090	0.016	0.703
IN	186	0.092	0.110	1.970	0.001	0.650
Not clear	10	0.385	0.315	0.326	0.040	0.898

Note. N = Number of cases where the type of nonresponse was reported. UN = unit nonresponse, AT = attrition, IN = item nonreponse.

Table 5. *Example of a Data Set With Incomplete Item Scores (Sijtsma & Van der Ark, 2003).*

Case	X_1	X_2	X_3	X_4	X_5
1	2	1	1	.	.
2	3	5	4	5	5
3	4	3	.	3	4
4	1	1	1	3	2
5	.	3	3	.	4
6	5	5	3	.	5
7	1	3	2	2	2
8	3	3	1	2	.

Table 6. *Example of Deterministic and Stochastic Variable Mean Imputation (Left), and Deterministic and Stochastic Regression Imputation (Right), in the Data Example from Sijtsma and Van der Ark (2003).*

Determistic Variable Mean Imputation

Case	X_1	X_2	X_3	X_4	X_5
1	2	1	1	**3**	**3.67**
2	3	5	4	5	5
3	4	3	**2.14**	3	4
4	1	1	1	3	2
5	**2.71**	3	3	**3**	4
6	5	5	3	**3**	5
7	1	3	2	2	2
8	3	3	1	2	**3.67**
Mean	2.71	3	2.14	3	3.67

Deterministic Regression Imputation

Case	X_1	X_2	X_3	X_4	X_5
1	2	1	1	**3**	**2.47**
2	3	5	4	5	5
3	4	3	**2.42**	3	4
4	1	1	1	3	2
5	**2.71**	3	3	**3.28**	4
6	5	5	3	**4.13**	5
7	1	3	2	2	2
8	3	3	1	2	**2.61**

Stochastic Variable Mean Imputation

Case	X_1	X_2	X_3	X_4	X_5
1	2	1	1	**0.72**	**1.38**
2	3	5	4	5	5
3	4	3	**2.28**	3	4
4	1	1	1	3	2
5	**4.52**	3	3	**2.71**	4
6	5	5	3	**0.71**	5
7	1	3	2	2	2
8	3	3	1	2	**3.59**
Mean	2.71	3	2.14	3	3.67
SD	1.50	1.51	1.21	1.22	1.37

Stochastic Regression Imputation

Case	X_1	X_2	X_3	X_4	X_5
1	2	1	1	**2.97**	**2.93**
2	3	5	4	5	5
3	4	3	**3.28**	3	4
4	1	1	1	3	2
5	**0.83**	3	3	**2.62**	4
6	5	5	3	**3.93**	5
7	1	3	2	2	2
8	3	3	1	2	**2.55**

Table 7. *Example of Multiple Imputation Using NORM (Schafer, 1998) in the Data Example From Sijtsma and Van der Ark (2003).*

Imputed Data Set #1

Case	X_1	X_2	X_3	X_4	X_5
1	2	1	1	**3.72**	**1.93**
2	3	5	4	5	5
3	4	3	**2.18**	3	4
4	1	1	1	3	2
5	**5.81**	3	3	**4.17**	4
6	5	5	3	**5.89**	5
7	1	3	2	2	2
8	3	3	1	2	**1.69**

Imputed Data Set #2

Case	X_1	X_2	X_3	X_4	X_5
1	2	1	1	**3.29**	**3.73**
2	3	5	4	5	5
3	4	3	**2.47**	3	4
4	1	1	1	3	2
5	**-0.03**	3	3	**2.86**	4
6	5	5	3	**3.59**	5
7	1	3	2	2	2
8	3	3	1	2	**1.99**

Imputed Data Set #3

Case	X_1	X_2	X_3	X_4	X_5
1	2	1	1	**4.82**	**3.17**
2	3	5	4	5	5
3	4	3	**-0.18**	3	4
4	1	1	1	3	2
5	**1.97**	3	3	**5.74**	4
6	5	5	3	**4.43**	5
7	1	3	2	2	2
8	3	3	1	2	**2.52**

Imputed Data Set #4

Case	X_1	X_2	X_3	X_4	X_5
1	2	1	1	**2.01**	**2.17**
2	3	5	4	5	5
3	4	3	**1.87**	3	4
4	1	1	1	3	2
5	**2.40**	3	3	**5.08**	4
6	5	5	3	**3.60**	5
7	1	3	2	2	2
8	3	3	1	2	**4.29**

Imputed Data Set #5

Case	X_1	X_2	X_3	X_4	X_5
1	2	1	1	**0.72**	**1.38**
2	3	5	4	5	5
3	4	3	**2.28**	3	4
4	1	1	1	3	2
5	**4.52**	3	3	**2.71**	4
6	5	5	3	**0.71**	5
7	1	3	2	2	2
8	3	3	1	2	**3.59**

Table 8. *Frequencies in Which Missing-Data Methods were Used in Studies from 24 Issues of Psychological Assessment, Personality and Individual Differences, and Journal of Personality Assessment.*

Missing-data method	Type of Nonresponse						
	UN	AT	IN	PL	Not clear	None reported	Total
LD	91	44	164	1	14	0	314
LD-CM	10	8	13	0	0	0	31
LD-CM-R	6	11	8	0	1	0	26
AC	1	11	64	1	4	0	81
AC-CM	0	4	1	0	0	0	5
IMP	0	0	18	1	0	0	19
DMLE	0	1	9	1	1	0	12
VD	1	0	36	0	0	0	37
FU	0	1	2	0	0	0	3
PRO	0	0	10	0	0	0	10
FU-LD-CM	1	1	0	0	0	0	2
Other	0	1	0	0	0	0	1
None reported	0	1	26	2	2	453	484
Total	110	83	351	6	22	453	1025

Note. UN = unit nonresponse, AT = attrition, IN = item nonreponse, PL = planned missingness. LD = listwise deletion. LD-CM = Listwise deletion with check for MCAR, MCAR not rejected. LD-CM-R = Listwise deletion with check for MCAR, MCAR rejected. AC = Available case analysis. AC-CM = Available-case analysis with check for MCAR, MCAR not rejected. IMP = Imputation. DMLE = Direct maximum likelihood estimation. VD = Variable deletion. FU = Follow-up PRO = Pro rating. FU-LD-CM = Combination of Follow-up and Listwise deletion with check for MCAR.

Table 9. *Example of Deterministic TW Imputation in the Data Example from Sijtsma and Van der Ark (2003).*

Case	X_1	X_2	X_3	X_4	X_5	PM_i
1	2	1	1	**1.45**	**2.12**	1.33
2	3	5	4	5	5	4.4
3	4	3	**2.76**	3	4	3.5
4	1	1	1	3	2	1.6
5	**3.17**	3	3	**3.45**	4	3.33
6	5	5	3	**4.62**	5	4.5
7	1	3	2	2	2	2
8	3	3	1	2	**3.04**	2.25
IM_j	2.71	3	2.14	3	3.67	OM = 2.88

Table 10. *Results of Statistical Analyses of the ACL data (Vorst, 1992) Without Missing Data (First Row), and With 5% of the Item Scores Removed According to Either MCAR (Second Row), MAR (Third Row), or NMAR (Fourth Row). Missing Data are Imputed Using Method TW.*

Data	alpha	Mean test score	Min. test score	Max. test score	Mean diff.	*t*	df	*p*
original	0.807	24.3764	5.00	40.00	-1.298	-2.261	431	.024
MCAR	0.811	24.3982	5.00	40.00	-1.196	-2.039	391	.042
MAR	0.810	24.3473	5.00	39.80	-1.307	-2.136	410	.033
NMAR	0.810	24.1621	5.00	40.00	-1.328	-2.264	402	.024

```
GET  FILE='C:\imputation\example.sav'.

DATASET DECLARE deterministic.
MVA
VARIABLES = X1 X2 X3 X4 X5
/EM ( TOLERANCE=0.001 CONVERGENCE=0.0001 ITERATIONS=25 )
/REGRESSION ( TOLERANCE=0.001 FLIMIT=4.0 ADDTYPE= NONE OUTFILE=stochastic )
.

GET  FILE='C:\imputation\example.sav'.

SET SEED = 2 .
DATASET DECLARE stochastic.
MVA
VARIABLES = X1 X2 X3 X4 X5
/EM ( TOLERANCE=0.001 CONVERGENCE=0.0001 ITERATIONS=25 )
/REGRESSION ( TOLERANCE=0.001 FLIMIT=4.0 ADDTYPE=RESIDUAL
OUTFILE=stochastic ) .
DESCRIPTIVES
VARIABLES=X1 X2 X3 X4 X5
/STATISTICS=MEAN STDDEV .
```

Figure 1: *SPSS Syntax for Applying Both Deterministic and Stochastic Regression Imputation in the Example Data Set from Sijtsma and Van der Ark (2003).*