Published in final edited form as:
Multivariate Behav Res. 2014 January 1; 49(1): 78–91. doi:10.1080/00273171.2013.855890.

Analysis of Variance of Multiply Imputed Data

Joost R. van Ginkel and **Pieter M. Kroonenberg**
Leiden University

Abstract

As a procedure for handling missing data, Multiple imputation consists of estimating the missing data multiple times to create several complete versions of an incomplete data set. All these data sets are analyzed by the same statistical procedure, and the results are pooled for interpretation. So far, no explicit rules for pooling *F*-tests of (repeated-measures) analysis of variance have been defined. In this paper we outline the appropriate procedure for the results of analysis of variance for multiply imputed data sets. It involves both reformulation of the ANOVA model as a regression model using effect coding of the predictors and applying already existing combination rules for regression models. The proposed procedure is illustrated using three example data sets. The pooled results of these three examples provide plausible *F*- and *p*-values.

Introduction

Missing data are ubiquitous in applied studies. Their presence needs to be addressed in research whenever differences between groups are the central focus of attention. Multiple imputation (Rubin, 1987, p. 2) is one of the preferred techniques for handling missing data but little seems to have been done with this technique in the analysis-of-variance context. In this paper we present a procedure for addressing this problem for both standard analysis of variance and its repeated-measures variants. Combination rules for other designs follow directly but not necessarily trivially, from the cases presented here.

Designs

Analysis of variance (ANOVA) is one of the most widely used statistical techniques where differences in means between groups are the primary focus. It is mainly used in experimental research based on factorial designs, but it is also regularly employed in non-experimental studies. The technique is especially geared towards testing statistical relationships between categorical independent or predictor variables on the one hand and a numerical dependent or response variable on the other hand.

When the same subjects are measured repeatedly at different time points and (possibly) under different conditions on the same predictor variables the technique is referred to as repeated-measures analysis of variance. Predictor variables that vary across persons are called *between-subjects* factors, while the repeated measurement of the response variable is evaluated as a *within-subjects* factor.

Corresponding author: Joost R. Van Ginkel, Child and Family Studies, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands, jginkel@fsw.leidenuniv.nl.

Handling Missing Data: Listwise Deletion

As any other statistical technique, the results from (repeated-measures) ANOVA are influenced by the presence of missing data. The simplest way to deal with this problem is to delete all cases from the analysis with at least one missing value. However, besides a loss of valuable information and power due to the reduction of the sample size, results may also be biased if this loss of subjects is not random but systematic. More specifically, when deleting incomplete cases, the missing values have to be *missing completely at random* (MCAR; Little & Rubin, 2002, p. 10) in order to obtain valid results.

Handling Missing Data: Multiple Imputation

Multiple imputation (Rubin, 1987) is an alternative missing-data procedure, which has become increasingly popular. The technique consists of substituting *M* plausible random values for each missing value so as to create *M* plausible complete versions of the incomplete data set. The *M* complete data sets are then analyzed by the statistical analysis of interest, and the results of these *M* analyses are pooled into one analysis, in which the additional uncertainty due to the missing data is incorporated.

Multiple imputation is implemented in several software packages such as Stata 10.0 (ICE; StataCorp, 2007), the MICE library in S-Plus (2007), SPSS 19.0 (SPSS, 2010), SAS 9.3 (2011) in the procedure PROC MI (Yuan, 2000), NORM (Schafer, 1998), Amos 7.0 (Arbuckle & Wothke, 2010), the missing-data library in S-Plus and several packages for R (Su, Gelman, Hill, & Yajima, 2011; Van Buuren & Groothuis-Oudshoorn, 2011).

Comparison Procedures

Multiple imputation has several advantages over Listwise deletion. Firstly, unlike Listwise deletion, Multiple imputation uses all available data and does not throw away any information. Secondly, Multiple imputation makes less stringent assumptions about the missingness mechanism. While Listwise deletion requires the data to be MCAR in order to obtain valid inferences, Multiple imputation will also lead to valid results if the missing data are *missing at random* (MAR; Little & Rubin, 2002, p. 10; Rubin, 1976). In this case the missing data are considered to have occurred at random conditional on one or more observed variables. For example, if people with high incomes tend to leave more questions open on the other variables than people with low incomes, people with the same income have the same probability of missing data on the other variables, and income is observed for all respondents, then the missingness for these questions is said to be missing at random (MAR), provided that income is included in the imputation model

Combination Rules for (Repeated-Measures) ANOVA

A problem of Multiple imputation in the context of analysis of variance is that to our knowledge the rules for pooling the significance tests of the *M* analyses of the completed data sets have never been explicitly discussed in the literature. Even SPSS 19.0, which performs Multiple imputation and pools results from multiply imputed data sets for several statistical techniques, does not provide pooled *F*-tests for any type of analysis of variance.

A possible reason for this is that Multiple imputation may not often be considered necessary in ANOVA. ANOVA is often used in experimental settings where a researcher has control over the situation, so that missing data will not occur frequently. In a repeated-measures ANOVA design missing data may be more common due to attrition but in this context researchers usually handle the missing data using multilevel with full information maximum likelihood (e.g., Hox, 2002; Snijders & Bosker, 1999). The necessity of Multiple imputation in an ANOVA context may thus not seem obvious at first.

However, ANOVA is used in non-experimental research as well. It frequently happens that researchers are interested in mean differences among the different sexes, different SES's, or different ethnicities. Moreover, a disadvantage of full information maximum likelihood is that for handling the missing data it normally cannot include variables outside the multilevel model. An exception to this is full information maximum likelihood in a structural equation modeling context, in which auxiliary variables may be used as saturated correlates (see, Graham, 2003). However, structural equation modeling is not the most obvious choice for carrying out ANOVA.

A consequence of this limitation of full information maximum likelihood is that it will only be guaranteed to provide unbiased results if the missingness depends on variables within the multilevel model. If the missingness depends on observed variables outside the multilevel model, the MAR assumption is violated for the specific analysis and valid results can no longer be guaranteed. Multiple imputation on the other hand can include auxiliary variables that are not part of the analysis, in the imputation model. Therefore it is also guaranteed to give unbiased results if the missingness depends on observed variables outside the multilevel model.

Besides non-experimental settings, Multiple imputation may also be useful in experimental settings, although its usefulness may not seem obvious at first. Unbalanced designs in which groups have unequal sample sizes, can be conceived of as a missing-data problem (e.g. Schafer, 1997, p. 21). For example, if a control group in an experiment has a sample size of $n_1 = 15$ and the experimental group has a sample size of $n_2 = 12$, this may also be conceived of as a balanced design in which 3 subjects in the experimental group have missing data on the dependent variable. One could decide to ignore the unbalancedness and carry out the ANOVA anyway. However, variations in the dependent variable due to overall main effects and interactions are only additive in balanced designs (Winer, 1971, pp. 402–404). In unbalanced designs additivity is lost. As a result, F-tests become less robust to unequal variances and loose power.

In short it can be argued that there is clearly a need for combination techniques for (repeated-measures) ANOVA of multiply imputed data sets. Although these rules have not explicitly been defined for ANOVA as of yet, it is straightforward to extend the already existing pooling rules to the F-tests of (repeated measures) analysis of variance, as will be shown in this paper. An earlier brief discussion of this idea was presented in an unpublished conference presentation (Van Ginkel & Kroonenberg, 2011).

It should be noted that both SAS 9.3 (MIAnalyze; Yuan, 2000, see the manual[1], pp. 4676–4678) and Stata 10.0 have the tools needed for carrying out our proposed procedure, but neither are their steps fully automated, nor do the complete procedures follow directly from the manuals or the existing literature. For example, the SAS manual does not contain the words “analysis of variance” or “ANOVA”. Thus, a SAS user would have to figure out the practicalities made explicit in this paper by him- or herself.

Once realized what is required the procedure itself is conceptually fairly straightforward. However, to figure out which steps are needed to get to the correct procedure intimate knowledge of Multiple imputation and effect coding is required, something which the average user of ANOVA does not have. One of the purposes of this paper is to explicitly supply guidelines for carrying out ANOVA on multiply imputed data sets.

In the next section we will discuss existing rules for combining the results of multiply imputed data sets. Then we will discuss how these rules can be applied in the context of both ANOVA and repeated-measures ANOVA. Finally we provide three empirical applications and indicate which conclusions may be drawn about the proposals.

Rules for Pooling Significance Tests of Multiply Imputed Data Sets

In this section we will first outline the theory behind pooling significance tests from multiply imputed datasets for single parameters followed by those for several parameters simultaneously. With this as a basis we will formulate the rules for analysis-of-variance designs in the next section.

Single Parameter Estimates

After constructing M completed versions of the incomplete data set, the same statistical analysis is applied to each of the imputed data sets. These M analyses are combined into one pooled result so that the uncertainty due to the missing data can be taken into account. When the analysis involves a single parameter estimate, such as a single regression coefficient or the difference between two sample means, the following pooling rules for M imputed data sets apply (Rubin, 1987): First, define $\hat{Q}$ as the parameter estimate of the parameter Q that would have been obtained if no data were missing, and $\sqrt{U}$ as its standard error. Each imputed data set m (m = 1,…, M) has an estimate of $\hat{Q}$, denoted $\hat{Q}_m$, and a standard error $\sqrt{U}_m$. An overall estimate of Q based on the M imputed data sets is simply the mean of the M estimates.

$$\overline{Q}=\frac{1}{M}\sum_{m=1}^{M}\hat{Q}_m. \quad (1)$$

The estimate of the overall variance T of $\overline{Q}$ consists of two parts, namely $\overline{U}$the *within-imputation variability* i.e. the mean of the squared standard errors within the imputed data sets

[1] http://support.sas.com/documentation/onlinedoc/stat/930/mianalyze.pdf

$$\overline{U}=\frac{1}{M}\sum_{m=1}^{M}U_m, \quad (2)$$

and *B* the *between-imputation variability* caused by the differences in imputed values across the data sets

$$B=\frac{1}{M-1}\sum_{m=1}^{M}(\hat{Q}_m-\overline{Q})^2. \quad (3)$$

The overall variance *T* associated with $\overline{Q}$ then becomes

$$T=\overline{U}+(1+M^{-1})B, \quad (4)$$

where the coefficient $(1 + M^{-1})B$ serves to correct for the extra variability due to the missing data. To test the null hypothesis that a parameter is equal to a specific value, i.e. $Q = Q_0$, the following statistic can be used:

$$t_\nu=\frac{\overline{Q}-Q_0}{\sqrt{T}}, \quad (5)$$

which has a *t*-distribution with ν degrees of freedom, or equivalently

$$F_{1,\nu}=\frac{(\overline{Q}-Q_0)^2}{T}, \quad (6)$$

can be used, which has an *F*-distribution with 1 degree of freedom in the numerator, and ν degrees of freedom in the denominator.

For ν we cannot use the same number of degrees of freedom as when no data are missing, because there is extra uncertainty due to the missing data. Thus, ν has to be adjusted downwards. Two approximations for ν exist. The first approximation, denoted $\nu^{\#}$, is given by

$$\nu^{\#}=(M-1)\left[1+\frac{\overline{U}}{(1+M^{-1})B}\right]^2 \quad (7)$$

(Rubin, 1987, p. 77). This is the approximation that is for instance used in the Multiple-imputation procedures in SPSS 19.0 (SPSS, 2010). Note that this approximation is totally independent of the sample size since it is not included in the formula. This is because the approximation was derived under the assumption that if no data were missing, statistical inferences would be based on statistical tests with infinite number of error degrees of freedom. More specifically, for the complete-data case, the statistical test should be a *z*-test rather than a *t*-test (see Barnard & Rubin, 1999). To correct for the sample size, Barnard and Rubin (1999) proposed an alternative approximation, which is implemented in SAS 9.3

(2011). Define ν_{com} as the number of degrees of freedom that would have been obtained if no data were missing. Then the approximation ν^* is given by

$$\nu^*=\left(\frac{1}{\nu\#}+\frac{1}{\nu_{obs}}\right)^{-1}\text{with } \nu_{obs}=\left[1-\frac{(1+M^{-1})B}{T}\right]\nu^*_{com}\text{and } \nu^*_{\text{com}}=\left(\frac{\nu_{com}+1}{\nu_{com}+3}\right)\nu_{com}. \quad (8)$$

This approximation is always smaller than ν_{com} and for an infinite sample size it reduces to Equation 7. For the complete derivation of this approximation we refer to Barnard & Rubin (1999).

Multiparameter Estimates—When more than one parameter is tested for significance simultaneously, we need a pooled statistic that simultaneously tests several parameters. In the context of regression, Rubin (1987, pp. 79–81) defined extensions of Equations 1–7 for multiparameter estimates. Suppose that $\hat{\mathbf{Q}}$ is a $p \times 1$ vector of parameter estimates of the parameter vector **Q** that would have been obtained if no data were missing, and **U** is a $p \times p$ covariance matrix of the parameter estimates. For a regression model this covariance matrix can be computed as follows: Let **X** be an $N \times p$ matrix containing the predictor variables, let s^2_ε be the error variance. The covariance matrix of the dependent variable **Y** is defined as

$$\mathbf{V}=s^2_\varepsilon\mathbf{I}_N. \quad (9)$$

The covariance matrix of the parameter estimate **Q** is

$$\mathbf{U}=(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \quad (10)$$

(See also, Rubin, 1987, p. 79). Each imputed data set m (m = 1,…, M) has an estimate for the parameters $\hat{\mathbf{Q}}$, denoted $\hat{\mathbf{Q}}_m$, and a covariance matrix $\hat{\mathbf{U}}_m$. An overall estimate for the parameter vector **Q** based on the parameter estimates from the M imputed data sets is (analogous to Equation 1)

$$\overline{\mathbf{Q}}=\frac{1}{M}\sum_{m=1}^{M}\hat{\mathbf{Q}}_m. \quad (11)$$

As for single parameter estimates, the overall covariance matrix of $\overline{\mathbf{Q}}$ consists of two parts, namely the within-imputation variance$\overline{\mathbf{U}}$

$$\overline{\mathbf{U}}=\frac{1}{M}\sum_{m=1}^{M}\hat{\mathbf{U}}_m. \quad (12)$$

(analogous to Equation 2) and the between-imputation variance **B**

$$\mathbf{B}=\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\mathbf{Q}}_m-\overline{\mathbf{Q}})(\hat{\mathbf{Q}}_m-\overline{\mathbf{Q}})'. \quad (13)$$

(analogous to Equation 3). The best multivariate estimate of the total covariance matrix, to date, is given by (analogously to Equation 4)

$$\mathbf{T}=(1+r)\overline{\mathbf{U}}, \text{with } r=(1+M^{-1})tr(\mathbf{B}\overline{\mathbf{U}}^{-1})/p, \quad (14)$$

where *r* denotes the average relative increase in variance due to nonresponse across the parameters of **Q** (see Rubin, 1987). More specifically, using some algebra it can be shown that *r* is the ratio of the increase in variance due to the missing data, defined as $\mathbf{T}-\overline{\mathbf{U}}$, to the average variance $\overline{\mathbf{U}}$. Analogously to the univariate case the null hypothesis of differences between the pooled parameter estimate and its population value can be tested by an *F*-statistic

$$F=(\overline{\mathbf{Q}}-\mathbf{Q}_0)\mathbf{T}^{-1}(\overline{\mathbf{Q}}-\mathbf{Q}_0)'/p \quad (15)$$

which has an approximate *F*-distribution with *p* and ν degrees of freedom. As for the univariate case, two approximations for the degrees of freedom for the denominator ν exist. Suppose that $q = p(M-1)$ and we are dealing with a large sample then the degrees of freedom $\nu_1^{\#}$ is approximated by

$$\nu_1^{\#}=4+(q-4)[1+(1-2q^{-1})r^{-1}]^2 \quad (16)$$

for $q > 4$, and

$$\nu_1^{\#}=q(1+p^{-1})(1+r^{-1})/2$$

for $q \le 4$ (Li, Raghunathan, & Rubin, 1991). Like Equation 7, Equation 16 was derived under the assumption that if no data were missing, the error degrees of freedom would be set at infinity. To correct for sample size, Reiter (2007) proposed a more accurate approximation, based on second-order Taylor series expansions. Since such approximations contain several terms, the complete formula is quite long and complex. Using ν_{com}^* as computed in Equation 8, the complete formula is given by:

$$\nu_1^*=4+1/z\text{with}$$

$$z=\frac{1}{\nu_{com}^*-4(1+a)}+\frac{1}{q-4}\left\{\frac{a^2[\nu_{com}^*-2(1+a)]}{(q+a)^2\{\nu_{com}^*-4(1+a)\}}\right\}+\frac{1}{q-4}\left\{\frac{8a^2[\nu_{com}^*-2(1+a)]}{(1+a)\{\nu_{com}^*-4(1+a)\}^2}+\frac{4a^2}{(1+a)\{\nu_{com}^*-4(1+a)\}}\right\}+ \quad (17$$

with

$$a=rq/(q-2).$$

For the derivation of this formula we refer to Reiter's paper (2007).

Although the distribution of *F* in Equation 15 was derived under the strong assumption that the fractions of missing information are equal across all *p* parameters in $\bar{\mathbf{Q}}$, a simulation study by Li, Raghunathan, and Rubin (1991) showed that this procedure is robust to violation of this assumption. As we will show below the above development can be used in analysis of variance by transforming such models into regression ones.

Applying the Rules for Multiple Imputation to (Repeated Measures) ANOVA

The pooling techniques described above can also be used to obtain pooled *F*-tests for (repeated measures) ANOVA. However, as far as we have been able to trace this has not yet been presented in the literature. Rubin (1987) provided an example of regression analysis (pp. 79–81) but did not explicitly extend this to (repeated-measures) ANOVA. Moreover, we have not been able to find published ANOVA applications for multiply imputed data sets. Furthermore, implementation in statistical software packages is either limited (SPSS 19.0) or not well documented and not fully automated (SAS 9.3, Stata 10.0).

Applying the Combination Rules for Multiple Imputation to Analysis of Variance

Factorial Designs

In this section we will describe how Van Ginkel and Kroonenberg's (2011) proposals for using Rubin's (1987) pooling procedures for regression can be used for the analysis of variance of two-way factorial designs. The procedure readily extends to ANOVA for other factorial designs. Even though analysis of variance is nothing but a regression analysis with categorical predictors the regression rules cannot directly be applied here, as the standard way of handling categorical variables in regression is by dummy coding. The crux of using regression analysis for analysis of variance in general is the use of *effect coding* (e.g., Edwards, 1985, pp. 146–150) rather than dummy coding. This will ensure appropriate estimates for the effects. Since pooled *F*-tests from ANOVA in Multiple imputation can only be obtained by using the parameters and covariance matrices of a regression analysis, the use of effect coding is essential in this context.

The Two-Way ANOVA Model for Independent Measures—In a two-way design the value of the dependent variable of person *i* (*i* = 1,…, *N*) in group *j* (*j* = 1,…, *J*) of the first factor, and group *k* (*k* = 1,…, *K*) of the second factor is described by

$$Y_{ijk}=\mu+\alpha_j+\beta_k+\alpha\beta_{jk}+\varepsilon_{ijk}, \quad (18)$$

where μ is the overall mean, α_j is the deviation of group *j* from the first factor from the overall mean, β_k is the deviation of group *k* from the second factor from the overall mean, $\alpha\beta_{jk}$ is the interaction between factors 1 and 2, and ε_{ijk} is an individual error term which is assumed to be normally distributed with mean zero and variance σ^2.

The Two-Way ANOVA Model in Regression Format Using Dummy Coding—For dummy coding of the effects the X_{1ij}'s (*j* = 1,…, *J* - 1) are coded as 1 if respondent *i* belongs to group *j* of Factor 1, and 0 otherwise. The same goes for X_{2ik} (*k* = 1,…, *K* - 1) for Factor 2.

If a respondent has 0’s for all dummy variables of a specific factor the individual belongs to the reference group. The intercept is written as β_0, and β_{1j}, β_{2k}, and β_{3jk} are the regression coefficients of group *j* of factor 1, group *k* of factor 2, and the interaction of factors 1 and 2, respectively. The two-way ANOVA model can now be formulated as a regression model:

$$Y_{ijk}=\beta_0+\sum_{j=1}^{J-1}\beta_{1j}X_{1ij}+\sum_{k=1}^{K-1}\beta_{2k}X_{2ik}+\sum_{j=1}^{J-1}\sum_{k=1}^{K-1}\beta_{3jk}X_{1ij}X_{2ik}+\varepsilon_{ijk}. \quad (19)$$

However, in this form the intercept is no longer the overall mean and the effects do not have the same meaning as in the standard ANOVA parameterization. Instead, the intercept is the mean of the respondents who are in the references groups for both independent variables, and the effects are the deviations of each group from the mean of the reference group. Consequently, when this parameterization is used for combining the results of ANOVA from multiply imputed data sets, firstly the *F*-test of the intercept does not test whether the overall mean differs significantly from zero but whether the mean of the reference category differs from zero. Secondly, the *F*-tests of the effects does not test whether all means differ from the overall mean but whether all means except the mean of the reference category differ significantly from the mean of the reference category. This interpretation differs from the interpretation of the *F*-tests in standard ANOVA.

The Two-Way ANOVA Model in Regression Format Using Effect Coding—In order to give the regression coefficients the standard meaning in analysis of variance, one has to use effect coding (Edwards, 1985, pp. 146–150). In this type of coding, the X_{1ij}’s (j = 1,…, J - 1) are coded as 1 if respondent *i* belongs to group *j* of Factor 1, −1 if the individual belongs to the reference group, and 0 otherwise. The same coding system is used for factor 2. With this coding system, the intercept β_0 equals the overall mean μ, and each effect $\beta_{q\ell}$ of factor *q* is the difference between the mean of group *l* and the overall mean. The regression coefficient of the reference group is defined as $\beta_{qL}=-\sum_{\ell=1}^{L-1}\beta_{q\ell}$. The interaction effects are defined as the product of the effect coded factors.

Covariance of parameter estimates—For an appropriate evaluation of the analysis of variance and its tests of significance we also need the covariance matrix of the parameter estimates. For each effect in the ANOVA this can be obtained by substituting the standard ANOVA mean square error (*MSE*) for the error variance s_ε^2 in Equation 9, and using for **X** the effect coded design matrix of the specific effect.

Example: The NICHD data

The pooling procedure for effect-coded factors will be illustrated with an analysis of a small subset of the data from the National Institute of Child Health and Human Development (NICHD) Study on Early Child Care (1996). The total dataset consists of information from *N* = 1364 children and it contains a sizeable number of missing data. The descriptives for the selected categorical variables are given in Table 1 and those for the numerical variables in Table 2.

To illustrate the above procedure for handling missing data in an ANOVA context 5 (Mother's ethnicity) × 2 (Child's gender) analyses of variance with the Child's school readiness as the response variable were carried out on the multiply imputed data sets.

Because school readiness and ethnicity were not completely observed their missing data were imputed five times within SPSS using Predictive Mean Matching (Little, 1988; Rubin, 1986). Roughly, this method first predicts the missing values for a numerical variable j using a regression model with the other variables as predictors. Next, for each respondent with a missing value it finds a matching respondent having variable j observed, who closely matches the respondent with a missing value on the predicted value. The observed value of this matching respondent is then used for the imputation of the missing value. For categorical variables Predictive Mean Matching does not use this matching procedure but imputes random values from a (multinomial) logistic regression model. All variables listed in Tables 1 and 2 were included in the imputation model.

In principle, it is possible to use other Multiple-imputation methods, for example, by generating random values from a normal linear regression model, or if the data are not normally distributed, by using transformations of the data, or other distributions (see, van Buuren, 2012, for an overview). However, the comparison of different Multiple-imputation methods is not the focus of this paper.

Using effect coding, the results of the five imputed data sets were combined employing Rubin's (1987) rules for multiparameter estimates. Because the standard ANOVA procedures in SPSS do not provide covariance matrices for the parameters, the ANOVAs on the imputed data sets had to be carried out using the SPSS Mixed Models procedure. The pooling procedure was carried out using an SPSS macro by Van Ginkel (2010a). Detailed information about how to use this macro and the instructions for carrying out the entire analysis can be found in the corresponding manual (Van Ginkel, 2010b).

For predictor variables with only two levels (here, Child's Gender) the rules for single-parameter estimates must be used (Equations 1–4, and Equation 6); for predictor variables with more than two levels (here, Mother's Ethnicity) the rules for multiparameter estimates must be used (Equations 11–15). For the error degrees of freedom, the approximation from Reiter (2007) was employed. The complete syntax code for the analysis can be downloaded from the first author's website[2]. Table 3 shows the pooled ANOVA results of the five imputed data sets (second row of each panel), together with the results of Listwise deletion (first row of each panel). For the results of Multiple imputation, $\bar{\lambda}$, the average proportion of variation in **Q** attributable to missing data is shown as well (Van Buuren, 2012, p. 41). This measure gives an indication of the severity of the missing-data problem.

One noticeable result is that the pooled number of error degrees of freedom of the imputed data is always smaller than the number of error degrees of freedom of Listwise deletion (second row). With the exception of the F-test of the total model (Table 3, second row), this difference between the degrees of freedom of Listwise deletion and Multiple imputation is

[2] http://www.socialsciences.leiden.edu/educationandchildstudies/childandfamilystudies/organisation/staffcfs/van-ginkel.html

more severe for large values of $\bar{\lambda}$ (last column). However, by increasing the number of imputations, the number of error degrees of freedom of the imputed data may come closer to, or even exceed the number of error degrees of freedom of Listwise deletion, since the approximation of the number of degrees of freedom is also influenced by the number of imputations. Schafer (1997, p. 111) explained this phenomenon in the context of a single parameter estimate $\bar{Q}$ as follows: the variance of $\bar{Q}$ contains a term that equals the between-imputation variance B divided by M (represented by M^{-1} in Equation 14). As the number of imputations increases, this term becomes smaller, resulting in less uncertainty about the missing data. This is also reflected in the number of degrees of freedom (see Equations 7 and 8). The same holds for the approximation of the degrees of freedom for multiparameter estimates (Reiter, 2007). However, it should be noted that the number of degrees of freedom based on Reiter's approximation (see Equation 17) can never exceed the number of degrees of freedom of the hypothetical complete data (see Reiter, 2007, p. 502).

As a check we carried out the same analyses again, only now using 100 imputations (Table 3, each third row). The results confirm our conjecture: with 100 imputations, the error degrees of freedom are larger than those of Listwise deletion. Also, note that the effect of Gender was significant for 5 imputations, but not anymore for 100 imputations.

The *F*- and *p*-values of Multiple imputation and Listwise deletion do not differ much. However, it is not clear in this example to what extent the differences between the results of Multiple imputation and Listwise deletion are systematic or can be attributed to sampling fluctuation. However, a simulation study by Li, Raghunathan, and Rubin (1991) within a regression context shows that this pooling procedure provides accurate *F*- and *p*-values.

As an aside with respect to the substantive outcomes it should be noted that the only significant predictor of School readiness, Mother's ethnicity, is a variable which is strongly related to other variables such as education and income (see for instance Bakermans-Kranenburg, Van IJzendoorn, & Kroonenberg, 2004). However, sorting out the most complete analysis goes beyond the illustrative use of these data for ANOVA on multiply imputed data sets.

Repeated-Measures Designs

Because repeated-measures ANOVA has more than one error term in the model, the pooling procedure for the Multiple imputation is hardly trivial and more complicated than standard ANOVA. Suppose we have a two-way model with only within-subject factors. First, let $\boldsymbol{\pi}_i$ be the random effect of person i with mean 0 and variance σ_{π}^2. Second, let $(\boldsymbol{\alpha\pi})_{ij}$ be the interaction term for the j-th level of Factor 1 of person i with mean 0 and variance $\sigma_{\alpha\pi}^2$, let $(\boldsymbol{\beta\pi})_{ik}$ be the interaction term for the k-th level of Factor 2 of person i with mean 0 and variance $\sigma_{\beta\pi}^2$, and let $(\boldsymbol{\alpha\beta\pi})_{jki}$ be the three-way interaction term of the j-th level of Factor 1, the k-th level of Factor 2, and respondent i with variance $\sigma_{\alpha\beta\pi}^2$. Finally, let $\boldsymbol{\varepsilon}_{ijk}$ be an error term for the j-th level of Factor 1 and the k-th level of Factor 2 for person i with variance σ_{ε}^2. The two-way repeated-measures ANOVA model can then be written as

$$Y_{ijk}=\mu+\pi_i+\alpha_j+(\alpha\pi)_{ij}+\beta_k+(\beta\pi)_{ik}+(\alpha\beta)_{jk}+(\alpha\beta\pi)_{jki}+\varepsilon_{ijk}. \quad (20)$$

See also, Maxwell & Delaney (2004, pp. 574–577), who discuss the features of this model and calculations of the *F*-tests in more detail. To conform to Rubin's (1987) combination rules the repeated measures ANOVA model has to be rewritten as a regression model as well. However, since the factors are nested within each person *i* and each factor has its own error term, this can only be realized by rewriting it as a special case of a multilevel model where the repeated measures are nested within persons. First, let β_{0i} be a random intercept of person *i* with mean γ_{00} and variance $\sigma^2_{\eta_0}$. Next, let β_{1ij} be the random coefficient of person *i* for group *j* of Factor 1, with mean γ_{j0} and variance $\sigma^2_{\eta_j}$. Third, let β_{2ik} be the random coefficient of person *i* for group *k* of Factor 2, with mean γ_{0k} and variance $\sigma^2_{\eta_k}$. Finally, let β_{3jk} be the regression coefficient of the interaction effect for group *j* of Factor 1 and group *k* of Factor 2 with mean 0 and variance $\sigma^2_{\eta_{jk}}$, and let σ^2_{ε} be the error variance of respondent *i* at level *j* of Factor 1 and level *k* of Factor 2. The multilevel model that describes the two-way repeated measures ANOVA model can be written as

$$\begin{aligned}
Y_{ijk}=\beta_{0i}+\sum_{j=1}^{J-1}\beta_{1ij}X_{1ij}+\sum_{k=1}^{K-1}\beta_{2ik}X_{2ik}+\sum_{j=1}^{J-1}\sum_{k=1}^{K-1}\beta_{3ijk}X_{1ij}X_{2ik}+\varepsilon_{ijk}.\\
\beta_{0i}=\gamma_{00}+\eta_{00i}, \eta_{00i}\sim N(0,\sigma^2_{\eta_0})\\
\beta_{1ij}=\gamma_{j0}+\eta_{j0i}, \eta_{j0i}\sim N(0,\sigma^2_{\eta_j})\\
\beta_{2ik}=\gamma_{0k}+\eta_{0ki}, \eta_{0ki}\sim N(0,\sigma^2_{\eta_k})\\
\beta_{3ijk}=\gamma_{jk}+\eta_{jki}, \eta_{jki}\sim N(0,\sigma^2_{\eta_{jk}})\\
\varepsilon_{ijk}\sim N(0,\sigma^2_{\varepsilon}),
\end{aligned} \quad (21)$$

where η_0, η_{j0i}, η_{0ki}, η_{jki}, and ε_{ijk} are the individual random terms of the intercept, the effect of Factor 1, the effect of Factor 2, the interaction, and the error respectively. Like standard ANOVA, the independent variables X_{1ij} and X_{2ik} have to be effect coded.

The covariance matrix of the parameter estimates is defined as follows. First, suppose *H* is the total number of random effects in the model, s^2_h is the estimated variance of the *h*-th random effect, and *D* is the number of levels within the *h*-th random effect. Next, let $\mathbf{W}_{hi}$ be an (*NJK*) × *D* design matrix for subject *i* with entry 1 indicating the *d*-th random term for subject *i*, and 0 otherwise. Third, let $\mathbf{W}_h$ be an (*NJK*) × (*ND*) design matrix of the *h*-th random effect for all subjects which can be contructed as $\mathbf{W}_h$= [$\mathbf{W}_{h1}$,…,$\mathbf{W}_{hN}$]. Fourth, define $\mathbf{G}_h=s^2_h\mathbf{I}_{(ND)}$. The (*NJK*) × (*NJK*) covariance matrix **V** of variable **Y** can now be computed as

$$\mathbf{V}=\sum_{h=1}^{H}\mathbf{W}_h\mathbf{G}_h\mathbf{W}^{'}_h+\mathbf{\Sigma} \quad (22)$$

with $\mathbf{\Sigma}=s_{\varepsilon}^{2}\mathbf{I}_{(NJK)}$. Finally, for each effect the covariance matrix of its parameter estimates is computed as Equation 10, using for **X** the effect coded design matrix of the specific effect (see, for example, Giesbrecht and Burns, 1985, and Robinson, 1991). From hereon the calculations proceed in exactly the same way as in an ANOVA with independent measures.

Example: The McArdle-Nesselroade Developmental Change Data

The use of the combination rules for a two-way repeated measures ANOVA with two within-subjects factors will be illustrated using the data from the McArdle and Nesselroade (1994) study. Children (*N* = 32) were measured on 25 occasions on 7 tasks (Table 4). Because the imputation model would become overparameterized if 25 × 7 = 175 variables were to be included given there were only 32 children, it was decided for illustrative purposes to use only the measurements on weeks 1, 13, and 25 for both the imputation procedure and the subsequent repeated measures ANOVA; for the descriptives see Table 5. To impute the data for all time points without the imputation model becoming overparameterized, more respondents would be needed.

Table 5 shows that not all children provided data on all tasks and on each occasion, so that the missing data needed to be imputed. This was done five times using Predictive Mean Matching. Since all variables were numeric, the underlying Multiple-imputation model for each variable was a linear regression model with the other time points being the (numerical) predictors. A 3 (Weeks) × 7 (Tasks) repeated measures ANOVA was carried out on the multiply imputed data with the scores on the specific tasks as the dependent variables. Again, the SPSS Mixed Models procedure was used. Using effect coding for the predictors, the results of the five imputed data sets were combined using Rubin’s (1987) rules for multiparameter estimates.

The pooling procedure was again carried out using the SPSS macro by Van Ginkel (2010a; 2010b) using the error degrees of freedom approximation from Reiter (2007). The complete syntax for the analysis with 5 imputations is provided on the first author’s website. Table 6 shows the pooled results of the imputed data sets for both 5 imputations and 100 imputations (third and fourth row of each panel). For comparison, the results of Listwise deletion are also shown. Because in multilevel modeling it is possible to estimate the model using Full information maximum likelihood (FIML) without throwing away incomplete cases, results for Full information maximum likelihood are also shown (second row of each panel).

Since no auxiliary variables are used for Multiple imputation in this example, FIML uses the same amount of information for handling the missing data as Multiple imputation does. Consequently, Multiple imputation has no advantages over FIML here so that FIML can serve as a gold standard. It should be noted though that normally it is recommendable to include auxiliary variables in the imputation model. A data set without auxiliary variables has been used here to keep the results comparable with FIML.

It can be seen that the Multiple imputation procedure provides plausible *F*-values in the sense that their values are close to the *F*-values that result from the gold standard FIML. Unlike the previous example the numbers of error degrees of freedom for 5 imputations are not smaller than for Listwise deletion. Additionally the numbers of error degrees of freedom for 5 imputations hardly differs from the numbers of error degrees of freedom for 100 imputations. This difference between this example and the previous one could lie in the fact that on average the proportion of variation attributed to missing data (last column) is lower than in the previous example.

Mixed Designs

Finally we will look at an ANOVA design with both a between-subjects factor α_j and a within-subjects factor β_k. Let $\pi_{i(j)}$ be random term for the effect of person *i* with mean 0 and variance σ_π^2 in group *j*, and let ε_{ijk} be an error term for the *j*-th level of Factor 1 and the *k*-th level of Factor 2 for person *i* with variance σ_ε^2. The mixed two-way ANOVA model is written as

$$Y_{ijk}=\mu+\alpha_j+\pi_{i(j)}+\beta_k+(\alpha\beta)_{jk}+(\pi\beta)_{ik(j)}+\varepsilon_{ijk}, \quad (23)$$

where the *j* between brackets indicates that *i* is nested within group *j* (also, see Maxwell & Delaney, 2004, pp. 594–597). Rewriting this model as a multilevel model goes as follows: First, we define the value of a between-subjects factor for respondent *i* in group *k* (*k* = 1,…, *K*) as Z_{ik}, and the value of a within-subjects factor for respondent *i* in group *k* at time point *j* (*j* = 1,…, *J*) as X_{ijk}. Next, let β_{0i} be a random intercept of person *i* with mean $\gamma_{00}+\sum_{k=1}^{K}\gamma_{0k}Z_{ik}$ and variance $\sigma_{\eta_0}^2$, and let β_{1ij} be the random coefficient for measure *j* of Factor 1, with mean $\gamma_{j0}+\sum_{k=1}^{K}\gamma_{jk}Z_{ik}$ and variance $\sigma_{\eta_j}^2$. Finally, let σ_ε^2 be the error variance of respondent *i* in group *k* at time point *j*. The multilevel model that describes the mixed two-way ANOVA model can be written as

$$\begin{aligned}
&Y_{ijk}=\beta_{0i}+\sum_{j=1}^{J-1}\beta_{1ij}X_{ijk}+\varepsilon_{ijk}\\
&\beta_{0i}=\gamma_{00}+\sum_{k=1}^{K-1}\gamma_{0k}Z_{ik}+\eta_{0i},\eta_{0i}\sim N(0,\sigma_{\eta_0}^2)\\
&\beta_{1ij}=\gamma_{j0}+\sum_{k=1}^{K-1}\gamma_{jk}Z_{ik}+\eta_{ji},\eta_{ji}\sim N(0,\sigma_{\eta_j}^2)\\
&\varepsilon_{ijk}\sim N(0,\sigma_\varepsilon^2),
\end{aligned} \quad (24)$$

where η_{0i}, η_{ji}, and ε_{ijk} are the individual error terms of the intercept, the effect of the within-subjects factor, and of the lowest level. Again, the independent variables Z_{ij} and X_{ijk} have to be effect coded. For each effect the covariance matrix of its parameter estimates can be computed as in Equations 22 and 10, using for **X** the effect coded design matrix of the specific effect. The only difference with a completely within-subjects design is that here there is only one design matrix $\mathbf{W}_h$.

Example: Anorexia Data

In our last example we will analyze a data set collected from *N* = 71 young girls suffering from Anorexia (Hand, Daly, Lunn, McConway, & Ostrowski, 1994, *p*. 229). In this study three different groups received three different kinds of treatment, namely Cognitive behavioral treatment (n_1 = 29), Standard treatment (control group; n_2 = 26), and Family therapy (n_3 = 17). For each girl body weight was measured before and after treatment. Thus, there was one between-subjects factor (treatment) and one within-subjects factor (time point). The descriptives for each group are given in Table 7.

Although this example contains no missing data it may be conceived of as a missing-data problem because the design is unbalanced. By imagining three additional cases in the Control group and twelve in the Family therapy group with missing values on body weight at both time points, we get a balanced design with missing data on the dependent variable.

Like in the other examples, Multiple imputation was done using Predictive Mean Matching. For each time point the underlying Multiple-imputation model was a linear regression model, using group membership as a categorical predictor. A 3 (Treatment) × 2 (Time point) mixed ANOVA was carried out on the multiply imputed data with body weight as the dependent variable. Again, the SPSS Mixed Models procedure was used. Using effect coding of the predictors **X**, the results of the imputed data sets were combined using Rubin's (1987) rules for multiparameter estimates using the SPSS macro by Van Ginkel (2010a; 2010b). The complete syntax for 5 imputations may also be found on the first author's website. Table 8 shows the pooled results of the imputed data sets for both 5 imputations and 100 imputations (second and third row of each panel). The results of the unbalanced ANOVA (equivalent to Listwise deletion) are shown for comparison (first row).

In this example it is even more noticeable that the number of error degrees of freedom of 5 imputed data sets is smaller than for Listwise deletion. For example, the number of error degrees of freedom of the intercept is as low as 7 whereas for Listwise deletion it is 68 (second row). When data are imputed 100 times the differences in number of degrees of freedom with Listwise deletion are smaller, and for most *F*-tests (though not all) the number of degrees of freedom are larger than for Listwise deletion.

In this example, the $\bar{\lambda}$s (last column) are substantially larger on average than in the previous examples. This is probably due to a combination of a higher percentage of missing data than in the previous examples, and a relatively small sample size, causing much uncertainty about the imputed values.

Additionally, the $\bar{\lambda}$s for 5 imputations are substantially larger than for 100 imputations. To see whether differences in $\bar{\lambda}$ between 5 imputations and 100 imputations were systematic or caused by sampling fluctuation, we reran the imputation procedure with 5 imputations with many different seed values. These runs showed that the values of $\bar{\lambda}$ varied substantially across different imputation runs with 5 imputations, and were sometimes larger and sometimes smaller than for 100 imputations. Thus, differences found between 5 and 100 imputations are most probably due to sampling fluctuation.

Finally, the main effect of Treatment is not significant for 5 imputations while it is significant for Listwise deletion and 100 imputations.

Discussion

In this paper the pooling of *F*- and *p*-values for (repeated measures) ANOVA on multiply imputed data sets was described in detail. We showed how the existing regression rules for pooling Multiple imputation significance tests (Rubin, 1987) can also be applied to (repeated-measures) ANOVA provided one uses effect coding rather than dummy coding. The procedures were illustrated with three data sets, one with two between-subjects factors, one with two within-subjects factors, and one with a between-subjects and a within-subjects factor.

The benefits of the described procedure are considerable. While it was not described in the past how to pool significance tests of (repeated measures) ANOVA for multiply imputed data sets, a procedure to do so was outlined in this paper, which makes use of already available procedures. This has the advantage that no simulations need to be carried out to study the robustness of the proposed methods, because our proposal is completely based on existing methods of which the robustness has already been discussed and studied in other contexts (Barnard & Rubin, 1999; Li, Raghunathan, & Rubin, 1991; Reiter, 2007).

One noticeable finding was that in two of the examples the number of error degrees of freedom of the *F*-tests was (substantially) smaller for five imputed data sets than for Listwise deletion. However, when data were imputed 100 times, the number of error degrees of freedom was often larger than for Listwise deletion. These findings suggest that Multiple imputation would introduce more uncertainty in the results than Listwise deletion. This is, however, an awkward conclusion since Multiple imputation uses more information in the data than Listwise deletion does. A more plausible interpretation would be that for high amounts of missing data the approximations of the degrees of freedom are (too) sensitive to the number of imputations.

Also, for five imputations it was found that an effect was significant which was not significant for 100 imputations and Listwise deletion, and the other way around. Moreover, in the Anorexia example the results turned out to vary substantially for 5 imputations when the multiple-imputation procedure was repeatedly carried out. The differences in findings for 5 and 100 imputations and the unstable results for 5 imputations suggest that 5 imputations are too few for obtaining stable results (also, see Bodner, 2008, who reached similar conclusions). When some *F*-tests of the ANOVA have *p*-values close to the significance level it is advisable using more than 5 imputations, in order to obtain more reliable results.

However, for the McArdle-Nesselroade developmental change data the numbers of error degrees of freedom obtained from 5 imputations were larger than for Listwise deletion, and did not differ substantially from the numbers of error degrees of freedom obtained from 100 imputations. We suspected that this difference with the other two examples was due to a relatively small sample size in combination with the chosen imputation method (Predictive mean matching). With a small sample size Predictive mean matching will often find the same matching cases over multiply imputed data sets, resulting in a small between-imputation variance. Due to this small between-imputation variance the number of error degrees of freedom is already quite large for five imputations. Consequently, increasing the number of imputations to 100 will not improve much on the stability.

To test this conjecture we also imputed the data using the linear regression method from MICE instead of Predictive mean matching (results not shown). Roughly said, this method imputes expected values from a regression model plus some random error term, rather than finding matching cases. Since the imputed values from this imputation method are not restricted to values from certain response patterns observed in the data, they can vary more than when Predictive mean matching is used, resulting in more between-imputation variance.

When using this method, the number of error degrees of freedom indeed dropped below the number of degrees of freedom for Listwise deletion, for each *F*-test. However, besides this drop in degrees of freedom, conclusions obtained from the *F*-tests also changed substantially (i.e., *p*-values changing from far below the significance level to higher than 0.50). When inspecting the data it turned out that many of the variables were skew-distributed. Thus, on second thought the linear regression approach did not seem to be the best imputation method for this data set either. On the other hand, the data set may be too small to find an appropriate number of matching cases for predictive mean matching. It seems like a proper handling of the missing data goes beyond the use of these data as an illustration. In fact, this would require a separate study.

Implementation in Existing Software Packages

To date, the implementation of the combination rules needed for ANOVA in statistical software packages has either been quite limited or very well hidden. SPSS 19.0 does not provide any pooled results for *F*-tests for multiply imputed data. Using an SPSS macro (Van Ginkel, 2010a) the procedure may be carried out in SPSS but this still requires substantial manual pre-processing. For repeated-measures ANOVA the procedure is even more complicated.

Although SAS 9.3 and Stata 10.0 have a procedure for carrying out the combination rules for ANOVA in Multiple imputation, the manual provides no clear practical guidance in how to use them in the specific context of ANOVA. The procedure outlined in this paper should also be of assistance to any SAS or Stata user who intends to use Multiple imputation in this context. In the appendix it is explained how exactly the procedures can be carried out in SAS 9.3, Stata 10.0, and SPSS 19.0 using Van Ginkel's (2010a) macro.

Considering the involved task needed to actually carry out the described procedure, its implementation in future releases of SPSS and other software packages, as well as a clear guidance with examples in the SAS and Stata manuals seems desirable if its use is to become common place for routine use.

Further improvements in Multiple imputation for ANOVA

A next step in extending the pooling of ANOVA results of multiply imputed data sets would be to derive the pooling techniques for post hoc tests, contrasts, and effect sizes. Although to our knowledge no literature on this exists as of yet, this will mostly be a matter of applying the existing rules to these situations, just like we did in this paper for the *F*-tests in ANOVA.

Acknowledgments

The authors would like to thank Dr. Alan Taylor for his help with Stata.

APPENDIX: Complete scheme for pooling significance tests of a two-way ANOVA of multiply imputed data sets

Scheme for independent measures:

1. Use effect coding for the coding of the independent variables. If effect coding is not implemented in the software package used, this requires creating additional effect coded variables which should then be used as predictors in the regression.
2. Estimate the regression model of Equation 19 using a procedure in a statistical software package that provides both its regression coefficients and the covariance matrix of Equation 10. In SPSS such a procedure is Mixed models, in SAS this can be done using the procedure PROC MIXED, and in Stata using the procedure XTMIXED.
3. For effects of variables with two categories, use the formulas from Equations 1–6, and 8 for pooling the *F*- and *p*-values, for effects of variables with more than two categories use the formulas from Equations 11–15, and 17. In SAS 9.3 these calculations are provided in the PROC MIANALYZE procedure using the MULT option within the TEST statement. Stata 10.0 can perform these computations using the MI ESTIMATE procedure. SPSS 19.0 does not support the combination rules needed but they may be applied using Van Ginkel's (2010a; 2010b) macro.

Scheme for repeated measures:

1. Store the data in stacked format with different time points below each other.
2. Use effect coding for the coding of both the between-subjects factors and within-subjects factors.
3. Carry out the multilevel regression model from either Equation 21 or 24, providing both its regression coefficients and the covariance matrix of Equation 22 (Mixed models in SPSS 19.0, PROC MIXED in SAS 9.3, and XTMIXED in Stata 10.0). If manually created effect coded variables are used, beware that these effect coded variables cannot be used for specifying the random effects. For example, whereas it is sufficient in Mixed models in SPSS 19.0 to use the indicator variables for both the fixed and the random effects when the standard implemented dummy coding is used, this will not work for effect coding. In the latter case, the effect coded variables will have to be entered as Covariates and specified as fixed effects, while the original indicator variables will have to be entered as Factors and specified as random effects.

4. Use the formulas from Equations 1–6, 8, 11–15, and 17 for pooling the *F*- and *p*-values (MIANALYZE in SAS 9.3, MI ESTIMATE in Stata 10.0, and using the macro by Van Ginkel in SPSS 19.0). Note: for repeated measures, carrying out the pooling procedure in SPSS 19.0 requires substantial pre-processing of the covariance matrix of Equation 22 since it is saved in a format that cannot readily be read by the macro. Examples are available from the first author upon request.

REFERENCES

Arbuckle, JL.; Wothke, W. AMOS 7.0 [Computer software]. Chicago, IL: Smallwaters; 2010.

Bakermans-Kranenburg MJ, Van IJzendoorn MH, Kroonenberg PM. Differences in attachment security between African-American and white children: Ethnicity or socio-economic status? Infant Behavior and Development. 2004; 27:417–433. (Correction: 28, 96–96, 2005);

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

Bodner TE. What improves with increased missing data imputations? Structural Equation Modeling. 2008; 15:651–675.

Edwards, AL. Multiple regression analysis and the analysis of variance and covariance. New York: Freeman; 1985.

Giesbrecht FG, Burns JC. Two-stage analysis based on a mixed model: Large sample asymptotic theory and small-sample simulation results. Biometrics. 1985; 41:477–486.

Graham JW. Adding missing-data-relevant variables to FIML-based structural equation models. Structural Equation Modeling. 2003; 10:80–100.

Hand, DJ.; Daly, F.; McConway, K.; Lunn, D.; Ostrowski, E. A Handbook of small data sets. Boca Raton, FL: Chapman & Hall; 1994.

Hox, JJ. Multilevel analysis: Techniques and applications. Mahwah, NJ: Erlbaum; 2002.

Li KH, Raghunathan TE, Rubin DB. Large-sample significance levels from multiply imputed data using moment based statistics and an*F*reference distribution. Journal of the American Statistical Association. 1991; 86:1065–1073.

Little RJA. Missing-data adjustments in large surveys. Journal of Business and Economic Statistics. 1988; 6:287–296.

Little, RJA.; Rubin, DB. Statistical analysis with missing data. 2nd ed. New York: Wiley; 2002.

Maxwell, SE.; Delaney, HD. Designing experiments and analyzing data. 2nd ed.. Mahwah, NJ: Lawrence Erlbaum Associates; 2004.

McArdle, JJ.; Nesselroade, JR. Using multivariate data to structure developmental change. In: Cohen, SH.; Reese, HW., editors. Life-span developmental psychology: Methodological contributions. Hillsdale, NJ: Lawrence Erlbaum Associates; 1994. p. 223-267.

NICHD Early Childcare Research Network. Characteristics of infant childcare: Factors contributing to positive caregiving. Early Childhood Research Quarterly. 1996; 11:269–306.

Reiter JP. Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. Biometrika. 2007; 94:502–508.

Robinson GK. That BLUP is a good thing: The estimation of random effects. Statistical Science. 1991; 6:15–32.

Rubin DB. Inference and missing data. Biometrika. 1976; 63:581–592.

Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business and Economic Statistics. 1986; 4:87–94.

Rubin, DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.

SAS Institute Inc. SAS® 9.3 [Computer software]. Cary, NC: SAS Institute Inc; 2011.

Schafer, JL. Analysis of incomplete multivariate data. London: Chapman & Hall; 1997.

Schafer JL. NORM: Version 2.03 for Windows 95/98/NT [Computer software]. 1998 Retrieved February 27th, 2011, from http://www.stat.psu.edu/~jls/misoftwa.html.

Snijders, T.; Bosker, R. Multilevel analysis. London: Sage; 1999.

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

SPSS Inc. SPSS 19.0 for Windows [Computer software]. Chicago: Author; 2010.

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

Su YS, Gelman A, Hill J, Yajima M. Multiple imputation with diagnostics (mi) in R: opening Windows into the black box. Journal of Statistical Software. 2011; 45:1–31.

Van Buuren, S. Flexible imputation of missing data. Boca Raton: Chapman & Hall/CRC Press; 2012.

Van Buuren S, Groothuis-Oudshoorn CGM. MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011; 45:1–67.

Van Ginkel, JR. MI-MUL2.SPS [Computer code]. 2010a. Retrieved October 21st, 2011, from http://www.socialsciences.leiden.edu/educationandchildstudies/childandfamily studies/organisation/staffcfs/van-ginkel.html

Van Ginkel, JR. MI-MUL2.pdf [Sofware manual]. 2010b. Retrieved February 27th, 2011, from http://www.socialsciences.leiden.edu/educationandchildstudies/childandfamily studies/organisation/staffcfs/van-ginkel.html

Van Ginkel, JR.; Kroonenberg, PM. Multiple Imputation and (Repeated Measures) Analysis of Variance; Presentation given at the 17th International Meeting of the Psychometric Society; July, 2011; Hong Kong. 2011.

Winer, BJ. Statistical principles in experimental designs. 2nd ed. New York: McGraw-Hill; 1971.

Yuan, YC. Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference (Paper, No. 267). Cary, NC: SAS Institute; 2000. Multiple imputation for missing data: Concepts and new development. Retrieved February 27th, 2011, from http://www.ats.ucla.edu/stat/sas/library/multipleimputation.pdf.

Table 1

Descriptive Statistics for Selected Categorical Variables from the NICHD Study of Early Child Care (1996).

Variable	Values	n
Gender	1. Male	705
	2. Female	659
	Total	1364
Mother’s ethnicity	1. American Indian, Eskimo, Aleutin	8
	2. Asian; Pacific Islander	30
	3. Black; Afro-American	174
	4. White	1127
	5. Other	25
	Total	1364
Mother’s education	1. seven years of schooling	3
	2. eight years of schooling	13
	3. nine years of schooling	20
	4. ten years of schooling	42
	5. eleven years of schooling	61
	6. High school diploma	287
	7. Some college; no degree	455
	8. Bachelor's degree	284
	9. Some graduate work or master's degree	161
	10. law degree	14
	11. PhD; several master’s degrees	23
	Total	1363

Table 2

Descriptive Statistics for Selected Numerical Variables from the NICHD Study of Early Child Care (1996)

Variable	***n***	**Min**	**Max**	***M***	***SD***
School Readiness	1159	1	17	9.02	2.89
Depression mother	1119	0	51	9.40	8.63
Mother's age	1364	18	46	28.11	5.63
Total family income ($)	1189	2500	400002	52374	41645
Mother IQ (Peabody PVT)	1167	40	159	99.01	18.35
Bayley Mental Develomental. Index	1162	50	150	92.15	14.64

Table 3

ANOVA Results of the Incomplete Data Set From the NICHD (1996) Project, Using Listwise Deletion, Five Imputations, and 100 Imputations. Significant Effects are Printed in Bold.

Effect	**Results obtained from:**	df_{between}	df_{within}	F	p	$\bar{\lambda}$
Model	Listwise deletion	9	1149	21.11	**<.001**	-
	5 imputations		670	21.69	**<.001**	.15
	100 imputations		1318	22.69	**<.001**	.13
Intercept	Listwise deletion	1	1149	984.63	**<.001**	-
	5 imputations		432	969.03	**<.001**	.08
	100 imputations		1257	1007.59	**<.001**	.05
Gender	Listwise deletion	1	1149	3.76	>.05	-
	5 imputations		655	3.94	**<.05**	.05
	100 imputations		1194	3.69	.06	.07
Mother’s ethnicity	Listwise deletion	4	1149	38.31	**<.001**	-
	5 imputations		350	39.17	**<.001**	.14
	100 imputations		1298	42.02	**<.001**	.11
Gender × Mother’s ethnicity	Listwise deletion	4	1149	0.41	.80	-
	5 imputations		394	0.25	.91	.13
	100 imputations		1280	0.33	.86	.13

Note: df_{within} (per imputed data set; all effects) = 1354.

Table 4

Descriptions of the Tasks in the Study From McArdle and Nesselroade (1994).

Task	Description
1.	Trial 1 on Board Game (delayed spatial recognition task)
2.	Trial 2 on Board Game
3.	Forward Digit Span
4.	Backward Digit Span
5.	True Gist Correct (Text memory task)
6.	False Gist Correct (Text memory task)
7.	Elaborations (Text memory task)

Table 5

Descriptive Satistics of the Variables in the Data Set from the McArdle and Nesselroade (1994).

	n	Min	Max	*M*	*SD*
Week 1, Task 1	27	0	16	8.26	4.588
Week 1, Task 2	27	0	15	6.37	3.432
Week 1, Task 3	30	4	11	7.37	1.712
Week 1, Task 4	30	3	10	6.40	1.754
Week 1, Task 5	32	2	8	6.66	1.310
Week 1, Task 6	32	3	8	5.56	1.318
Week 1, Task 7	32	3	7	5.37	1.129
Week 13, Task 1	31	3	16	8.03	4.637
Week 13, Task 2	31	2	15	7.84	4.677
Week 13, Task 3	31	4	11	7.55	1.929
Week 13, Task 4	31	3	11	7.42	2.262
Week 13, Task 5	29	6	8	7.45	.736
Week 13, Task 6	29	3	8	6.17	1.513
Week 13, Task 7	29	3	8	6.45	1.298
Week 25, Task 1	31	0	16	8.84	4.677
Week 25, Task 2	30	0	15	8.13	5.070
Week 25, Task 3	31	0	11	7.42	2.680
Week 25, Task 4	29	2	13	7.52	2.516
Week 25, Task 5	30	5	8	7.03	1.066
Week 25, Task 6	27	1	8	6.19	1.520
Week 25, Task 7	26	4	7	5.85	1.223

Table 6

ANOVA Results of the Incomplete Data Set From McArdle and Nesselroade (1994), Using Listwise Deletion, Full Information Maximum Likelihood, Five Imputations, and 100 imputations. Significant Effects are Printed in Bold.

Effect	Results obtained from:	df_{between}	df_{within}	F	p	$\bar{\lambda}$
Intercept	Listwise deletion	1	22	824.29	**<.001**	-
	FIML		31	1266.98	**<.001**	-
	5 imputations		29	1270.61	**<.001**	.02
	100 imputations		29	1296.90	**<.001**	.02
	Per imputed data set		31	-	-	-
Task	Listwise deletion	6	132	7.24	**<.001**	-
	FIML		189	7.12	**<.001**	-
	5 imputations		176	7.15	**<.001**	.06
	100 imputations		184	7.24	**<.001**	.05
	Per imputed data set		186	-	-	-
Week	Listwise deletion	2	44	7.74	**<.01**	-
	FIML		52	5.22	**<.01**	-
	5 imputations		57	4.19	**.02**	.04
	100 imputations		60	4.26	**.02**	.08
	Per imputed data set		62	-	-	-
Task × Week	Listwise deletion	12	264	0.90	.55	-
	FIML		354	0.76	.69	-
	5 imputations		347	0.65	.80	.08
	100 imputations		369	0.55	.88	.08
	Per imputed data set		372	-	-	-

Table 7

Descriptive Statistics of the Anorexia Data (Hand et al., 1994).

	Cognitive behavioral treatment (n = 29)				***Control (n = 26)***				***Family therapy (n = 17)***			
	Min	**Max**	***M***	***SD***	**Min**	**Max**	***M***	***SD***	**Min**	**Max**	***M***	***SD***
Pretest	70.00	94.90	82.69	4.85	70.50	91.80	81.82	5.67	73.40	94.20	83.23	5.02
Posttest	71.30	103.60	85.70	8.35	73.00	89.60	80.88	4.70	75.20	101.60	90.49	8.48

Table 8

ANOVA Results of the Anorexia Data (Hand et al., 1994), no Multiple Imputation (Unbalanced), Five Imputations, and 100 Imputations. Significant Effects are Printed in Bold.

Effect	**Results obtained from:**	df_{between}	df_{within}	F	p	$\bar{\lambda}$
Intercept	Unbalanced	1	68	18136.73	**<.001**	-
	5 imputations		7	7978.55	**<.001**	.64
	100 imputations		42	12610.90	**<.001**	.45
Treatment	Unbalanced	2	68	5.96	**<.01**	-
	5 imputations		10	2.78	.11	.54
	100 imputations		72	4.20	**.02**	.43
Time point	Unbalanced	1	68	11.83	**<.01**	-
	5 imputations		10	5.94	**.04**	.55
	100 imputations		54	8.98	**<.01**	.31
Treatment × Time point	Unbalanced	2	68	6.21	**<.01**	-
	5 imputations		14	4.12	**.04**	.42
	100 imputations		76	4.56	**.01**	.35

Note: df_{within} (per imputed data set; all effects) = 84.