Pooling ANOVA Results From Multiply Imputed Datasets: A Simulation Study

Simon Grund, Oliver Lüdtke, and Alexander Robitzsch

Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Centre for International Student Assessment, Germany

Abstract

The analysis of variance (ANOVA) is frequently used to examine whether a number of groups differ on a variable of interest. The global hypothesis test of the ANOVA can be reformulated as a regression model in which all group differences are simultaneously tested against zero. Multiple imputation offers reliable and effective treatment of missing data; however, recommendations differ with regard to what procedures are suitable for pooling ANOVA results from multiply imputed datasets. In this article, we compared several procedures (known as D_1, D_2 and D_3) using Monte Carlo simulations. Even though previous recommendations have advocated that D_2 should be avoided in favor of D_1 or D_3, our results suggest that all procedures provide a suitable test of the ANOVA's global null hypothesis in many plausible research scenarios. In more extreme settings, D_1 was most reliable, whereas D_2 and D_3 suffered from different limitations. We provide guidelines on how the different methods can be applied in one- and two-factorial ANOVA designs and information about the conditions under which some procedures may perform better than others. Computer code is supplied for each method to be used in freely available statistical software.

Keywords: multiple imputation, missing data, multiparameter test, pooling, ANOVA.

Pooling ANOVA Results From Multiply Imputed Datasets: A Simulation Study

The analysis of variance (ANOVA) is a popular method for analyzing data in many fields of psychology and the social sciences (Cohen, Cohen, West, & Aiken, 2003; Maxwell & Delaney, 2004). One of the major goals of an ANOVA is to examine whether a number of groups (e.g., demographic features, experimental conditions) differ with respect to some variable of interest. The global null hypothesis, according to which all groups stem from the same population, is tested by comparing the portions of variance that reside between and within groups. Under the null hypothesis, the ratio of the mean squares between and within groups follows an F distribution. If group differences are reasonably large compared with individual differences, the global null hypothesis is rejected, and groups are believed to differ with respect to the variable of interest.

Missing data are a pervasive problem in the social sciences. Deleting the missing values (e.g., listwise deletion) is an easy but inefficient way of dealing with missing data that can seriously distort statistical analyses (Little & Rubin, 2002). Other techniques such as multiple imputation (Rubin, 1987) promise a more reliable and efficient treatment of missing data (Schafer & Graham, 2002). Multiple imputation (MI) draws a number of M replacements for the missing values from their posterior predictive distribution, given the observed data and a statistical model. The completed datasets are then analyzed using regular complete-data methods, and the parameter estimates are pooled according to the rules described in Rubin (1987) to form final parameter estimates and inferences.

Rubin's rules are easily applied to one-dimensional estimands such as means or regression coefficients, but multidimensional estimands (e.g., comparing multiple groups in the ANOVA's F test) call for different methods. Several such methods are discussed in the literature, and clear

recommendations can be found in various books and articles (Little & Rubin, 2002; Marshall, Altman, Holder, & Royston, 2009; Reiter & Raghunathan, 2007; Schafer, 1997). However, some authors' conclusions are less than definite and they emphasize the need for further research concerning realistic applications of these methods (Enders, 2010; Snijders & Bosker, 2012; van Buuren, 2012). In addition, previous studies have often focused on a technical understanding of these methods without considering specific research designs. Using computer simulations, we compared several pooling methods for the F test in one- and two-factorial ANOVA designs. We examined the robustness of these methods as well as the conditions under which some methods may be more trustworthy than others. We attempted to complement the existing literature with simulation results that can be easily applied to research practice. Computer code is given for each method to be used in freely available software.

Pooling ANOVA Results

The one-factorial ANOVA can be reformulated as a regression model in which the outcome variable is regressed on a number of dummy variables that represent the membership in a group i ($i = 1, \ldots, I$). For I groups, the group membership can be coded by $K = I - 1$ dummy variables such that the regression coefficients reflect differences between groups. In complete datasets, the Wald test of the K-dimensional vector of regression coefficients (without the intercept) is equivalent to testing the ANOVA's null hypothesis that there are no differences between groups (e.g., Cohen et al., 2003). Over the past years, several methods have become available for carrying out multiparameter hypothesis tests in multiply imputed datasets (e.g., Enders, 2010; Little & Rubin, 2002; Schafer, 1997; van Buuren, 2012). These methods build on different aspects of the completed-data analyses and thus differ in behavior and ease of application. Here, we provide a

brief overview of the procedures featured in our study, illustrated for the one-factorial ANOVA. The procedures extend naturally to two-factorial designs, with effect coding instead of dummy coding.

Moment based statistics (D_1 and D_1^*). The D_1 procedure extends Rubin's rules to multidimensional estimands such as the K-dimensional vector of regression coefficients in the ANOVA. Using D_1, the vectors of regression coefficients and their associated covariance matrices are pooled across the imputed datasets. Given a set of coefficient vectors $\hat{Q}_m$ ($m = 1, \ldots, M$) and estimates of their sampling covariance matrix $\hat{U}_m$, the D_1 statistic reads

$$D_1 = \frac{(\bar{Q} - Q_0)^T \bar{U}^{-1} (\bar{Q} - Q_0)}{K(1 + \text{ARIV}_1)} , \tag{1}$$

where $K = I - 1$ is the number of regression coefficients that represent group differences, $\bar{Q}$ and $\bar{U}$ are the average point and covariance estimates, and Q_0 is the vector of regression coefficients expected under the null hypothesis. The ARIV_1 denotes the average relative increase in variance due to nonresponse, that is, the extent to which the sampling variance of the estimator has increased due to missing data

$$\text{ARIV}_1 = \frac{(1 + M^{-1})\text{tr}(B\bar{U}^{-1})}{K} , \tag{2}$$

where B is the covariance matrix of the estimates $\hat{Q}_m$ across the imputed datasets (see Enders, 2010, for an illustration). The ARIV is conceptually related to the fraction of missing information (FMI; Rubin, 1987), which denotes the portion of the total sampling variance of an estimator that is due to missing data[1]. Rubin (1987) and Li, Raghunathan, and Rubin (1991) derived an

[1]Estimates of the FMI were based on estimates of the ARIV such that FMI = ARIV/(1+ARIV).

F reference distribution for D_1, along with K numerator and v_1 denominator degrees of freedom. For $a = K(M-1)$, the denominator degrees of freedom are calculated as

$$v_1 = \begin{cases} 4 + (a-4)[1 + (1-2a^{-1})\,\mathrm{ARIV}_1^{-1}]^2 & \text{if } a > 4 \\ (K+1)(M-1)(1+\mathrm{ARIV}_1^{-1})^2/2 & \text{otherwise} \end{cases}. \tag{3}$$

In its original formulation, the degrees of freedom for D_1 were derived under the assumption of infinite complete-data degrees of freedom. Reiter (2007) proposed a correction formula that adjusts the denominator degrees of freedom v_1 for finite samples. The resulting test is henceforth called D_1^*. Calculating D_1 and D_1^* requires pooling the point and variance estimates across datasets, a task that is relatively simple and well documented (see Enders, 2010).

The D_1 procedure is frequently recommended in the literature (e.g., Allison, 2001; Enders, 2010; Graham, 2012; Little & Rubin, 2002; Schafer, 1997; van Buuren, 2012). Li, Raghunathan, and Rubin (1991) showed that D_1 is reliable and robust unless the FMI is very large and variable across parameters. Reiter (2007) showed that D_1^* produced accurate Type I error rates even in small samples. Licht (2010) proposed an adjustment of D_1 and replicated the favorable results of Li, Raghunathan, and Rubin (1991) for finite samples and larger K. van Ginkel and Kroonenberg (2014) illustrated the use of D_1^* in empirical datasets. However, simulation results regarding D_1 and D_1^* are still relatively scarce, and van Buuren (2012) suggests evaluating them "in more general settings" (p. 157). Enders (2010) found it "difficult to assess the trustworthiness of the D_1 statistic in realistic research scenarios" (p. 236).

***p* values from Wald-like hypothesis tests (D_2).** Li, Meng, Raghunathan, and Rubin (1991) developed a test statistic that is computed from a series of Wald tests (or their p values,

equivalently) rather than from point and variance estimates. This is especially useful if K is large or variance estimates (e.g., standard errors) are not available. Given a number of M Wald-like test statistics W_m, the D_2 statistic reads

$$D_2 = \frac{\overline{W}K^{-1} - (M+1)(M-1)^{-1}\text{ARIV}_2}{1 + \text{ARIV}_2} , \tag{4}$$

where $\overline{W}$ is the average test statistic across datasets and K is again the number of parameters that represent group differences. The ARIV_2 is another estimate of the average relative increase in variance that is based solely on the individual test statistics W_m

$$\text{ARIV}_2 = (1 + M^{-1}) \left[\frac{1}{M-1} \sum_{m=1}^{M} \left(\sqrt{W}_m - \overline{\sqrt{W}} \right)^2 \right] , \tag{5}$$

where $\overline{\sqrt{W}}$ denotes the average $\sqrt{W}_m$ across the imputed datasets (see Enders, 2010). Li, Meng, et al. (1991) proposed an F reference distribution for D_2 with K numerator and v_2 denominator degrees of freedom

$$v_2 = K^{-3/M}(M-1)(1 + \text{ARIV}_2^{-1})^2 . \tag{6}$$

In order to apply D_2, the individual test statistics (W_m) should follow a χ^2 distribution. Hence, in ANOVA models, the F values for all datasets (F_m) must be transformed such that $W_m = KF_m$, each of which approach a χ^2 distribution as the denominator degrees of freedom go to infinity. The D_2 statistic is easily calculated by pooling the test statistics across datasets. No specialized software or programming skills are required in order to calculate D_2, and only the M test statistics from the imputed datasets must be entered into the formulae, which are routinely included in the output of most statistical software.

However, the literature often advises against D_2. Li, Meng, et al. (1991) suggested that it be used only as a rough guide because its Type I error rates can be too high or too low depending on the FMI. It is usually recommended that D_1 be used whenever possible because D_2 is less precise, less powerful, and only loosely correlated with the "more nearly optimal" D_1 (Schafer, 1997, p. 116; Enders, 2010; Little & Rubin, 2002). Nonetheless, D_2 has been acknowledged for its ease of implementation because it operates directly on the test statistics (e.g., Allison, 2001; Snijders & Bosker, 2012). Van Buuren (2012) advised that D_2 may be used if nothing but the test statistics are available but that D_2 is "considerably less reliable" than other pooling methods (p. 159).

Pooled likelihood-ratio tests (D_3). Coming from the perspective of model comparison, hypotheses about a set of parameters can be tested using likelihood-ratio tests (LRTs). The D_3 procedure was developed by Meng and Rubin (1992) to enable LRTs with multiply imputed datasets. The procedure does not require variance estimates; instead, it operates on the likelihood. Meng and Rubin (1992) showed that it is not sufficient to simply combine the individual LRT statistics L_m into an average $\bar{L}$. In addition, the LRT statistic needs to be evaluated at the average estimates of the model parameters for all imputed datasets. The D_3 statistic reads

$$D_3 = \frac{\tilde{L}}{K(1 + \text{ARIV}_3)} , \tag{7}$$

where $\tilde{L}$ is the mean LRT statistic across the imputed datasets evaluated at the average parameter estimates, and K is the number of parameters being tested. Estimating the ARIV_3 includes the two pooled LRTs evaluated at the individual and pooled estimates, respectively (see Enders, 2010)

$$\text{ARIV}_3 = \frac{M + 1}{K(M - 1)}(\bar{L} - \tilde{L}) . \tag{8}$$

According to Meng and Rubin (1992), the F reference distribution for D_3 has K numerator and v_3 denominator degrees of freedom. For $a = K(M-1)$,

$$v_3 = \begin{cases} 4 + (a-4)[1 + (1-2a^{-1})\,\mathrm{ARIV}_3^{-1}]^2 & \text{if } a > 4 \\ (K+1)(M-1)(1+\mathrm{ARIV}_3^{-1})^2/2 & \text{otherwise} \end{cases}. \quad (9)$$

Calculating D_3 can be tedious because it requires that users have access to the likelihood function and that it is possible to evaluate it at user-defined values. Due to its complexity, the procedure is not frequently used, but it has been implemented in likelihood-oriented software such as M*plus* (Asparouhov & Muthén, 2008), SAS (Mistler, 2013) and the `semTools` package for R (Pornprasertmanit, 2014). The D_3 statistic is frequently recommended when D_1 cannot be calculated, that is, in the absence of standard errors (Little & Rubin, 2002; van Buuren, 2012). It has been argued that D_1 and D_3 should behave similarly, and more reliably than D_2, because the two are approximately equal (Meng & Rubin, 1992; Schafer, 1997). However, Enders (2010) pointed out that "virtually no research studies have compared the two test statistics" (p. 241).

Present Study

Even though recommendations regarding D_1, D_1^*, D_2 and D_3 can be found in the literature, the behavior of these methods is still not fully understood. Earlier studies focused on the general properties of these methods, and simulation studies considered the FMI as a pivotal point (e.g., Li, Meng, et al., 1991; Li, Raghunathan, & Rubin, 1991). Their usual recommendation is that, in general, some procedures should be preferred (D_1, D_1^*, D_3), while others should be avoided (D_2). However, in the present article, we argue that all of these methods provide suitable tests for ANOVA models in most conditions that are encountered in psychological research. We conducted

computer simulations that explore their performance from the perspective of practical research. Our results are intended to complement the existing literature with results that can be easily applied to practical research, and to assist researchers in their statistical decision making.

We examined the Type I error rates and the statistical power of the four pooling methods. Study 1 features a fully crossed simulation design in which the number of groups, the group size, the effect size, the missing data mechanism, and the amount of information available from an auxiliary variable were varied. This design allowed us to examine possible interactions between the simulation factors. However, in order to reduce computational effort, some of its conditions had to be restricted. The conditions were chosen to mimic what frequently occurs in applications of the ANOVA in psychological research. Two additional studies were conducted that relaxed some of the restrictions made in Study 1. This made it possible to examine specific findings in greater detail. Study 2a provides details on how including an auxiliary variable into the imputation model may influence statistical power (Collins, Schafer, & Kam, 2001). For this purpose, we varied the correlation between outcome and auxiliary variable in very fine steps, thus exploring the conditions in which the ANOVA might benefit from using MI. In Study 2b, we examined the effects of larger FMIs on the Type I error rates, that is, for larger amounts of missing data and given different amounts of auxiliary information. In this context, we elaborate on the "link" between the simulation factors and the FMI in our simulation design. This was deemed helpful for judging the severity of missing data problems in research practice and for providing a reference frame for the results of earlier studies. Study 3 extends the paradigm of Study 1 to two-factorial ANOVA designs. In the two-factorial design we took special interest in testing the overall interaction effect, which, especially in large ANOVA designs, may involve a large number of parameters.

Study 1

The first simulation study was conducted to assess the performances of D_1, D_1^*, D_2, and D_3 under conditions that are commonly encountered in one-factorial ANOVA designs. All simulation factors were fully crossed in order to examine the factors that drive the performance of these methods.

Simulation Procedure

Data generating model. The ANOVA provided the foundation for the data generating model. A continuous outcome Y was simulated from a normal distribution given the group means μ_i for a factor A with groups $i = 1, \ldots, I$, that is,

$$Y = \mu_i + \epsilon \quad \text{with} \quad \epsilon \sim N(0, \sigma_\epsilon^2), \tag{10}$$

where σ_ϵ^2 denotes the variance within groups. According to Cohen (1988), the variance of the group means around the population mean (i.e., the grand mean) $\bar{\mu}$ can be defined as

$$\sigma_A^2 = \frac{\sum_{i=1}^{I} (\mu_i - \bar{\mu})^2}{I}. \tag{11}$$

The sum of the two variances (σ_A^2 and σ_ϵ^2) was defined to be one. The population mean was assumed to be zero. Differences between groups were simulated according to Cohen's (1988) f, here

$$f_A = \frac{\sigma_A}{\sigma_\epsilon}\ . \tag{12}$$

Thus, the two variances followed as

$$\sigma_A^2 = \frac{f_A^2}{1 + f_A^2} \quad \text{and} \quad \sigma_\epsilon^2 = 1 - \frac{f_A^2}{1 + f_A^2} \, . \tag{13}$$

Different patterns of group means were simulated in order to mimic plausible research scenarios. This was achieved by rephrasing all group means as $\mu_i = p_i d_A$, where the p_i form a pattern of group means $p_A = (p_1, \ldots, p_I)$ that sums to zero, and d_A is a scaling factor that enlarges this pattern so that it would imply the correct portions of variance as given by Equation 13. The scaling factor d_A was derived by rearranging Equation 11, which yields

$$d_A = \sigma_A \sqrt{\frac{I}{\sum_{i=1}^{I} p_i^2}} \, . \tag{14}$$

We simulated two patterns of group means labeled "difference" and "trend," respectively, in which either one third of the groups differed greatly from the others or all groups differed in such a way that they formed a linear trend. For example, with $I = 3$ groups, the two patterns can be written $p_{A,\text{difference}} = (-1/2, 1, -1/2)$ and $p_{A,\text{trend}} = (-1, 0, 1)$, respectively. To illustrate, suppose we wanted to establish an effect of size $f_A = .40$ forming a difference pattern $p_A = (-1/2, 1, -1/2)$. This implies a variance of group means $\sigma_A^2 = 0.16/1.16 = 0.14$; thus, the scaling factor would become $d_A = 0.37\sqrt{3/(0.25 + 1 + 0.25)} = 0.53$. Finally, the group means μ_i would be $(-0.26, 0.53, -0.26)$.

A second continuous variable X was simulated to allow for different missing data mechanisms and to mimic situations in which auxiliary information can be included in the imputation model. The covariate X was simulated as

$$X = \rho_{xy} Y + \epsilon_X \quad \text{with} \quad \epsilon_X \sim N(0, 1 - \rho_{xy}^2) \, , \tag{15}$$

where ρ_{xy} denotes the correlation between X and Y. Table 1 provides an overview of the simulation design of all studies. In Study 1, we varied the number of groups ($I = 3, 6, 12$), the sample size within each group ($n = 25, 50, 100$), the effect size ($f_A = 0, .10, .25, .40$), and the correlation between X and Y ($\rho_{xy} = 0, .35, .70$).

Imposition of missing values. Missing data were imposed on the outcome Y, whereas the covariate X and the group membership of each person were fully observed. Different missing data mechanisms were defined according to Rubin (1976). In this classification, the hypothetical complete data Y are divided into observed and unobserved portions, Y_{obs} and Y_{mis}, respectively. An indicator variable R denotes which values in Y are observed. Rubin (1976) introduced several broad classes of missing data mechanisms. If the missing values are simply a random sample of the hypothetical completely observed Y, then the values are missing completely at random (MCAR), that is, $P(R|Y_{\text{obs}}, Y_{\text{mis}}) = P(R)$. If the chance of observing Y depends on the observed data but does not further depend on the missing part, then the values are missing at random (MAR), that is, $P(R|Y_{\text{obs}}, Y_{\text{mis}}) = P(R|Y_{\text{obs}})$. The two are often called *ignorable* missing data because the exact missing data mechanism need not be known in order to perform MI. Treating nonignorable missing data requires making strong assumptions about the missing data mechanism and thus was not considered in this study (see Carpenter & Kenward, 2013).

The missing values were simulated using a latent response variable R^*, which determined whether values in Y were missing dependent on the covariate X under a linear model

$$R^* = \lambda X + \epsilon_{R^*} \quad \text{where} \quad \epsilon_{R^*} \sim N(0, 1 - \lambda^2) \; . \tag{16}$$

Values in Y were set missing if $R^* < z$, where z is a quantile of the standard normal distribution

according to the desired probability of missing data (e.g., $z = -0.67$ for 25% missing data). As presented in Table 1, we varied the effect of X on the latent response indicator to simulate different missing data mechanisms. For Y to be MCAR, we set $\lambda = 0$, and for Y to be MAR given X, we set $\lambda = .35$ or .70. The probability of missing data was held constant at 25% but was varied in Study 2b. Note that our simulation design implicitly varies the FMI by varying population and sample characteristics that influence the FMI. This is in contrast to previous studies, in which the FMI was varied explicitly (e.g., Li, Meng, et al., 1991; Li, Raghunathan, & Rubin, 1991; Licht, 2010). As mentioned before, the simulation design was chosen to mimic situations that are encountered in real-world applications of the ANOVA. Thus, we manipulated the severity of the missing data problem in terms of the design factors (e.g., amount of missing data, presence of auxiliary variables) rather than the FMI. This perspective was chosen so that the simulation design would directly relate to research practice, whereas the FMI would occur only insofar as it emerged from the simulated conditions.

Imputation and analysis. Imputations were carried out using the `mice` package (van Buuren & Groothuis-Oudshoorn, 2011) in the statistical software R (R Core Team, 2014). The "norm" imputation method was used; therefore, missing values on Y were assumed to be normally distributed given the group membership and the covariate X. Following recent recommendations, we created $M = 100$ imputed datasets for each simulated dataset (Bodner, 2008; Graham, Olchowski, & Gilreath, 2007). However, all analyses were repeated with different subsets of M, that is, with the first 5, 10, 20, and 50 of the total 100 datasets, respectively (see Table 1). The ANOVA model was fitted by dummy coding the grouping variable and regressing the outcome Y on the $K = I - 1$ dummy variables. All methods—D_1, D_1^*, D_2, and D_3—were implemented in the software R. The computer code is provided in the supplemental online material along with an

example application to artificial data (see also Grund, Robitzsch, & Lüdtke, 2016). In addition, listwise deletion (LD) was included as a strategy for handling missing data because it is still frequently used in research practice.

We compared the pooling methods with respect to Type I error rates and their power to detect nonzero effects. The Type I error rate is the relative frequency with which the null hypothesis is rejected when the population effect (f) is zero. Ideally, the Type I error rate should be close to the predefined significance level α (e.g., 5% or 1%). A procedure was considered liberal or conservative when its Type I error rate was higher or lower, respectively, than the nominal α. Bradley (1978) suggested a criterion for robustness, according to which Type I error rates within $\alpha \pm 0.5\alpha$ are considered acceptable (e.g., within 2.5% and 7.5% for $\alpha = 5\%$). In addition, we calculated the Type I error rates for the complete datasets (i.e., before imposing missing values) to provide a benchmark for the different pooling methods. The statistical power is the relative frequency with which the null hypothesis is rejected when the population effect is *not* zero. Assessing differences in statistical power is difficult because the expected power is not a fixed value for all conditions (Cohen, 1988). Thus, the expected power itself served as a benchmark for the pooling methods.

Results

The first study featured six simulation factors and 648 conditions in total. All conditions were replicated 10,000 times to ensure that the Type I error rates and the power to detect nonzero effects had stabilized. Reporting all results was not feasible due to the large number of conditions and because not all factors influenced the performance of the pooling methods. The complete results for $M = 100$ imputations are provided in the supplemental online material, intended as a

repository for interested readers. We focus on the "difference" pattern of group means, and assume a level of $\alpha = 5\%$ throughout this section. The results were similar for $\alpha = 1\%$ and will be discussed whenever necessary.

Type I Error Rate. In all conditions and for all pooling methods, the Type I error rates varied within a reasonable range, that is, below 6.1% (D_2) and above 4.2% (D_3). Thus, no violations of Bradley's criterion for robustness were observed at $\alpha = 5\%$. Some methods were found to be liberal in some cases (D_1 and D_2), whereas others were slightly conservative (D_1^* and D_3). The extent to which the pooling methods were conservative or liberal was mostly influenced by the group size (n) and the number of groups (I). Figure 1 illustrates the Type I error rates of all procedures for different group sizes and different numbers of groups, when the correlation between X and Y ($\rho_{xy} = 0$) and the effect of X on missingness ($\lambda = 0$) were held constant.

The D_1 statistic was slightly liberal in small samples (i.e., small n or I) but otherwise provided nearly optimal results. The error rates obtained with D_1^* were nearly optimal under all conditions. D_2 was the most liberal of all pooling methods, but even for D_2, the Type I error rates were not seriously inflated. Contrary to D_1, however, D_2 remained somewhat liberal in larger samples, especially when the number of groups was large. Finally, D_3 produced reasonable Type I error rates but was somewhat conservative if the number of groups was large and the groups were relatively small (e.g., $I = 12$ and $n = 25$). Results obtained with LD were generally close to the ideal solutions and usually close to those obtained with D_1^*.

With increasing group size, the Type I error rates of the four pooling methods became more similar; that is, D_1 and to a lesser extent D_2 became less liberal, whereas D_3 became less conservative. Effects of the number of groups were more diverse because an increase in I increased both the sample size and the number of parameters of the global null hypothesis test.

For D_1, D_1^*, and D_3, an increase in I led to more conservative results. Type I error rates for D_1^* and D_3 sometimes fell below those obtained from complete datasets and below the nominal α. D_2 on the other hand remained somewhat liberal for larger values of I unless the group size was very large in comparison (e.g., $I = 12$ and $n = 25$).

A lower level of $\alpha = 1\%$ did not change the picture as a whole; that is, all pooling methods performed similarly when compared with one another. Bradley's criterion demands that Type I error rates vary within 0.5% and 1.5% in this case. Type I error rates could be as high as 1.5% (D_1) for smaller groups, thus violating Bradley's criterion for robustness, but they were usually close to the nominal value in larger samples (see the supplemental online material).

Statistical power. Assessing the power of the pooling methods entailed certain limitations due to floor and ceiling effects, that is, when the power approached 5% and 100%, respectively. Differences between methods were found to be consistent regardless of effect size, but naturally, these became smaller when the power approached its upper or lower bounds. Especially for large effects ($f_A = .40$), choosing a particular method became less important because the power was effectively 100% for all methods unless the samples were very small. Therefore, we will focus on small and moderate effect sizes ($f_A = .10$ and .25) in order to describe the results on a scale that is informative and meaningful for applied researchers (power between 60% and 80%).

The more liberal methods (D_1 for smaller samples, D_2) also scored highest in statistical power. Most importantly, the power obtained with MI was higher than with LD whenever the covariate X was somewhat informative about the missing values on Y, where a higher correlation between X and Y (ρ_{xy}) led to higher power when MI was used. The effect of X on missingness (λ) did not greatly influence the power by itself but moderated the aforementioned effects such that higher values of λ intensified the differences between LD and MI (see Collins et al., 2001).

Figure 2 illustrates the interplay of the correlation between X and Y and the effect of X on missingness in larger samples (n = 100, I = 12, f_A = .10). All pooling methods and LD were equally capable of detecting nonzero effects when the covariate carried no information about the missing outcome (ρ_{xy} = 0). As soon as the covariate provided information (ρ_{xy} = .35 or .70), higher statistical power was observed when MI was used. Similar results were obtained for moderate samples with small and large groups, as presented in Table 2. For small groups (n = 25, I = 12, f_A = .25), the more liberal pooling methods (D_1 and D_2) provided higher statistical power. With larger groups (n = 50, I = 3, f_A = .25), the difference between the pooling methods became smaller. Again, higher power was observed for MI when the covariate provided information about the missing Y. The conservative methods had lower power in general and thus relied more heavily on such information. Nonetheless, even the conservative methods had higher power than LD, given sufficient auxiliary information.

Number of imputed datasets. The number of imputations was varied within each simulation condition in order to provide an insight into how the results would have changed if fewer than M = 100 imputations had been used. The initial recommendation that M = 5 imputations would suffice for most applications of MI (Rubin, 1987) has been modified in the past by several authors (e.g., Bodner, 2008; Graham et al., 2007; Harel, 2007).

Interpreting the effect of different values of M proved to be challenging because its effect depended on the group size, the number of groups, the correlation between X and Y, and also differed between pooling methods. Figure 3 shows the results for different M in selected conditions. The results obtained with D_1, D_1^*, and D_3 were relatively insensitive to the number of imputations but were best when M was at least 20. For D_2, however, the performance changed substantially when more than 20 imputations were generated: Type I error rates from D_2 became

slightly higher, and statistical power was much larger with M > 20, especially when the number of groups was large and the covariate X did not provide information about the missing Y ($\rho_{xy} = 0$). With fewer imputations ($M \leq 20$), D_2 tended to be conservative and suffered from a substantial loss of power. With a sufficient number of imputations, the power of the four methods was almost identical.

Discussion

The first simulation study compared different pooling methods for testing the global null hypothesis of the ANOVA with multiply imputed datasets. Differences emerged in terms of Type I error rates: Some methods tended to be slightly liberal (D_1 and D_2) or conservative (D_1^* and D_3), but no procedure led to Type I error rates far above or below the nominal value. The liberal methods also tended to detect nonzero effects more frequently. The biggest difference, however, emerged between MI and LD when a covariate that provided information about the missing values was included in the imputation model. In such cases, using MI could be highly beneficial whereas potential losses from using MI when the covariate carried no information were not observed (see Collins et al., 2001).

Our study was able to replicate previous findings on the performances of D_1 and D_1^*, which were found to be stable and reliable in most cases (Li, Raghunathan, & Rubin, 1991; Reiter, 2007). Although seldom recommended, D_2 provided very reasonable results within the scope of the first study. Moreover, our results suggest that D_2 is equally powerful as D_1 and D_1^* when the number of imputations is sufficiently large. This is in stark contrast to current recommendations regarding D_2, which suggest that D_2 should generally be avoided because it was optimized for $M = 3$ imputations, less powerful than D_1, and unlikely to improve with larger M (Schafer, 1997;

van Buuren, 2012). Our findings suggest that, due to its ease of application, D_2 might be a viable alternative in many applications of multiparameter tests, such as in the ANOVA, despite being theoretically less convincing than D_1. The D_3 procedure also provided good results but was unnecessarily conservative in small samples. Given that D_3 is rather difficult to implement, we believe that D_1 and D_1^* are better choices for ANOVA models unless researchers intend to use likelihood-based statistical software that already offers D_3 (see Enders, 2010). Care should be taken when the pooling methods are applied under more extreme conditions. The D_2 procedure was slightly more liberal when the number of groups I (and hence the number of parameters) was large. In such cases, D_3 was quite conservative unless the groups were very large in comparison.

Several limitations are noteworthy. First, due to the large simulation design, not all factors could be varied in very great detail. The simulation suggested that MI benefits when information about the missing values is available, but, at this point, it is unclear how much information a covariate must provide in order to be helpful. Thus, the purpose of Study 2a was to explore the potential gains in statistical power. Second, we chose a fixed value for the probability of missing data. The chosen value of 25% is quite large for many applications of the ANOVA, but the number of missing values can sometimes be higher depending on how the data were collected (e.g., Graham, Taylor, Olchowski, & Cumsille, 2006). Especially D_2 has been shown to be sensitive to very small and large values of the FMI (Li, Meng, et al., 1991). In order to close the gap between our results and the existing literature, it must be elaborated upon how the amount of missing data and the presence of auxiliary variables influence the FMI and, as a result, the robustness of the pooling methods. This was the purpose of Study 2b. Finally, Study 1 was limited to one-factorial ANOVA designs. Therefore, Study 3 was conducted, which extended the paradigm of Study 1 to two-factorial ANOVA designs and the test of interaction effects.

Study 2a

To examine the effects of including a more or less useful covariate in the imputation model, we varied the correlation between X and Y in steps of .05, ranging from $\rho_{xy} = 0$ to $\rho_{xy} = .95$. Either 25% or 50% missing values were introduced into the dataset. The remaining factors were held constant, as shown in Table 1. One hundred imputations were created. These values were chosen to reflect practical research but also to avoid influences of sampling error and boundary conditions. The results were cross-checked for different conditions, but the main pattern of results was found to be comparable.

Figure 4 shows the statistical power to detect moderate effects ($f_A = .25$) for all pooling methods and LD as a function of the strength of the relationship between the covariate and the outcome (ρ_{xy}). The performance of the pooling methods differed only when the correlation was small and became increasingly similar as the correlation grew larger. This is not surprising because the FMI was largest when X and Y were uncorrelated (see Study 2b). Listwise deletion was comparably powerful as long as X was only weakly correlated with Y. For larger values of the correlation ($\rho_{xy} = .35$ and above), the pooling methods consistently outperformed LD in terms of statistical power. Whereas the advantages of using MI remained modest for correlations below .50, larger correlations greatly improved statistical power. When 50% of the data were missing, the differences between the pooling methods grew larger, especially when X and Y were only weakly correlated. In this case, D_2 appeared to be more powerful than D_1 and D_1^*, and D_3 appeared to be less powerful, essentially reflecting differences in Type I error rates.

This illustrates that the conclusions of Study 1 cannot be generalized to arbitrarily harsh conditions, and that more severe missing data problems must be met with more sophisticated

methods (e.g., D_1 or D_1^*). Previous research has expressed these conditions in terms of the FMI. In Study 2b, we elaborate on how the FMI is related to the amount of missing data and auxiliary information, and how one's assessment of the missing data problem may guide one's choice among the pooling methods.

Study 2b

The FMI in our study was influenced by the amount of missing data and the correlation between X and Y. Figure 5 illustrates the relationship between these measures in our study. If auxiliary information was not available ($\rho_{xy} = 0$), then the FMI was equal to the amount of missing data. Therefore, the FMI could be manipulated directly when $\rho_{xy} = 0$ by varying the missing data rate. However, if the covariate is predictive of the missing values, then the FMI is lowered depending on the strength of that relationship. In other words, the missing data problem becomes less severe the more information can be included into the imputation model. In Study 1, the missing data rate was fixed to 25%, which is already quite large for many applications of the ANOVA. As can be seen from Figure 5, this corresponds to an FMI of only .25 if $\rho_{xy} = 0$, or less if $\rho_{xy} = .35$ or .70. In earlier studies, values for the FMI up to .50 were often considered (see Figure 4). In Study 2b, we investigated the effects of the FMI more thoroughly by including different portions of missing data, ranging from 5% to 80% in increments of 5%, as well as different values for the correlation of X and Y, effectively varying the FMI between .03 and .80 (see Table 1). Type I error rates were calculated for each condition.

Figure 6 shows the Type I error rates for all methods in smaller and larger samples, given different amounts of missing data and a more or less useful covariate. D_1 and D_1^* were robust even when large portions of data were missing and when the covariate did not provide information

about the missing data. In such extreme cases, as predicted by Li, Meng, et al. (1991), D_2 was less reliable, and increasingly liberal in larger samples. The results remained acceptable for up to 50% missing data, at which point Bradley's liberal criterion for robustness was violated (FMI of .50). However, if the correlation between X and Y was large, then D_2 was more reliable, and the results remained acceptable for up to 65% missing data (also FMI of .50). Notice that, in smaller samples, D_2 became less liberal again when the amount of missing data became very large (above 70%)[2]. Surprisingly, D_3 was also affected by larger FMIs such that, for large amounts of missing data (above 40%), D_3 became more and more conservative. These results occurred most strongly in smaller samples, where results remained acceptable for up to 65% missing data when X provided no information about Y. This effect too became smaller as the correlation between X and Y grew larger.

Study 3

The third study was conducted in order to assess whether our results could be generalized to two-factorial ANOVA designs and, in particular, to tests of the interaction effect. For this purpose, we extended the procedure of Study 1 to two-factorial designs in which two factors A and B, with I and J levels, respectively, could influence the outcome Y. The two main effects and the interaction effect were each assigned an effect pattern, denoted p_A, p_B and p_{AB}, respectively, and an effect size, denoted f_A, f_B and f_{AB}, respectively. The difference pattern was employed for the two main effects. The interaction effect was defined in a similar fashion such that groups on the

[2]The behavior of D_2 for large FMIs appeared to be a result of two compensatory mechanisms. Liberal behavior of D_2 was associated with F-values slightly larger than 1. The inflation of F values was associated with values of the ARIV_2 that were lower than the respective ARIV_1 (see Equation 4), especially in larger samples. Conservative behavior of D_2, on the other hand, seemed to be induced by the denominator degrees of freedom, v_2, which tended to be smaller than v_1, and noticeably so in smaller samples (see the first term in Equation 6; cf. Equations 3 and 9).

main diagonal of the $I \times J$ design would have larger values in Y compared to the off-diagonal groups. Scaling factors for each pattern were derived by the same logic as in Study 1. We chose similar values for the remaining simulation factors, as can be seen in Table 1. We examined the interaction effect in a 3×3 and 5×5 design with a different number of persons per group. Since the total sample size increased rapidly with I and J, we simulated smaller groups of size 10, 30 and 50, respectively, so that the range in total sample size was similar to Study 1.

The results for the main effects were consistent with those of Study 1. Therefore, we only report our findings concerning the interaction effect, that is, the Type I error rates if $f_{AB} = 0$ ($\alpha = 5\%$) and the power to detect nonzero interaction effects, given that the main effects were both zero. The test of the interaction effect involved 4 parameters in the 3×3 design and 16 parameters in the 5×5 design. Larger designs were not considered because they are rarely found in practice.

Figure 7 shows the Type I error rates obtained from the different pooling methods and LD when all effects are zero, and $\rho_{xy} = 0$ as well as $\lambda = 0$. For moderate ($n = 30$) and larger groups ($n = 50$) all methods were found to be robust. As in Study 1, the Type I error rates of D_1 and D_2 were slightly above those of D_1^* and D_3. For smaller groups ($n = 10$), D_1 and D_2 were found to be somewhat liberal in the 3×3 design (5.7% and 5.8%) and slightly conservative in the 5×5 design (4.8% and 4.1%), whereas D_1^* and D_3 performed conservatively in both cases (4.9% and 4.0% in the 3×3 design; 4.1% and 2.8% in the 5×5 design, respectively).

Similar differences were observed for the power to detect nonzero interaction effects, as is shown in Table 3. For smaller groups ($n = 10$, $I = J = 5$), D_1 and D_2 had greater power to detect nonzero interaction effects, whereas D_1^* and especially D_3 were less powerful; a pattern that was most pronounced if the covariate X did not provide information about the missing data ($\rho_{xy} = 0$). For moderate groups ($n = 30$, $I = J = 3$), the differences between the pooling methods were much

smaller. In comparison with LD, and in larger samples, the power of the pooling methods was low if the covariate did not provide information about the missing values ($\rho_{xy} = 0$), but higher than with LD if the covariate was predictive of the missing data ($\rho_{xy} = .70$). In smaller samples, such low power was only observed for D_3. The missing data mechanism did not influence the power obtained with the pooling methods, but LD showed lower power if Y was MAR ($\lambda = .70$).

General Discussion

By means of Monte Carlo simulation, we examined the performance of different pooling methods for the global null hypothesis test of the ANOVA with multiply imputed datasets. The goal of the present article was to complement the existing literature with simulation results that argue from the perspective of applied researchers. Similar to previous studies, we can conclude that D_1 and D_1^* are the most reliable pooling methods available and that D_3 behaves similarly in larger samples. However, we found that the use of D_2, at least for hypothesis tests in the ANOVA, is perfectly supported by many conditions that commonly occur in research practice. All pooling methods provided large potential gains over LD in terms of statistical power when a useful covariate could be included in the imputation model, provided that the number of imputations was sufficiently large. Whereas the increase in statistical power depended on the presence of useful covariate information, there was usually no harm in using MI when the covariate did not provide any information at all. We hope that the simulation approach taken in this study will aid researchers in judging the severity of the missing data problem and in choosing the procedure which is the most fitting for their purpose.

In general, D_1 and D_1^* provided the most reliable hypothesis tests for the ANOVA, which replicates what previous studies concluded about D_1. Their Type I error rates varied within a

small range around the optimal value, and reasonable gains in statistical power arose from including auxiliary variables. The slightly liberal behavior of D_1 was limited to small samples. Both methods appeared to be reliable even when large portions of data were missing.

The D_2 procedure performed well in Study 1 and Study 3. We observed similar gains in statistical power for D_1 and D_2, but the power of D_2 was much lower if the number of imputed datasets was not large enough. However, unlike previous research suggested, the power of D_2 improved drastically when the number of imputations was increased and was ultimately equal to that of D_1 (cf. Schafer, 1997; van Buuren, 2012). In line with previous research, we found that when the FMI was large, the robustness of D_2 was compromised (see Li, Meng, et al., 1991). Our simulation study suggests that researchers should refrain from using D_2 if large portions of the data are missing and no auxiliary variables can be included to compensate for the loss of information (e.g., 50% missing data, low correlation with other variables); the more information is supplied by covariates, the more missing data may be tolerated by D_2. All in all, the D_2 statistic appeared to be a reasonable choice for most applications of the ANOVA in psychological research. This is an encouraging result for applied researchers because D_2 is very easy to calculate using the test statistics alone, without requiring specialized software or programming experience.

The likelihood-based D_3 procedure performed well in most conditions, but it was quite conservative unless the samples were very large. This behavior was intensified if large portions of the data were missing. In general, the D_3 statistic can be recommended; however, at least for hypothesis tests in the ANOVA, larger gains in statistical power can be obtained using D_1, D_1^* and D_2, which are often easier to implement.

Our results also have important implications for applications of MI in which large portions of the data are missing, for example, in “planned missing data” designs (Graham et al., 2006). In

such designs, both MI and LD provide approximately unbiased parameter estimates because the data are MCAR. However, based on our results, it seems crucial that "planned missing data" designs include auxiliary variables which are correlated with the variables of interested, thus providing more informed imputations of missing values. Otherwise, analyses based on MI will be no more efficient than those based on LD (see Rhemtulla, Savalei, & Little, 2016). In other words, hypothesis tests based on MI can be much more powerful than those based on LD, but only if useful covariates are available that can be included in the imputation model.

As is true for any computer simulation, our study was limited in a number of ways. First, the complex simulation design limited the number of levels that could be studied for each factor in a fully crossed manner. In Study 1, we fixed the probability of missing data to 25%, and most other factors had a small number of levels. We addressed this problem by varying some simulation factors in two additional studies to explore their effects in better detail. Nonetheless, not all conditions were fully crossed, and therefore our results should not be generalized too readily to the vast diversity of conditions that can occur in practical research. Second, likelihood-based methods may be considered, which offer some advantages over LD, for example, to include auxiliary variables or to condition on possible causes of missing data (see Enders, 2010; Little & Rubin, 2002; von Hippel, 2007). Third, we assumed the covariate and the grouping variable to be fully observed at all times. This is often unlikely in practice, in which case, more general missing data methods must be considered (e.g., Enders, 2008; Little, 1992). Even though imputation itself was of minor interest in our study, results may differ for multivariate missing data problems. Finally, there are further alternatives to the four pooling methods considered here and they should be subjects of future research. Raghunathan and Dong (2011) proposed a pooling method which is solely based on the sum of squares. Variations and applications of D_1 and D_3 have been

considered by Licht (2010), Kientoff (2011), and Consentino and Claeskens (2010).

In future studies, researchers may wish to address ANOVA designs with multiple or nested factors, interaction effects, repeated measurements, or random effects (see van Ginkel & Kroonenberg, 2014). Effect size measures should be considered to allow for a more exhaustive treatment of missing data in ANOVA designs (Harel, 2009). However, the procedures featured in our study are not limited to the ANOVA. In structural equation modeling, researchers may utilize the same procedures that are featured in our study for various multiparameter tests with multiply imputed data (see Enders, 2010, for an overview). We believe that all pooling methods have good potential for reliable and efficient statistical inference when faced with missing data. The computer code for these methods is provided in the supplemental online material. We encourage researchers to use and extend these methods to thereby promote a wider application of missing data methods in psychology and the social sciences.

References

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Asparouhov, T., & Muthén, B. O. (2008). *Chi-square statistics with multiple imputation* (Technical Appendix).

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 651–675. doi: 10.1080/10705510802339072

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. Hoboken, NJ: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). New York, NY: Routledge.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351. doi: 10.1037/1082-989X.6.4.330

Consentino, F., & Claeskens, G. (2010). Order selection tests with multiply imputed data. *Computational Statistics & Data Analysis*, *54*, 2284–2295. doi: 10.1016/j.csda.2010.04.009

Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information

maximum likelihood-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 434–448. doi: 10.1080/10705510802154307

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Graham, J. W. (2012). *Missing data: Analysis and design*. New York, NY: Springer.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*, 206–213. doi: 10.1007/s11121-007-0070-9

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*, 323–343. doi: 10.1037/1082-989X.11.4.323

Grund, S., Robitzsch, A., & Lüdtke, O. (2016). mitml: Tools for multiple imputation in multilevel modeling (Version 0.3-1) [Computer software].

Harel, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*, *4*, 75–89. doi: 10.1016/j.stamet.2006.03.002

Harel, O. (2009). The estimation of R^2 and adjusted R^2 in incomplete data sets using multiple imputation. *Journal of Applied Statistics*, *36*, 1109–1118. doi: 10.1080/02664760802553000

Kientoff, C. J. (2011). *Development of weighted model fit indexes for structural equation models using multiple imputation* (Doctoral dissertation). Iowa State University.

Li, K.-H., Meng, X.-L., Raghunathan, T. E., & Rubin, D. B. (1991). Significance levels from repeated *p*-values with multiply-imputed data. *Statistica Sinica*, *1*, 65–92.

Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an *F* reference distribution.

Journal of the American Statistical Association, *86*, 1065–1073. doi: 10.1080/01621459.1991.10475152

Licht, C. (2010). *New methods for generating significance levels from multiply-imputed data* (Doctoral dissertation). Universität Bamberg.

Little, R. J. A. (1992). Regression with missing *X*'s: A review. *Journal of the American Statistical Association*, *87*, 1227–1237.

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

Marshall, A., Altman, D. G., Holder, R. L., & Royston, P. (2009). Combining estimates of interest in prognostic modelling studies after multiple imputation: Current practice and guidelines. *BMC Medical Research Methodology*, *9*, 57. doi: 10.1186/1471-2288-9-57

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). New York, NY: Psychology Press.

Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, *79*, 103–111. doi: 10.1093/biomet/79.1.103

Mistler, S. A. (2013). A SAS macro for computing pooled likelihood ratio tests with multiply imputed data. In *Proceedings of the SAS Global Forum.*

Pornprasertmanit, S. (2014). semTools: Useful tools for structural equation modeling (Version 0.4-6) [Computer software].

R Core Team. (2014). R: A language and environment for statistical computing (Version 3.1.2) [Computer software].

Raghunathan, T., & Dong, Q. (2011). *Analysis of variance from multiply imputed data sets*. Unpublished manuscript, University of Michigan, Ann Arbor, MI.

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika*, *94*, 502–508. doi: 10.1093/biomet/asm028

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, *102*, 1462–1471. doi: 10.1198/016214507000000932

Rhemtulla, M., Savalei, V., & Little, T. D. (2016). On the asymptotic relative efficiency of planned missingness designs. *Psychometrika*, *81*, 60–89. doi: 10.1007/s11336-014-9422-0

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592. doi: 10.1093/biomet/63.3.581

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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: CRC Press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. doi: 10.1037//1082-989X.7.2.147

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

van Ginkel, J. R., & Kroonenberg, P. M. (2014). Analysis of variance of multiply imputed data. *Multivariate Behavioral Research*, *49*, 78–91. doi: 10.1080/00273171.2013.855890

von Hippel, P. T. (2007). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, *37*, 83–117. doi:

10.1111/j.1467-9531.2007.00180.x

Table 1

Simulation Design of the Different Simulation Studies

Design conditions	Study 1	Study 2a	Study 2b	Study 3
Group size (n)	25, 50, 100	25	25, 100	10, 30, 50
Levels of A and B (I, J)	3, 6, 12	12	12	3×3, 5×5
Main effect A (f_A)	0, .10, .25, .40	.25	0	0, .10, .25
Main effect B (f_B)	–	–	–	0
Interaction effect (f_{AB})	–	–	–	0, .10, .25
Effect patterns	difference, trend	difference	difference	difference
Correlation XY (ρ_{xy})	0, .35, .70	0, .05, . . . , .95	0, .20, .35, .50, .70, .90	0, .35, .70
MD effect of X (λ)	0, .35, .70	0	0	0, .35, .70
MD probability	25%	25%, 50%	5%, 10%, . . . , 80%	25%
Number of Imputations	5, 10, 20, 50, 100	100	100	5, 10, 20, 50, 100

Note. The correlation ρ_{xy} and the MD probability were varied in steps of .05 and 5% in Studies 2a and 2b, respectively. MD = missing data.

Table 2

Power to Detect Nonzero Effects (α=5%) for all Pooling Methods and LD

	λ = 0					λ = .70				
	LD	D_1	D_1^*	D_2	D_3	LD	D_1	D_1^*	D_2	D_3
	n = 25, I = 12, f_A = .25 (PE = .836)									
ρ_{xy} = 0	.683	.687	.669	.692	.658	.675	.680	.664	.685	.649
ρ_{xy} = .35	.675	.697	.682	.702	.674	.662	.698	.679	.711	.668
ρ_{xy} = .70	.677	.761	.747	.756	.749	.630	.758	.744	.762	.745
	n = 50, I = 3, f_A = .25 (PE = .780)									
ρ_{xy} = 0	.644	.646	.635	.649	.634	.645	.646	.635	.648	.631
ρ_{xy} = .35	.649	.671	.660	.668	.658	.631	.654	.644	.660	.640
ρ_{xy} = .70	.640	.704	.694	.704	.695	.611	.713	.704	.720	.703

Note. PE = power expected; n = group size; I = number of groups; f_A = size of main effect A; ρ_{xy} = correlation between X and Y; λ = effect of X on missingness; D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

Table 3

Power to Detect Nonzero Interaction Effect (α=5%) in a Two-Factorial Design for all Pooling Methods and LD

	λ = 0					λ = .70				
	LD	D_1	D_1^*	D_2	D_3	LD	D_1	D_1^*	D_2	D_3
	$n = 10$, $I = J = 5$, $f_{AB} = .25$ (PE = .653)									
$\rho_{xy} = 0$	.489	.447	.415	.427	.355	.488	.438	.407	.435	.340
$\rho_{xy} = .35$	.488	.464	.430	.447	.394	.476	.448	.415	.458	.362
$\rho_{xy} = .70$	.474	.530	.500	.504	.508	.470	.543	.505	.546	.516
	$n = 30$, $I = J = 3$, $f_{AB} = .25$ (PE = .922)									
$\rho_{xy} = 0$	.803	.805	.802	.808	.796	.817	.815	.809	.822	.800
$\rho_{xy} = .35$	.807	.818	.809	.814	.806	.800	.817	.810	.820	.806
$\rho_{xy} = .70$	.797	.870	.863	.871	.866	.791	.880	.877	.888	.878

Note. PE = power expected; n = group size; I = number of groups by factor A; J = number of groups by factor B; f_{AB} = size of interaction effect; ρ_{xy} = correlation between X and Y; λ = effect of X on missingness; D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

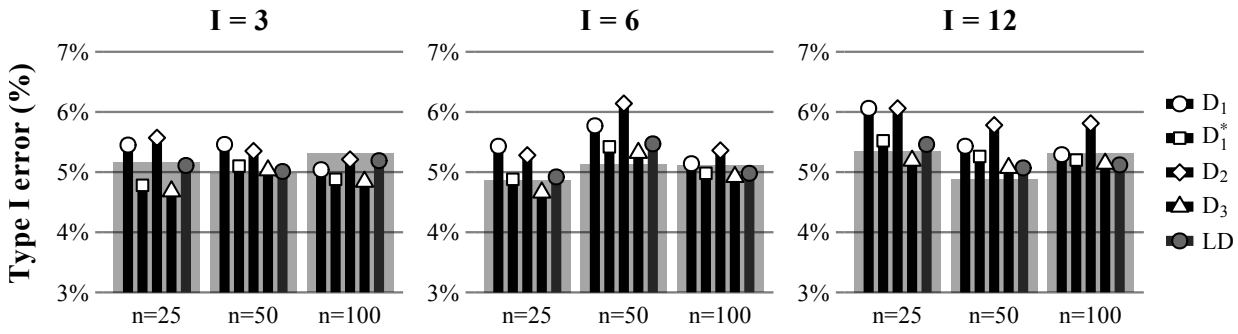

Figure 1. Type I error rates for different pooling methods and LD ($\alpha = 5\%$) depending on group size (n) and number of groups (I), given MCAR data ($\lambda = 0$) with no auxiliary information ($\rho_{xy} = 0$). The grey boxes indicate the Type I error rates obtained from complete datasets. D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

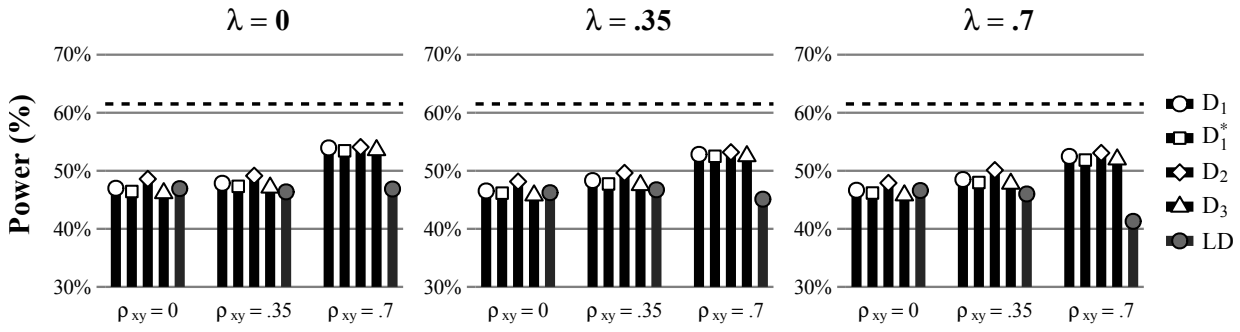

Figure 2. Power to detect nonzero main effect ($f_A = .10$) in larger samples ($n = 100$, $I = 12$) depending on the missing data mechanism. The expected power is indicated by a dashed line. ρ_{xy} = correlation between X and Y; λ = effect of X on missingness; D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

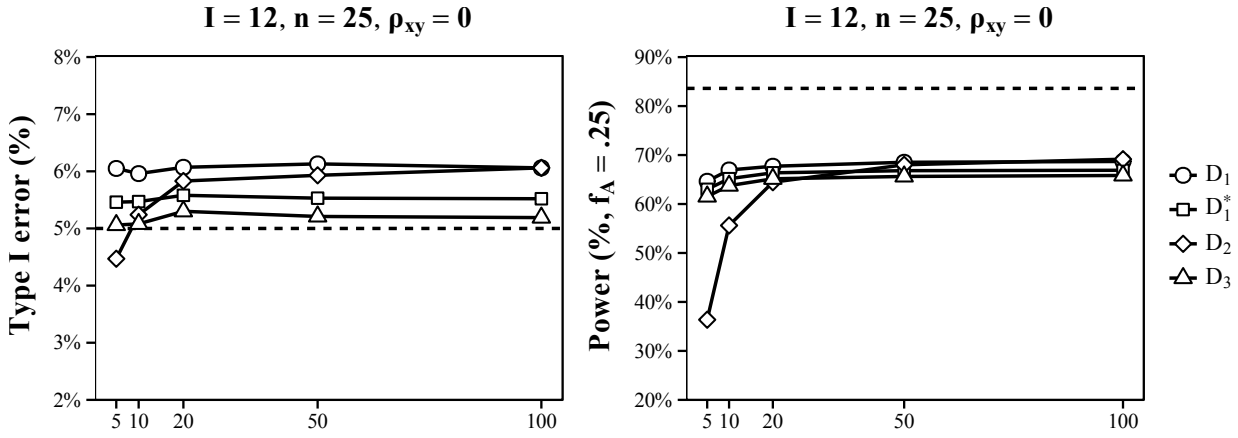

Figure 3. Type I error and statistical power of all pooling methods depending on the numbers of imputations. I = number of groups; n = group size; ρ_{xy} = correlation between X and Y; f_A = size of main effect A; D_1, D_1^*, D_2, D_3 = pooling methods.

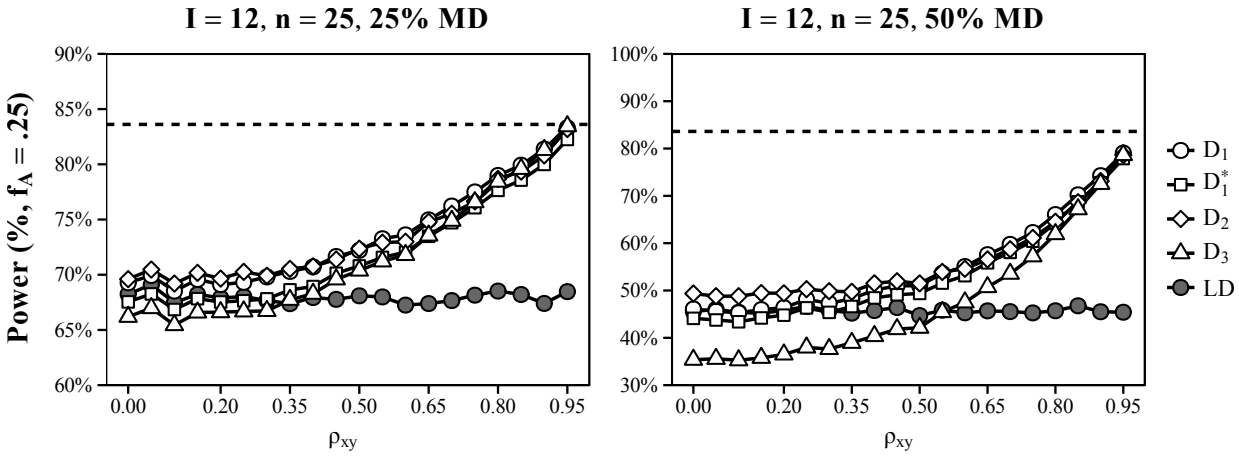

Figure 4. Power to detect main effect for all pooling methods and LD depending on the correlation between X and Y and the amount of missing data. The expected power is indicated by a dashed line. I = number of groups; n = group size; f_A = size of main effect A; ρ_{xy} = correlation between X and Y; D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

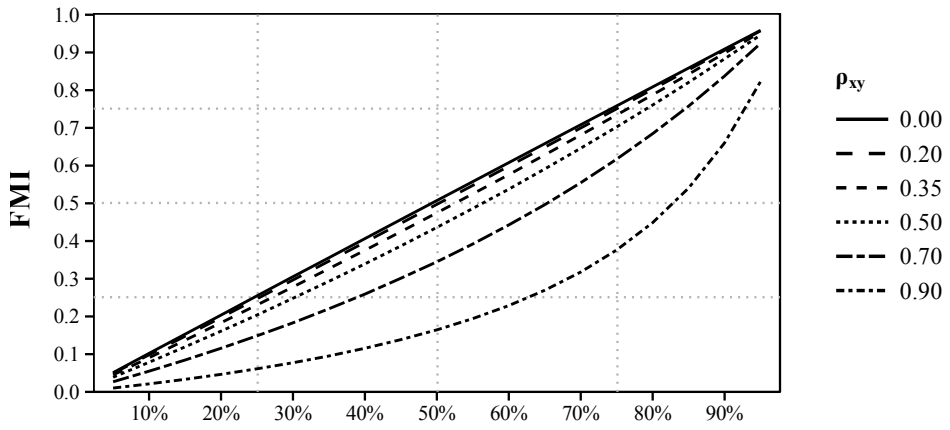

Figure 5. Estimates of the FMI obtained from D_1 in larger samples (n = 100, I = 12) with zero main effect (f_A = 0) depending on the amount of missing data and the correlation between X and Y. FMI = fraction of missing information; ρ_{xy} = correlation between X and Y.

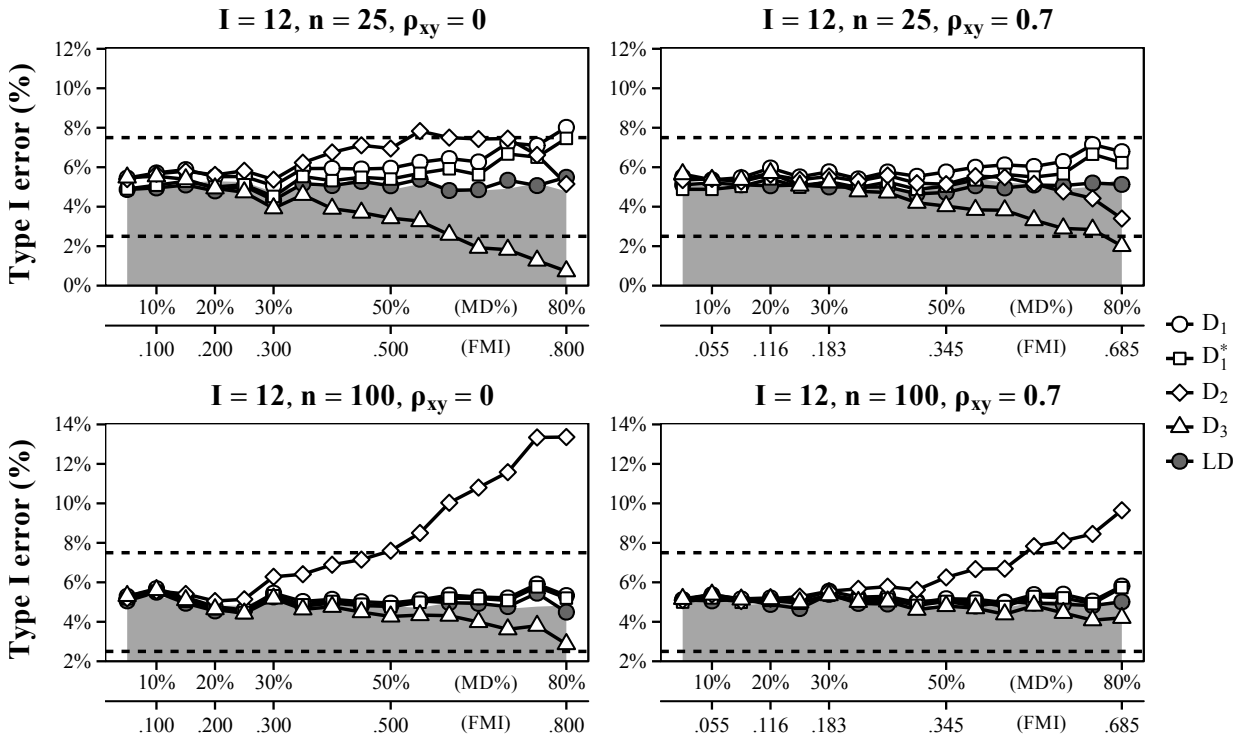

Figure 6. Type I error rates of all pooling methods and LD in moderate and larger samples dependent on the amount of missing data and the correlation between X and Y. The grey area indicates the Type I error obtained from complete datasets. I = number of groups; n = group size; ρ_{xy} = correlation between X and Y; FMI = fraction of missing information; D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.

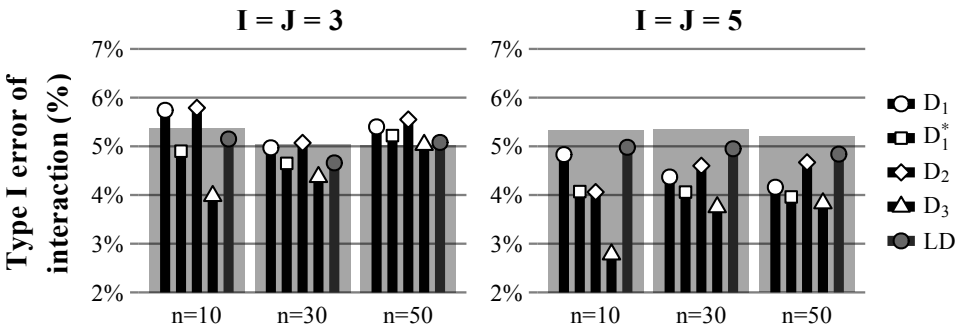

Figure 7. Type I error rates for different pooling methods and LD (α = 5%) for the interaction effect in the two-factorial design, depending on group size (n) and number of groups per factor ($I = J$). The grey boxes indicate the Type I error rates obtained from complete datasets. D_1, D_1^*, D_2, D_3 = pooling methods; LD = listwise deletion.