Multiple imputation to balance unbalanced designs for two-way analysis of variance

Ginkel, J.R. van; Kroonenberg, P.M.

METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Multiple Imputation to Balance Unbalanced Designs for Two-Way Analysis of Variance

Joost R. van Ginkel[a], Pieter M. Kroonenberg[b]

[a] *Department of Psychology, Leiden University, Leiden, The Netherlands.* **[b]** *Department of Child and Family Studies, Leiden University, Leiden, The Netherlands.*

Methodology, 2021, Vol. 17(1), 39–57, https://doi.org/10.5964/meth.6085

Received: 2021-02-04 • **Accepted:** 2021-03-16 • **Published (VoR):** 2021-03-31

Corresponding Author: Joost R. van Ginkel, Methodology and Statistics, Department of Psychology, Leiden University, PO Box 9500, 2300 RB Leiden, The Netherlands. Tel.: +31(0)71 527 3620, E-mail: jginkel@fsw.leidenuniv.nl

Supplementary Materials: Materials [see Index of Supplementary Materials]

Abstract

A balanced ANOVA design provides an unambiguous interpretation of the F-tests, and has more power than an unbalanced design. In earlier literature, multiple imputation was proposed to create balance in unbalanced designs, as an alternative to Type-III sum of squares. In the current simulation study we studied four pooled statistics for multiple imputation, namely D_0, D_1, D_2, and D_3 in unbalanced data, and compared them with Type-III sum of squares. Statistics D_1 and D_2 generally performed best regarding Type-I error rates, and had power rates closest to that of Type-III sum of squares. Additionally, for the interaction, D_1 produced power rates higher than Type-III sum of squares. For multiply imputed datasets D_1 and D_2 may be the best methods for pooling the results in multiply imputed datasets, and for unbalanced data, D_1 might be a good alternative to Type-III sum of squares regarding the interaction.

Keywords

unbalanced designs, multiple imputation, two-way analysis of variance, missing data, type-III sum of squares

In an experiment where two-way analysis of variance is the intended analysis, unforeseen circumstances may occur which may cause the design to be unbalanced. Unbalanced data may also occur in non-experimental research when group sizes are unequal by themselves. One important consequence of imbalance is that due to the resulting multicollinearity *F*-tests in ANOVA have less power than in balanced designs (e.g., Shaw & Mitchell-Olds, 1993, p. 1643). One of the most widely used methods for calculating *F*-tests in unbalanced data is Type-III sum of squares. This method has several advantages over

other types of sum of squares, such as Type-I and Type-II. For example, in Type-III sum of squares, the *F*-values are not influenced by the order in which each effect is entered in the ANOVA, as in Type-I sum of squares. For a discussion of the advantages of Type-III sum of squares over the other types, see Maxwell & Delaney (2004). However, despite these advantages, the power of the *F*-tests in Type-III sum of squares is still lower than in balanced designs.

According to Schafer (1997, p. 21) an unbalanced design can be considered a missing-data problem by imagining a number of additional cases with missing data on the outcome variable, which in the appropriate conditions would result in a balanced design (also, see Shaw & Mitchell-Olds, 1993, p. 1640, and Winer, Brown, & Michels, 1991, pp. 479-481). Considering an unbalanced design a missing-data problem creates new ways to handle the problems that come with unbalanced designs by using methods for dealing with missing data. Shaw and Mitchell-Olds (1993, p. 1641) discuss imputation (i.e., estimation) of the missing data to balance the design. They point out that imputation will give unbiased parameter estimates of the ANOVA model, but biased significance tests. This bias is caused by the fact that the imputed values are treated as observed values, and do not incorporate uncertainty of the missing values in the analysis (e.g., Schafer, 1997, p. 2). Thus, a method for estimating the missing values is needed that does take this uncertainty into account. One such method is *multiple imputation* (Rubin, 2004; Van Buuren, 2012). In multiple imputation, 1) missing values are estimated multiple times (*M*), resulting in *M* different complete versions of the incomplete dataset. Next, 2) these complete datasets are analyzed using the statistical analysis of interest, resulting in *M* outcomes of the same analysis, and 3) the results are combined into one pooled analysis, such that the uncertainty about the missing values is taken into account. For this pooling, formulas are used which will henceforth be called *combination rules*.

For balancing unbalanced data, multiple imputation may be used as follows. First, by adding a number of additional cases to specific groups such that all groups have equal size, the dataset is now balanced where in some cells cases have missing data on the outcome variable. These missing data are then multiply imputed using factors A and B as categorical predictors of the missing data on the outcome variable. Procedures for how to generate multiply imputed values for the missing data are described in, for example, Van Buuren (2012, Chapter 3) and Raghunathan, Lepkowski, Van Hoewyk, and Solenberger (2001).

Once multiply imputed datasets have been obtained, the ANOVAs can be applied to these datasets, and the results can be combined using specific combination rules. However, Van Ginkel and Kroonenberg (2014) argue that the *M F*-tests of an effect in ANOVA in *M* multiply imputed datasets cannot be combined directly, because for the combination rules that are suited for ANOVA (to be discussed shortly) one needs *M* sets of regression coefficients and *M* covariance matrices of these regression coefficients. See, Rubin (2004, pp. 79-81) and Van Ginkel (2019) who show how to calculate a pooled *F*-sta-

tistic testing several regression coefficients for significance simultaneously for multiply imputed datasets.

However, one can reformulate the two-way ANOVA model as a regression model, so that the combination rules can be applied to the regression coefficients. Van Ginkel and Kroonenberg (2014) show that this can be done by recoding the categorical variables and their interaction into effect-coded variables, and using these effect-coded variables as predictors in the regression model (also, see Winer et al., 1991, pp. 959-963). Once estimates of the regression coefficients and their covariances have been obtained for each of the M imputed datasets, the results may be combined.

For pooling the results of two-way ANOVA Rubin (2004) discusses four methods that may be good candidates. Schafer (1997) called three of these methods D_1, D_2, and D_3. The fourth method was not given a name, probably because Schafer noted some problems of this method, which made him set it aside altogether. However, for the current application of these combination rules, the statistic that was not given a name may still be useful, for reasons discussed later on. Because the method with no name was chronologically the first method in Schafer's book, we will call this statistic D_0. Each of these methods are discussed next.

Statistic D_0

Let $\widehat{\mathbf{Q}}_m$ be a vector of length p, containing p parameter estimates of a statistical model, and let $\mathbf{U}_m$ be a covariance matrix of the p parameter estimates, in imputed dataset m. A pooled set of parameter estimates across M imputed datasets is computed as:

$$\overline{\mathbf{Q}} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbf{Q}}_m. \tag{1}$$

The pooled covariance matrix has two components, namely a within-imputation covariance matrix $\overline{\mathbf{U}}$, and a between-imputation covariance matrix $\mathbf{B}$:

$$\overline{\mathbf{U}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{U}_m, \tag{2}$$

$$\mathbf{B} = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\mathbf{Q}}_m - \overline{\mathbf{Q}}\right)\left(\widehat{\mathbf{Q}}_m - \overline{\mathbf{Q}}\right)'. \tag{3}$$

The total covariance matrix of the estimate $\overline{\mathbf{Q}}$ is computed as

$$\mathbf{T} = \overline{\mathbf{U}} + \left(1 + M^{-1}\right)\mathbf{B}. \tag{4}$$

To test all p parameter estimates in $\overline{\mathbf{Q}}$, statistic D_0 (Rubin, 2004, pp. 77-78) is computed as:

$$D_0 = \overline{\mathbf{Q}}'\mathbf{T}^{-1}\overline{\mathbf{Q}}/p, \tag{5}$$

which has an approximate F-distribution with p model degrees of freedom, and $(1 + p^{-1})\hat{v}/2$ error degrees of freedom, where an approximation for $\hat{v}$ by Reiter (2007) may be used. The formula for $\hat{v}$ is rather long and complex, so we will not give it here. For computational details, see Reiter (2007).

A pooled F-test for the effect of factor A may be computed by filling in the pooled regression coefficients of the effect coded variables that together form factor A in $\widehat{\mathbf{Q}}_m$ in Equation 1 and Equation 3, filling in the covariances of these coefficients in Equation 2, and use the resulting $\overline{\mathbf{Q}}$, $\overline{\mathbf{U}}$ and $\mathbf{B}$ in Equation 4 and Equation 5. Similarly, F-tests for factor B and the interaction may be computed. Computational details on how to obtain the covariance matrix $\mathbf{U}_m$ for the effects of factors A, B and the interaction are given in Van Ginkel and Kroonenberg (2014).

Statistic D_1

Because $\mathbf{B}$ is a noisy estimate of the between-imputation variance and does not even have full rank if $M < p$, Schafer (1997, p. 113) advises against using statistic D_0. Rather than using $\mathbf{T}$ as an estimate for the total covariance matrix, a more stable total covariance matrix is:

$$\widetilde{\mathbf{T}} = (1 + r_1)\overline{\mathbf{U}} \tag{6}$$

$$r_1 = (1 + M^{-1})tr\left(\mathbf{B}\overline{\mathbf{U}}^{-1}\right)/p$$

where r_1 is the *relative increase in variance due to nonresponse*. The resulting F-value testing all elements in vector $\overline{\mathbf{Q}}$ for significance simultaneously is:

$$D_1 = \overline{\mathbf{Q}}'\widetilde{\mathbf{T}}^{-1}\overline{\mathbf{Q}}/p, \tag{7}$$

which, under the assumption that r_1 is equal across all elements of $\overline{\mathbf{Q}}$, has an approximate F-distribution with p numerator degrees of freedom, and $\hat{v}$ denominator degrees of freedom.

Despite the assumption of equal r_1 across all elements of $\overline{\mathbf{Q}}$, Li, Raghunathan and Rubin (1991b) showed that violation of this assumption does not substantially influence the Type-I error rate of D_1. Because of this finding, and because of the disadvantages of statistic D_0, D_1 has been widely implemented in different statistical packages, such as SAS (SAS, 2013; Yuan, 2011), Stata (StataCorp, 2017), the `miceadds` package in R (Robitzsch, Grund, & Henke, 2017), and an SPSS (IBM SPSS, 2017) macro by Van Ginkel (2010). Using the procedure outlined by Van Ginkel and Kroonenberg (2014), Grund et al.

(2016) showed that D_1 generally produces Type-I error rates close to the theoretical α in both one-way and two-way ANOVAs.

Statistic D_2

Define

$$d_{W,m} = \widehat{\mathbf{Q}}_m' \mathbf{U}_m^{-1} \widehat{\mathbf{Q}}_m \tag{8}$$

as the Wald statistic of imputed dataset m,

$$\bar{d}_W = \frac{1}{M} \sum_{m=1}^{M} d_{W,m} \tag{9}$$

as the average Wald statistic across imputed datasets, and

$$r_2 = \left(1 + M^{-1}\right) \left[\frac{1}{M-1} \sum_{m=1}^{M} \left(\sqrt{d_{W,m}} - \overline{\sqrt{d_W}} \right)^2 \right] \tag{10}$$

as an alternative estimate of the relative increase in variance due to nonresponse. Statistic D_2 is given by:

$$D_2 = \frac{\bar{d}_W p^{-1} - (M+1)(M-1)^{-1} r_2}{1 + r_2} \tag{11}$$

As a reference distribution for D_2 an F-distribution with p numerator degrees of freedom and $\nu_2 = p^{-\frac{3}{m}}(m-1)\left(1 + r_2^{-1}\right)^2$ denominator degrees of freedom is used. Applied to two-way ANOVA, significance tests for factors A, B, and the interaction can be obtained by substituting the relevant coefficients for $\widehat{\mathbf{Q}}_m$, and by substituting the covariance matrices of the relevant coefficients for $\mathbf{U}_m$, and substitute these quantities in Equation 8.

An advantage of D_2 is that it is easier to compute than D_0 and D_1 as it can be calculated from M separate Wald statistics alone and no separate combining of the covariance matrices is needed (Schafer, 1997, p. 115). Li, Meng, Raghunathan, and Rubin (1991a) however, show that D_2 occasionally produces Type-I error rates as high as 8% when using α = .05. Li et al. (1991a) therefore argue that D_2 only be used as a rough guide. However, Grund et al. (2016) found Type-I error rates for D_2 that were closer to the theoretical 5% than found by Li et al. (1991a). Thus, how well D_2 performs regarding Type-I error rates in which situations, remains unclear.

Statistic D_3

Statistic D_3 (Meng & Rubin, 1992) is a combined likelihood-ratio test of M likelihood-ratio statistics from M imputed datasets. Define $d_{L,m}$ as a likelihood-ratio test of imputed dataset m, and

$$\bar{d}_L = \frac{1}{M}\sum_{m=1}^{M} d_{L,m} \tag{12}$$

as the average $d_{L,m}$ across imputed datasets. Next, define $d^*_{L,m}$ as a likelihood-ratio test of imputed dataset m but now evaluated at the average parameter estimates of the model across M imputed datasets. The average $d^*_{L,m}$ is computed as

$$\tilde{d}_L = \frac{1}{M}\sum_{m=1}^{M} d^*_{L,m} \tag{13}$$

Statistic D_3 is computed as

$$D_3 = \frac{\bar{d}_L}{p(1+r_3)} \tag{14}$$

$$r_3 = \frac{M+1}{p(M-1)}\left(\bar{d}_L - \tilde{d}_L\right)$$

The reference distribution that is used for D_3 is an F-distribution with p numerator degrees of freedom and a denominator degrees of freedom given by:

$$\begin{aligned} \nu_3 &= 4 + (t-4)\left[1 + \left(1 - 2t^{-1}\right)r_3^{-1}\right]^2 \quad \text{if } t = p(M-1) > 4 \\ \nu_3 &= t\left(1 + p^{-1}\right)\left(1 + r_3^{-1}\right)^2/2 \qquad\qquad \text{otherwise.} \end{aligned} \tag{15}$$

For a specific effect in the model (factor A, factor B, interaction) D_3 is computed by letting $d_{L,m}$ and $d^*_{L,m}$ be the likelihood-ratio tests that test the full model against the full model, minus the terms that together form the specific effect.

Goal of the Current Study

Several authors (Schafer, 1997; Shaw & Mitchell-Olds, 1993; Winer et al., 1991) have pointed out that unbalanced designs can be considered a missing-data problem. Shaw and Mitchell-Olds (1993) discussed imputation as a possible solution for imbalance and at the same time noted some problems of the imputation approach regarding p-values. However, they did not mention the use of multiple imputation to overcome the problems of single imputation in unbalanced designs. Although Schafer (1997) never explicitly

suggested to balance an unbalanced design using multiple imputation, the suggestion implicitly followed from the fact that Schafer's book is about multiple imputation. Van Ginkel and Kroonenberg (2014, p. 79) made the suggestion explicit, and used it in an empirical data example (pp. 87-88). The reasoning behind their suggestion was that balanced designs generally have more power than unbalanced designs, and that consequently multiple imputation in unbalanced designs could increase power.

However, Van Ginkel and Kroonenberg (2014) did not carry out any simulations to back up this claim. Additionally, one aspect that Van Ginkel and Kroonenberg did not include in their reasoning is that multiple imputation does not actually add information that could increase power. Indeed multiple imputation simulates additional cases, but imputing M values of the outcome variable for each additional case creates additional variation in the parameter estimates of the ANOVA model as well, represented by covariance matrix $\mathbf{B}$ or the relative increase in variance due to nonresponse (r_1, r_2, and r_3). The statistics D_0 to D_3 include these quantities in their calculation such it lowers their power, which may consequently undo the increased power as a result of creating additional cases. Also, Van Buuren (2012, p. 48) and Von Hippel (2007) argue that when missing data only occur on the outcome variable, analyzing only the complete cases may be preferred over multiple imputation. Since unbalanced data can be considered a situation with missings on only the outcome variable, and that Type-III sum of squares in this context is equivalent to analyzing only complete cases, it would follow that in unbalanced data Type-III sum of squares is preferred over multiple imputation.

In short, both valid arguments for balancing unbalanced data using multiple imputation prior to two-way ANOVA, and simulation studies that confirm its usefulness seem to be lacking. However, the fact that this suggestion has been made in the literature or even just the fact that unbalanced data are often described as a missing-data problem and that multiple imputation is a highly recommended procedures for dealing with missing data, calls for a simulation study to investigate the usefulness of this suggestion. In the current paper we will carry out such a simulation study. Consequently, the first research question is whether there is some benefit in using multiple imputation for balancing an unbalanced design prior to a two-way ANOVA after all.

Furthermore, Grund et al. (2016) already investigated statistics D_1, D_2, and D_3 in a wide variety of situations. In their study they found that in most situations all statistics produced Type-I error rates close to the theoretical 5%. However, in more extreme situations (e.g., small sample size, high percentages of missing data), D_2, and D_3 were somewhat conservative and had lower power than D_1. Based on this study, D_1 is the preferred statistic for combining the results of multiply imputed data in ANOVA.

However, the question is to what extent the results by Grund et al. (2016) generalize to a situation where unbalanced data are balanced using multiple imputation. D_1 showed the most promising results but has one potential weak point, and which may especially be problematic in unbalanced data: it assumes that the fraction of missing information

is equal across the p parameter estimates in $\overline{Q}$. Unequal fractions of missing information across parameter estimates in $\overline{Q}$ are inherently related to unbalanced designs as each regression coefficient of an effect-coded variable of a specific factor represents a specific group of this factor. Unequal group sizes imply that for smaller groups more additional values of the dependent variable have to be multiply imputed than for larger groups, resulting in unequal fractions of missing information across parameter estimates.

In a situation which inherently has unequal fractions of missing information across parameter estimates, a statistic assuming equal fractions of missing information across parameter estimates (D_1) is not the most convenient choice, especially when an alternative version of this statistic exists (D_0) that does not make this assumption (when p = 1, or when the fractions of missing information across parameter estimates are equal, both statistics are equivalent). Given that D_1 generally gives the best results regarding power and Type-I error rates, and that D_0 is equivalent to D_1 but without the assumption of equal fractions of missingness across parameters, D_0 might be the ideal candidate for calculating pooled F-tests in unbalanced data when imbalance is handled using multiple imputation. Although Schafer (1997) and Li et al. (1991b) advised against the use of D_0, the argument against its use was mainly that for small M it may not perform well. However, this problem may easily be resolved by increasing M.

Furthermore, although Li et al. (1991b) showed that violation of equal fractions of missing information is not necessarily problematic for D_1, in their study the unequal fractions of missing information randomly varied across parameters of the simulation model and across replicated datasets. However, in unbalanced data unequal fractions of missing information may not always randomly vary across parameters. For example, in a field experiment it may be more difficult to collect data for one experimental condition than for the other, or in a clinical trial more people may drop out in one condition than in the other. Also, in non-experimental studies some categorical variables may have unequal groups in the population, for example, ethnicity. When drawing a random sample from this population, the different ethnic groups will have unequal sizes in the sample as well.

When fractions of missing information randomly vary across parameters, the fractions of missing information may not be equal within one replication, but the average fractions of missing information across replicated datasets are. Consequently, the negative effect of unequal fractions of missing information may cancel itself out across replications. However, in situations where the differences in fractions of missing information across parameters do not vary across replicated datasets, a statistic might be needed that allows for different fractions of missing data across parameter estimates.

Thus, a second research question is how the different pooling statistics from Grund et al. (2016) will behave in unbalanced data where the unequal fractions of missing information do not vary across replications, now also including statistic D_0. To this end, rejection rates of statistics D_0 to D_3 were studied along with the rejection rates

for Type-III sum of squares under the null- and alternative hypothesis. For comparison, rejection rates were also studied for balanced data with the same total sample size.

In the next section, the setup of the simulation study is described. In the section that follows, results of the simulation study are shown. Finally, in the discussion section conclusions will be drawn about the usefulness of multiple imputation for balancing unbalanced designs, and implications for which statistic to use will be discussed.

Method

Data were simulated according to a two-way ANOVA model in the form of a regression model with effect coded predictors. Some of the properties of the data were held constant while some were varied (discussed next). The properties that were varied resulted in several design cells. Within each design cell 2500 replications were drawn (based on studies by Harel, 2009, and Van Ginkel, 2019).

The simulations were programmed in R (R Core Team, 2018). Multiple imputation was performed using Fully Conditional Specification (FCS; e.g., Van Buuren, 2012) in the `mice` package (Van Buuren & Groothuis-Oudshoorn, 2011). For continuous variables there are two different versions of FCS, namely regression imputation (FCS-R; e.g., Little & Schenker, 1995, p. 60; Van Buuren, 2012, p. 13) and Predictive Mean Matching (FCS-PMM; Rubin, 1986; Van Buuren, 2012, pp. 68-74). The default method in `mice` is FCS-PMM because it generally preserves distributional shapes better when data are not normally distributed than FCS-R (Marshall, Altman, & Holder, 2010; Marshall, Altman, Royston, & Holder, 2010). However, for imputing an outcome variable with a normally distributed error (as in our simulation model) FCS-R may be a better alternative as it explicitly imputes values according to a normal linear model. Thus, the method for multiple imputation was FCS-R. For the imputation of outcome variable Y, factors A and B plus their interaction were used as predictors. Type-III sum of squares were computed using the `car` package (Fox & Weisberg, 2019). Statistic D_3 was calculated using the `mitml` package (Grund et al., 2016). The other statistics D_0, D_1, and D_2 were programmed by the first author.

Like in many other simulation studies, decisions regarding properties of the simulation design were to some extent arbitrary. However, prior to running the simulations, some test runs were done to see what properties would make the effects of imbalance and the differences between the different statistics most clearly visible, and which were also likely to occur in practice. The properties of the simulation design that are going to be discussed next, are mostly the result of these test runs.

Fixed Design Characteristics

The number of levels of factor A was $I = 2$. The error term of the ANOVA model was normally distributed with $\mu_\varepsilon = 0$ and $\sigma_\varepsilon = 18.37$.

Independent Variables

Number of Levels of Factor B

The number of levels of factor B was $\mathcal{J} = 3, 4, 5$.

Parameters of the Two-Way ANOVA Model

For each $\mathcal{J}$, two sets of parameters of the two-way ANOVA model were studied: one in which there were no effects of factor A, factor B, and the interaction in the population (the null model), and one in which there were effects (the alternative model).

For $\mathcal{J} = 3$ the two sets of parameter estimates were:

$\beta = \left(\beta_0, \beta_{1(A)}, \beta_{2(B)}, \beta_{3(B)}, \beta_{4(AB)}, \beta_{5(AB)}\right) = (27,0,0,0,0,0)$ and

$\beta = \left(\beta_0, \beta_{1(A)}, \beta_{2(B)}, \beta_{3(B)}, \beta_{4(AB)}, \beta_{5(AB)}\right) = (27,-1.5,-3,0,1,-0.5)$.

For $\mathcal{J} = 4$ the two sets were:

$\beta = \left(\beta_0, \beta_{1(A)}, \beta_{2(B)}, \beta_{3(B)}, \beta_{4(B)}, \beta_{5(AB)}, \beta_{6(AB)}, \beta_{7(AB)}\right) = (27,0,0,0,0,0,0,0)$ and

$\beta = \left(\beta_0, \beta_{1(A)}, \beta_{2(B)}, \beta_{3(B)}, \beta_{4(B)}, \beta_{5(AB)}, \beta_{6(AB)}, \beta_{7(AB)}\right) = (27,-1.5,-3,-1,1,1,0,-0.5)$.

Finally, for $\mathcal{J} = 5$ the two sets were:

$\beta = \left(\beta_0, \beta_{1(A)}, \beta_{2(B)}, \beta_{3(B)}, \beta_{4(B)}, \beta_{5(B)}, \beta_{6(AB)}, \beta_{7(AB)}, \beta_{8(AB)}, \beta_{9(AB)}\right) = (27,0,0,0,0,0,0,0,0,0)$ and

$\beta = \left(\beta_0, \beta_{1(A)}, \beta_{2(B)}, \beta_{3(B)}, \beta_{4(B)}, \beta_{5(B)}, \beta_{6(AB)}, \beta_{7(AB)}, \beta_{8(AB)}, \beta_{9(AB)}\right) =$
$(27,-1.5,-3,-1,0,1,1,0.25,-0.25,-0.5)$.

Sample Size

Small, medium, and large sample sizes were studied. Because $\mathcal{J}$ also affects the number of design cells of the two-way ANOVA, sample size also partly depended on the size of $\mathcal{J}$. For small N, the average cell size was 10, for medium N it was 20, and for large samples it was 30. Given these average cell sizes, sample sizes were $N = 60, 120, 180$ for $\mathcal{J} = 3$, $N = 80, 160, 240$ for $\mathcal{J} = 4$, and $N = 100, 200, 300$ for $\mathcal{J} = 5$.

Degree and Structure of Imbalance

Four different degrees of imbalance were simulated, along with balanced data, for comparison. The degree of imbalance was varied as follows: for a specific design cell the cell size was either increased or decreased by each time adding the same number to, or subtracting the same number from the original cell size in the balanced case. The increasing and decreasing of cell sizes was done such that the total sample size remained the same.

Additionally, to study whether it mattered which cells increased or decreased in size, an additional situation of imbalance was created where the cell sizes of the most severe

case of imbalance were randomly redistributed across design cells. The cell sizes for each degree of imbalance are displayed for small N in Table 1. For medium and large N the entries must be multiplied with 2 and 3 respectively.

Table 1

Cell Sizes for Different Degrees of Imbalance Under a Small Sample Size

		Imbalance				
Cell size	**Balanced**	**Small**	**Medium**	**Severe**	**Extra severe**	**Extra severe, order shuffled**
			No. levels factor B: 3			
n_{11}	10	8	6	4	2	18
n_{12}	10	10	10	10	10	10
n_{13}	10	12	14	16	18	2
n_{21}	10	11	12	13	14	6
n_{22}	10	10	10	10	10	10
n_{23}	10	9	8	7	6	14
			No. levels factor B: 4			
n_{11}	10	8	6	4	2	10
n_{12}	10	10	10	10	10	18
n_{13}	10	10	10	10	10	10
n_{14}	10	12	14	16	18	2
n_{21}	10	11	12	13	14	10
n_{22}	10	10	10	10	10	6
n_{23}	10	10	10	10	10	10
n_{24}	10	9	8	7	6	14
			No. levels factor B: 5			
n_{11}	10	8	6	4	2	10
n_{12}	10	10	10	10	10	18
n_{13}	10	10	10	10	10	10
n_{14}	10	10	10	10	10	10
n_{15}	10	12	14	16	18	2
n_{21}	10	11	12	13	14	10
n_{22}	10	10	10	10	10	6
n_{23}	10	10	10	10	10	10
n_{24}	10	10	10	10	10	10
n_{25}	10	9	8	7	6	14

Method for Handling Imbalance

Nine methods for handling imbalance were used: Type-III sum of squares, and two versions of each of the statistics D_0, D_1, D_2, and D_3. Because D_0 is argued to be sensitive to a small number of imputations, two different sizes of M were studied within each of the statistics, namely $M = 5$ and $M = 100$. This defines the eight procedures based on multiple imputation: $D_{0,M=5}$, $D_{0,M=100}$, $D_{1,M=5}$, $D_{1,M=100}$, $D_{2,M=5}$, $D_{2,M=100}$, $D_{3,M=5}$, and $D_{3,M=100}$.

Dependent Variables

For each of the F-tests in the two-way ANOVA, the number of times the null hypothesis was rejected was studied, denoted the *rejection rate*. As already noted, when $p = 1$, D_0 and D_1 are equivalent. Thus, for factor A, methods $D_{0,M=5}$, and $D_{0,M=100}$ will not be displayed.

Results

To get a rough impression of how close the Type-I error rates were to $\alpha = .05$ under the null hypothesis, it was tested whether the empirical rejection rates differed significantly from 0.05, using an Agresti and Coull (1998) confidence interval. Under the alternative hypothesis, it was tested whether the empirical rejection rates of the multiple-imputation methods differed significantly from the rejection rates of Type-III, assuming that the latter represent the "real" power. A method based on multiple imputation may be considered successful when under the null hypothesis it gives rejection rates not significantly different from 0.05, and when under the alternative hypothesis it gives rejection rates significantly higher than the rejection rates of Type-III. To avoid redundancy in discussing the results, we will mainly focus on the results of factor B and less on the results of factor A and the interaction.

Eighteen tables were needed to report all the results. Because results showed similar patterns across different J and different N, results of the remaining independent variables are only reported for $J = 3$ and a small N. Results for $J > 3$ and larger N are provided in Supplementary Materials. Table 2 shows the results for all levels of imbalance, all methods for handling imbalance, and all effects in the ANOVA, under the null model. Table 3 shows the results for all levels of imbalance, all methods for handling imbalance, and all effects in the ANOVA, under the alternative model.

Table 2

Rejection Rates for Each Effect Under the Null Model, a Small Sample Size, Three Levels of Factor B, for Different Methods for Handling Imbalance, and Different Degrees of Imbalance

		Imbalance				
Method	**Balanced**	**Small**	**Medium**	**Severe**	**Extra severe**	**Extra severe, order shuffled**
			Effect A			
Type-III	.052	.049	.051	.054	.055	.055
$D_{1,M=5}$		.049	.056	.059[a]	.061[a]	.058
$D_{1,M=100}$		.051	.051	.054	.054	.053
$D_{2,M=5}$		.047	.052	.052	.052	.054
$D_{2,M=100}$		.052	.053	.056	.060[a]	.058
$D_{3,M=5}$		.045	.034[a]	.027[a]	.019[a]	.016[a]
$D_{3,M=100}$		.044	.036[a]	.024[a]	.006[a]	.005[a]
			Effect B			
Type-III	.047	.049	.054	.049	.048	.048
$D_{0,M=5}$		.054	.062[a]	.067[a]	.077[a]	.082[a]
$D_{0,M=100}$		.053	.059[a]	.053	.053	.053
$D_{1,M=5}$		.051	.046	.054	.050	.054
$D_{1,M=100}$		.053	.058	.051	.052	.053
$D_{2,M=5}$		.050	.055	.058	.054	.057
$D_{2,M=100}$		.055	.060[a]	.054	.050	.051
$D_{3,M=5}$		.042	.034[a]	.026[a]	.024[a]	.023[a]
$D_{3,M=100}$		.048	.042	.026[a]	.014[a]	.018[a]
			Effect A × B			
Type-III	.058	.048	.054	.056	.045	.045
$D_{0,M=5}$		.055	.057	.070[a]	.083[a]	.084[a]
$D_{0,M=100}$		.056	.054	.058	.050	.048
$D_{1,M=5}$		.047	.049	.048	.053	.052
$D_{1,M=100}$		.057	.055	.056	.050	.052
$D_{2,M=5}$		.048	.050	.052	.056	.055
$D_{2,M=100}$		.058	.056	.059[a]	.046	.048
$D_{3,M=5}$		.041[a]	.031[a]	.028[a]	.021[a]	.022[a]
$D_{3,M=100}$		.048	.037[a]	.031[a]	.012[a]	.015[a]

[a]Significantly different from theoretical significance level of α = .05.

Table 3

Rejection Rates for Each Effects Under the Alternative Model, a Small Sample Size, Three Levels of Factor B, for Different Methods for Handling Imbalance, and Different Degrees of Imbalance

		Imbalance				
Method	**Balanced**	**Small**	**Medium**	**Severe**	**Extra severe**	**Extra severe, order shuffled**
			Effect A			
Type-III	.764	.742	.722	.686	.549	.549
$D_{1,M=5}$		.728	.666[a]	.585[a]	.415[a]	.406[a]
$D_{1,M=100}$		.748	.728	.680	.543	.541
$D_{2,M=5}$		.715[a]	.636[a]	.547[a]	.388[a]	.370[a]
$D_{2,M=100}$		.751	.731	.687	.558	.556
$D_{3,M=5}$		.698[a]	.546[a]	.370[a]	.140[a]	.142[a]
$D_{3,M=100}$		.732	.671[a]	.555[a]	.261[a]	.253[a]
			Effect B			
Type-III	.976	.972	.957	.922	.826	.836
$D_{0,M=5}$		.966	.934[a]	.870[a]	.729[a]	.745[a]
$D_{0,M=100}$		.976	.960	.925	.829	.838
$D_{1,M=5}$		.963[a]	.926[a]	.850[a]	.645[a]	.665[a]
$D_{1,M=100}$		.974	.964	.934[a]	.830	.837
$D_{2,M=5}$		.944[a]	.843[a]	.687[a]	.426[a]	.426[a]
$D_{2,M=100}$		.976	.961	.924	.807[a]	.818[a]
$D_{3,M=5}$		.947[a]	.861[a]	.706[a]	.384[a]	.398[a]
$D_{3,M=100}$		.972	.945[a]	.886[a]	.643[a]	.642[a]
			Effect A × B			
Type-III	.179	.176	.162	.150	.101	.148
$D_{0,M=5}$		.174	.157	.160	.131[a]	.173[a]
$D_{0,M=100}$		.189	.175	.152	.111	.157
$D_{1,M=5}$		.165	.144[a]	.126[a]	.102	.095[a]
$D_{1,M=100}$		.192	.183[a]	.169[a]	.137[a]	.120[a]
$D_{2,M=5}$		.150[a]	.125[a]	.115[a]	.075[a]	.117[a]
$D_{2,M=100}$		.192	.177	.149	.098	.164[a]
$D_{3,M=5}$		.146[a]	.106[a]	.077[a]	.050[a]	.042[a]
$D_{3,M=100}$		.180	.151	.107[a]	.049[a]	.036[a]

[a]Significantly different from Type-III, assuming Type-III is the "true" power.

Under the null hypothesis (Table 2), methods $D_{3,M=5}$, and $D_{3,M=100}$ tend to underestimate the Type-I error rate for factor B somewhat, which becomes worse as the degree of imbalance increases. This undercoverage is worst for $D_{3,M=100}$. Method $D_{0,M=5}$ tends to overestimate the Type-I error rate, which becomes worse as the degree of imbalance increases. Type-III, $D_{1,M=5}$, $D_{1,M=100}$, $D_{0,M=100}$, $D_{2,M=5}$, and $D_{2,M=100}$ produce Type-I error rates that do not differ significantly from 5% in most cases.

All methods based on $M = 5$ and method $D_{3,M=100}$ have lower power than Type-III sum of squares (Table 3). The difference in power between these methods and Type-III sum of squares becomes larger as the degree of imbalance increases. Methods $D_{0,M=100}$ and $D_{2,M=100}$ generally have power values close to that of Type-III sum of squares. Like methods $D_{0,M=100}$ and $D_{2,M=100}$, method $D_{1,M=100}$ has power values similar to that of Type-III sum for the main effects. However, for the interaction effect, $D_{1,M=100}$ has significantly higher power than Type-III sum of squares.

Finally, when the order of the cell sizes is shuffled, we only see changes in results for the interaction in the alternative model (Table 2): For methods $D_{1,M=5}$ and $D_{1,M=100}$ the power drops below the power level of Type-III sum of squares; for $D_{2,M=100}$ it raises up to a higher level than that of Type-III sum of squares.

Discussion

Van Ginkel and Kroonenberg (2014) suggested multiple imputation to increase the power of a two-way ANOVA in unbalanced designs compared to Type-III sum of squares. In the current study it was studied whether multiple imputation would indeed do this. For the main effects, methods $D_{0,M=100}$, $D_{1,M=100}$ and $D_{2,M=100}$ had power similar to those of Type-III sum of squares, but not higher power than Type-III sum of squares. However, for the interaction effect, method $D_{1,M=100}$ had higher power than Type-III sum of squares for all degrees of imbalance, except when the order of cell sizes was shuffled. In the latter case, only method $D_{2,M=100}$ had higher power than Type-III sum of squares. The other methods based on multiple imputation had lower power than Type-III sum of squares in all situations. Additionally, $D_{0,M=5}$ overestimated the Type-I error rate.

The main conclusion of this study is that there may be some benefit in doing multiple imputation for handling unbalanced data in two-way ANOVA after all. When using the appropriate statistics (D_1 and D_2) with a large number of imputations ($M = 100$), the test for the interaction may have higher power than Type-III sum of squares. However, the specific structure of imbalance (i.e. which cells have a high number of observations and which ones have a small number of observations) also seems to matter. Under all four degrees of imbalance $D_{1,M=100}$ was the methods that had highest power for the interaction. However, when the sizes of the cells were shuffled, $D_{2,M=100}$ had the highest power. Thus, it remains unclear how and when multiple imputation leads to higher power rates for the interaction in unbalanced two-way ANOVA.

Although there seems to be some benefit in multiple imputation over Type-III sum of squares, it may be wondered whether this benefit outweighs the costs. Multiple imputation is more work, the benefit seems to only concern the interaction, and it is not even entirely clear when it has higher power rates than Type-III sum of squares.

However, although the benefit of multiple imputation over Type-III sum of squares is relatively small, the results of this study are still important for other reasons. Previously it was only assumed that it was better to use D_1 than D_0 because D_0 used a noisy estimate of the total covariance matrix in its computation, which would especially be problematic for small M, but never demonstrated. The current study showed that for $M = 5$, D_0 indeed performed poorly regarding Type-I error rates (overestimation of the Type-I error rate).

Furthermore, Li et al. (1991b) showed that D_1 was robust to violation of the assumption of equal fractions of missing information across parameter estimates when the unequal fractions randomly varied across replications. The current study showed that this result also seems to hold when the unequal fractions were fixed across replications. Thus, based on the current results there does not seem to be a need for a version of D_1 that allows for unequal fractions of missing information, but which performs poorly for small M (D_0).

As for statistics D_2 and D_3, although Li et al. (1991a) showed that D_2 could sometimes be too liberal, both Grund et al. (2016) and the current study show satisfactory Type-I error rates of this method. Additionally, the current study also shows that the power of D_2 is comparable to that of Type-III sum of squares when $M = 100$. Considering these results and the fact that D_2 is relatively easy to compute, it may be a good candidate for pooling F-tests from ANOVA in multiply imputed data. However, as long as it is not clear when D_2 may overestimate the Type-I error rate, we must be cautious concluding that D_2 is generally the best alternative for pooling the results of ANOVA in multiply imputed data.

The results of D_3 were disappointing: in unbalanced data this method (severely) underestimated the Type-I error rates. Additionally, of all statistics, D_3 generally had the lowest power. An additional disadvantage of D_3 is that is not easily computed as it requires the likelihood of imputed dataset m evaluated at the average set of model parameters. The latter makes the implementation in software packages complex. Considering the disappointing results for D_3 and its complex implementation, this statistic is not the most convenient method for pooling results of ANOVA.

The results of the current study imply that software packages do not need to replace D_1 with D_0, not even in case of unequal fractions of missing information that are inherently related to the parameters in question. While it was previously shown that D_1 produces accurate Type-I error rates when fractions of missing information are on average equal across replications (Li et al, 1991b), the current study shows that D_1 also produces accurate Type-I error rates for unequal average fractions of missing information across replications.

In conclusion, it may be a bit premature to conclude that multiple imputation is a good alternative to Type-III sum of squares in unbalanced data, given the extra amount of work and the fact that its benefits only seem to show in the interaction. Finally, as most other studies have already indicated, we recommend using either D_1 or D_2 in multiply imputed data, with a slight preference of D_1.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no support to report.

Competing Interests: The authors have declared that no competing interests exist.

Supplementary Materials

For this article the following supplementary materials are available (Van Ginkel & Kroonenberg, 2021):

- Tables providing results for $J > 3$ and larger N.
- Programming code of the simulation study.

Index of Supplementary Materials

Van Ginkel, J. R., & Kroonenberg, P. M. (2021). *Supplementary materials to: Multiple imputation to balance unbalanced designs for two-way analysis of variance* [Tables, code]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.4714

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician, 52*, 119-126. https://doi.org/10.2307/2685469

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Thousand Oaks, CA, USA: SAGE.

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Pooling ANOVA results from multiply imputed datasets: A simulation study. *Methodology, 12*(3), 75-88. https://doi.org/10.1027/1614-2241/a000111

Harel, O. (2009). The estimation of R^2 and adjusted R^2 in incomplete datasets using multiple imputation. *Journal of Applied Statistics, 36*(10), 1109-1118. https://doi.org/10.1080/02664760802553000

IBM SPSS. (2017). *SPSS* (Version 25.0 for Windows) [Computer software]. Author.

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

Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991b). 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. https://doi.org/10.2307/2290525

Little, R. J. A., & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39-75). New York, NY, USA: Plenum Press. https://doi.org/10.1007/978-1-4899-1292-3_2

Marshall, A., Altman, D. G., & Holder, R. L. (2010). Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazard's model: A resampling study research article open access. *BMC Medical Research Methodology, 10*, Article 112. https://doi.org/10.1186/1471-2288-10-112

Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data with prognostic modelling studies: A simulation study. *BMC Medical Research Methodology, 10*, Article 7. https://doi.org/10.1186/1471-2288-10-7

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data* (2nd ed.). Mahwah, NJ, USA: Erlbaum.

Meng, X. L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika, 79*(1), 103-111. https://doi.org/10.1093/biomet/79.1.103

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*, 85-95.

R Core Team. (2018). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika, 94*(2), 502-508. https://doi.org/10.1093/biomet/asm028

Robitzsch, A., Grund, S., & Henke, T. (2017). *Package 'miceadds'*. Retrieved from https://cran.r-project.org/web/packages/miceadds/miceadds.pdf

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics, 4*(1), 87-94. https://doi.org/10.1080/07350015.1986.10509497

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (2nd ed.). New York, NY, USA: Wiley.

SAS. (2013). *SAS* (Version 9.4) [Computer software]. Author.

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

Shaw, R. G., & Mitchell-Olds, T. (1993). Anova for unbalanced data: An overview. *Ecology, 74*(6), 1638-1645. https://doi.org/10.2307/1939922

StataCorp. (2017). *Stata Statistical Software* (Version 15.0) [Computer software]. Author.

Van Buuren, S. (2012). *Flexible imputation of missing data.* Boca Raton, FL, USA: Chapman & Hall/CRC Press.

Van Buuren, S., & Groothuis-Oudshoorn, C. G. M. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1-67. https://doi.org/10.18637/jss.v045.i03

Van Ginkel, J. R. (2010). *MI-MUL2.SPS* [Computer code]. Retrieved from https://www.universiteitleiden.nl/en/staffmembers/joost-van-ginkel#tab-1

Van Ginkel, J. R. (2019). Significance tests and estimates for R^2 for multiple regression in multiply imputed datasets: A cautionary note on earlier findings, and alternative solutions. *Multivariate Behavioral Research, 54*(4), 514-529. https://doi.org/10.1080/00273171.2018.1540967

Van Ginkel, J. R., & Kroonenberg, P. M. (2014). Analysis of variance of multiply imputed data. *Multivariate Behavioral Research, 49*(1), 78-91. https://doi.org/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*(1), 83-117. https://doi.org/10.1111/j.1467-9531.2007.00180.x

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental designs* (3rd ed.). New York, NY, USA: McGraw-Hill.

Yuan, Y. C. (2011). Multiple Imputation using SAS Software. *Journal of Statistical Software, 45*(6), 1-25. https://doi.org/10.18637/jss.v045.i06