Combination rules for homoscedastic and heteroscedastic MANOVA models from multiply imputed datasets

Guillermo Vallejo¹ · M. Paula Fernández¹ · Pedro Rosário²

Published online: 17 August 2020
© The Psychonomic Society, Inc. 2020

Abstract

Classical MANOVA tests do not pose any difficulty when the assumptions on which they are based are satisfied, while the modified Brown-Forsythe (MBF) procedure has low sensitivity to the lack of multivariate normality and homogeneity of covariance matrices. Both methods assume complete data for all subjects. In this paper, we present combination rules for the MANOVA and MBF procedures with multiply imputed datasets. These rules are illustrated by pooling the results obtained with a two-factor multivariate design after applying the two approaches to each of the imputed datasets when the covariance matrices were equal (MI-MANOVA) and when the covariance matrices were unequal (MI-MBF). A Monte-Carlo study was carried out to compare the proposed solution, in terms of type I error rates and statistical power, with the MANOVA and MBF approaches without missing data, and with listwise deletion of missing data followed by the MANOVA approach (LD-MANOVA) and listwise deletion followed by the MBF procedure (LD-MBF). Simulations showed that the type I error rates in all analyses on datasets with missing values (with or without imputation) were well controlled. We also found that the MI-MANOVA approach was substantially more powerful than LD-MANOVA. Moreover, the power of the MI-MANOVA was generally comparable to that of its complete data counterpart. Similar results were obtained for the MI-MBF procedure when covariance matrices were unequal. We conclude, based on the current evidence, that the solution presented performs well and could be of practical use. We illustrate the application of combination rules using a real dataset.

Keywords Multivariate analysis of variance · Heteroscedasticity · Missing data · Multiple imputation

Comparisons among treatment means and treatment combinations of one or more factors on p dependent variables considered simultaneously play a central role in psychological research. As these comparisons often entail many potentially correlated dependent variables, the classical multivariate general linear model has often been used to perform tests of the omnibus or global effects, as well as multiple comparison hypotheses related to these effects. The widely used multivariate analysis of variance (MANOVA) is subsumed under this model (He et al., 2017). At present, the use of this procedure does not pose any difficulty, assuming that the observations of the dependent variables are independently drawn from a multivariate normal distribution with a common variance-covariance matrix of the dependent variables, as most

software programs automatically provide appropriate MANOVA test statistics.

Even though the test criteria most often used in the literature for testing the null hypothesis in MANOVA—Wilks' likelihood-ratio (WLR) test, Roy's largest-root test, Lawley-Hotelling's trace (LHT) test, and Bartlett-Nanda-Pillai's (BNP) trace test, among others—are based on the characteristic roots of matrices obtained from the product-moment matrices samples drawn from a normal population (i.e., sum of squares and cross-product matrices between-groups and within-groups), moderate departures from normality due to skewness do not constitute a serious problem if it is only the normality assumption that is not met (Blanca et al., 2017; Mardia, 1971). The classical MANOVA procedures, however, can become either too conservative or too liberal when the assumption of homogeneity of the population covariance matrices (i.e., homoscedasticity) is seriously violated (Konietzschke et al., 2015; Olson, 1974; Vallejo and Ato, 2012; Zhang, 2011).

Because these assumptions are often not met in practice, researchers have proposed several alternatives to the classical MANOVA models. As noted by Zhang et al. (2016), these

✉ Guillermo Vallejo
gvallejo@uniovi.es

¹ Department of Psychology, University of Oviedo, Oviedo, Spain

² School of Psychology, University of Minho, Braga, Portugal

solutions generally fall into two categories: simulation and bootstrap resampling methods, on the one hand, and approximation-based approaches on the other. For example, Gamage et al. (2004) proposed a simulation-based test using generalized P -values to obtain exact F -statistics, and Krishnamoorthy and Lu (2010) proposed a parametric bootstrap test, which was subsequently extended by Xu (2015) to the heteroscedastic one-way MANOVA and to the heteroscedastic two-way MANOVA. Later, Konietzschke et al. (2015) developed parametric and nonparametric bootstrap methods for multi-factor multivariate data. Although the above methods do not make assumptions about the equality of covariance matrices, a drawback of Monte Carlo-based approaches is that they are often time-consuming, especially when the sample sizes are large. In addition, and under some conditions of departures from normality, bootstrap methods—both parametric and nonparametric—are not always robust.

As mentioned above, another general strategy for dealing with unequal covariance matrices is to use approximation-based approaches. For example, Harrar and Bathke (2012) proposed the approximation of the null distributions of the hypothesis and error matrices involved in the modified WLR, LHT, and BNP tests using Wishart distributions with degrees of freedom estimated from the data by matching the first two moments of the aforementioned matrices to those of a p -dimensional central Wishart distribution. To overcome some drawbacks found with Harrar and Bathke's modified MANOVA tests, Zhang and Xiao (2012) proposed two other modifications of these tests using matching of covariance matrix variance components and affine-invariant covariance matrix transformations. Another procedure that has been shown to be invariant under different choices of the matrices used to define the general linear hypothesis is the modified Brown-Forsythe (MBF) approach proposed by Vallejo and Ato (2012). The authors compared this procedure with a multivariate version of the general linear model given in Vallejo et al. (2007) and two direct extensions of the works of Brunner et al. (1997) and Johansen (1980). The authors favored the MBF procedure for analyzing multivariate data collected in two-way layouts, in terms of the mean absolute relative error based on Zhang (2011).

In general, previous findings (e.g., Blanca et al., 2017; Mardia, 1971; Olson, 1974) indicate that MANOVA test criteria perform well in terms of inference when the assumption of homoscedasticity is satisfied, even under non-normality. When the assumption of homoscedasticity is violated, the MBF procedure is most robust (Vallejo and Ato, 2012). Despite these advantages, an important limitation of these approaches is that they assume complete measures with no missing values in multivariate responses, a common problem in psychological research. Examples may be found in Van Ginkel et al. (2020) and references therein. Usually the missingness for a given participant occurs only in a subset of the variables being measured. However, the standard

MANOVA and MBF procedures cannot utilize information in partially observed cases.

When there are participants with missing values in any of the dependent variables used, the researchers can use any suitable procedure designed for complete data after first excluding all participants with incomplete response vectors or, alternatively, the idea of multiple imputation (MI) proposed by Rubin (1987). The first strategy, which is known as complete case or listwise deletion (LD) analysis, relies on the assumption that data are missing completely at random (MCAR). This approach is easy to implement and works with any standard method for complete data but has the disadvantage of being less powerful and accurate as well as potentially biased, because information from partially observed cases is discarded. It should be noted that if many outcome variables are included in a model, only a small number of complete cases may remain in the after-deletion datasets for the analysis, even with low rates of missingness in the dependent variables. The second strategy, MI analysis, involves three steps: (1) replacing each missing data point with a set of values generated from its predictive distribution given the observed data; (2) analyzing each complete dataset separately using standard complete-data analysis methods, including the usual MANOVA when the assumption of homoscedasticity is tenable or the MBF procedure when this assumption is not tenable; and (3) combining the results according to appropriate rules to produce a single global estimate, which allows the uncertainty associated with the imputation of missing values to be taken into account. The MI approach usually assumes that the data are missing at random (MAR).

Several combining rules have been developed for obtaining MI inference with a variety of statistical quantities. These rules have been used, for example, for pooling the point estimates of the parameters of interest, such as a single regression coefficient or the difference between two sample means (Rubin and Schenker, 1986), multivariate tests of hypotheses (Li et al., 1991a), P -values (Li et al., 1991b), likelihood ratio tests (Meng and Rubin, 1992; Shaffer, 1997), and F -tests in the context of both ANOVA and repeated-measures ANOVA (Grund et al., 2016; Raghunathan, 2016; Raghunathan and Dong, 2011; Van Ginkel and Kroonenberg, 2014). Also, Vallejo et al. (2018) extended the Raghunathan and Dong procedure to support unequal dispersion matrices in the repeated measures context. These authors compared the MI-MBF procedure based on estimates by ordinary least squares and the linear mixed model based on maximum likelihood estimates for testing the interaction effect when the missing data have been generated under a monotone MAR mechanism. However, the widespread application of the MI technique poses new challenges for researchers. To our knowledge, no clear-cut solution has yet been proposed for this problem in the literature, either for the homoscedastic MANOVA model or for the heteroscedastic MANOVA model.

Therefore the primary aims of this study were the following: (a) to show the combination rules for pooling the results obtained after applying the same standard MANOVA analysis to each of the imputed datasets of a classic two-factor multivariate design, (b) to extend the proposed combination method in the multiple-imputation framework for the usual MANOVA to the heteroscedastic MBF procedure, (c) to evaluate the operating characteristics of the MANOVA analysis on datasets in which the missing data have been imputed (henceforth MI-MANOVA) when the assumption of homogeneity of covariance matrices is satisfied, and (d) to examine the operating characteristics of the MBF analysis based on the data completed by MI (henceforth MI-MBF) when the assumption of homogeneity of covariance matrices is violated. We investigated whether the MI-MANOVA and the MI-MBF procedures perform reasonably well in terms of controlling type I error probabilities (i.e., no treatment effects are claimed when none are present) and power (i.e., when effects are present there is no failure to detect them). As well as the MI-MANOVA and MI-MBF procedures, we also studied the type I error control and the power of MANOVA and MBF without missing data (referring to the dataset that arises when there are no missing values) and the listwise deletion of missing data (henceforth LD-MANOVA and LD-MBF).

Definition of the statistical tests

First, we summarize the homoscedastic and heteroscedastic MANOVA models in the context of factorial designs, and then we formulate the proposed combination rules for MANOVA procedures in multiply imputed datasets.

Homoscedastic two-factor MANOVA model

Consider a two-factor MANOVA model in which the levels of factors A and B are completely crossed, where A has a levels and B has b levels. Suppose at the ij th cell, we have a p -dimensional random sample $\mathbf{y}_{ijk}$ of size n_{ij} , where the subscript i refers to the level of factor A ($i = 1, \dots, a$), the subscript j refers to the level of factor B ($j = 1, \dots, b$), and the subscript k refers to the k th participant within the ij th cell ($k = 1, \dots, n_{ij}$). Assume that the vectors of observation in different cells are independent and identically distributed with mean vector $\boldsymbol{\mu}_{ij}$ and common unknown covariance matrix $\boldsymbol{\Sigma}$ (i.e., the covariance matrices are the same for all cells in the population). Then, in terms of the full rank (FR) or cell means model, the response vectors $\mathbf{y}_{ijk}$ can be modeled by

$$\mathbf{y}_{ijk} = \boldsymbol{\mu}_{ij} + \mathbf{e}_{ijk}. \quad (1)$$

The $\mathbf{e}_{ijk}$ vectors are mutually independent, each having a p -variate normal distribution with $E(\mathbf{e}_{ijk}) = \mathbf{0}$ and $V(\mathbf{e}_{ijk}) = E(\mathbf{e}_{ijk}\mathbf{e}_{ijk}') = \boldsymbol{\Sigma}$, where the prime denotes transpose.

The fact that $\boldsymbol{\Sigma}$ does not depend upon ij means that the error vectors all have the same covariance matrix. This assumption is the multivariate version of variance homogeneity.

In some cases, the cell mean vectors of the FR model are decomposed into the form $\boldsymbol{\mu}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij}$, where $\boldsymbol{\mu}$ is the grand mean vector, $\boldsymbol{\alpha}_i$ is the main effect of the i th level of A , $\boldsymbol{\beta}_j$ is the main effect of the j th level of B , and $\boldsymbol{\gamma}_{ij}$ is the interaction effect of the i th level of A with the j th level of B . Then, the less-than-full-rank (LFR) or effects model can be written as

$$\mathbf{y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij} + \mathbf{e}_{ijk}. \quad (2)$$

To uniquely define the model parameters and their least squares estimates, it is customary to impose the constraints $\sum_{i=1}^a \boldsymbol{\alpha}_i = \sum_{j=1}^b \boldsymbol{\beta}_j = \sum_{i=1}^a \boldsymbol{\gamma}_{ij} = \sum_{j=1}^b \boldsymbol{\gamma}_{ij} = \mathbf{0}$.

Both the cell means model and the effects model can be formulated in more compact form within the framework of the multivariate linear model. Let $\mathbf{Y}$ be the $N \times p$ random matrix of observations on p dependent variables generated by stacking all individual response vectors into $\mathbf{Y}$, where the N response vectors $\mathbf{y}_{111}, \dots, \mathbf{y}_{abn_{ab}}$ are assumed independent of one another. Let $\mathbf{X}$ be the $N \times q$ design matrix of independent variables related to $\mathbf{Y}$ with rank $R = R(\mathbf{X}) = ab < q$ ($1 + a + b + ab$) for the LFR model and $R(\mathbf{X}) = ab = q$ for the FR model. Then, assuming linearity of the relationship, we get

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \quad (3)$$

where $\mathbf{B}$ is a $q \times p$ matrix of unknown fixed effects common to all participants and $\mathbf{E}$ is an $N \times p$ random matrix of errors. The model specification is completed by assuming that the rows of the error term $\mathbf{E}$ are independently distributed as multivariate normal $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with an unknown positive definite covariance matrix. The assumption of equality of covariance matrices helps to provide estimators of the elements of $\mathbf{B}$ that have minimum variance. The unknown parameters in model (3) are $\mathbf{B}$ and $\boldsymbol{\Sigma}$. It is well known that the unbiased estimators of $\mathbf{B}$ and $\boldsymbol{\Sigma}$ are given by $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and $\hat{\boldsymbol{\Sigma}} = \mathbf{E}/(N-q)$, where $\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}\mathbf{X}'\mathbf{Y}$.

It is common practice in two-way MANOVA to make inferences about the main and interaction effects, i.e., simple differences in the vector of marginal means and the vector of cell means. Under the FR model, the corresponding null hypotheses can be written as

$$H_0(A) : \boldsymbol{\mu}_{1\cdot} - \boldsymbol{\mu}_{2\cdot} = \mathbf{0}, \boldsymbol{\mu}_{1\cdot} - \boldsymbol{\mu}_{3\cdot} = \mathbf{0}, \dots, \boldsymbol{\mu}_{1\cdot} - \boldsymbol{\mu}_{a\cdot} = \mathbf{0},$$

$$H_0(B) : \boldsymbol{\mu}_{\cdot 1} - \boldsymbol{\mu}_{\cdot 2} = \mathbf{0}, \boldsymbol{\mu}_{\cdot 1} - \boldsymbol{\mu}_{\cdot 3} = \mathbf{0}, \dots, \boldsymbol{\mu}_{\cdot 1} - \boldsymbol{\mu}_{\cdot b} = \mathbf{0},$$

$$H_0(AB) : \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = \mathbf{0} \text{ for all } i \neq i' \text{ and } j \neq j'.$$

In general, it is more useful to formulate the null hypotheses concerning linear combinations of parameters $\mathbf{B}$ of the form $H_0(t) : \mathbf{C}_t \mathbf{B} \mathbf{A} = \mathbf{\Gamma}_0$, $t = A, B$, or AB . Let $\mathbf{C}_t$ be the set of full-rank matrices of known constants defining between-subject contrasts, having ranks $(a-1)$, $(b-1)$, and $(a-1)(b-1)$, respectively. Also, let $\mathbf{A} = \mathbf{I}_p$ and $\mathbf{\Gamma}_0 = \mathbf{0}$, where $\mathbf{I}_p$ denotes the p -dimensional identity matrix. Then, the above null hypotheses can be equivalently expressed as follows,

$$\begin{aligned} H_0(A) : \mathbf{C}_A \mathbf{B} \mathbf{A} &= \mathbf{0}, \text{ with } \mathbf{C}_A \\ &= (\mathbf{1}_{a-1} \vdots -\mathbf{I}_{a-1}) \otimes (1/b) \mathbf{1}_b', \\ H_0(B) : \mathbf{C}_B \mathbf{B} \mathbf{A} &= \mathbf{0}, \text{ with } \mathbf{C}_B \\ &= (1/a) \mathbf{1}_a' \otimes (\mathbf{1}_{b-1} \vdots -\mathbf{I}_{b-1}), \\ H_0(AB) : \mathbf{C}_{AB} \mathbf{B} \mathbf{A} &= \mathbf{0}, \text{ with } \mathbf{C}_{AB} \\ &= (\mathbf{1}_{a-1} \vdots -\mathbf{I}_{a-1}) \otimes (\mathbf{1}_{b-1} \vdots -\mathbf{I}_{b-1}), \end{aligned}$$

where $\mathbf{I}_r$ and $\mathbf{1}_r$ denote the identity matrix of size r and the r -dimensional vector of ones, respectively, the symbol $\vdots$ represents the augmented matrix obtained by appending the columns of two given matrices, and $\otimes$ denotes the direct product operation.

A formal significance test of $H_0(t) : \mathbf{C}_t \mathbf{B} \mathbf{A} = \mathbf{0}$ versus $H_1(t) : \mathbf{C}_t \mathbf{B} \mathbf{A} \neq \mathbf{0}$ can be carried out through a function of the characteristic roots of the quantity $\mathbf{H}_t \mathbf{E}^{-1}$, where $\mathbf{E} = \mathbf{Y}' \mathbf{Y} - \hat{\mathbf{B}}' \mathbf{X}' \mathbf{Y}$ and $\mathbf{H}_t = (\mathbf{C}_t \hat{\mathbf{B}} \mathbf{A})' [\mathbf{C}_t (\mathbf{X}' \mathbf{X})^{-1} \mathbf{C}_t']^{-1} (\mathbf{C}_t \hat{\mathbf{B}} \mathbf{A})$. Later, we will show how these matrices are used to determine whether departures occur from the $H_0(t)$ being tested. For the purposes of this paper, we will restrict our discussion to the F test based on Wilks' Lambda. If homoscedasticity is assumed, Wilks' Lambda is commonly used as the MANOVA test statistic (Mertler and Vannatta, 2017).

Heteroscedastic two-factor MANOVA model

It is sometimes unrealistic to make the strong assumption that all observation vectors have the same covariance matrix. For this reason, we next consider a heteroscedastic MANOVA model; that is, we assume that the rows of $\mathbf{Y}$ are normally and independently distributed at each possible combination of the factor levels, with mean vector μ_{ij} and covariance matrix Σ_{ij} . The unbiased estimators of Σ_{ij} are $\hat{\Sigma}_{ij} = \mathbf{E}_{ij} / (n_{ij} - 1)$, where $\mathbf{E}_{ij} = \mathbf{Y}_{ij}' \mathbf{Y}_{jk} - \hat{\beta}_{ij}' \mathbf{X}_{ij}' \mathbf{Y}_{ij}$ is distributed independently as Wishart $W_p(n_{ij} - 1, \Sigma_{ij})$ and $\hat{\beta}_{ij}$ is the ordinary least squares estimator of vector β_{ij} (Nel, 1997). We also assume that $n_{ij} - 1 \leq p$, so that $\hat{\Sigma}_{ij}^{-1}$ exists with probability one. For the analysis of multivariate

heteroscedastic two-factor models, we use here the variant of the MBF approach proposed by Vallejo and Ato (2006).

Practical implementation of the MBF procedure requires estimation of the degree of freedom of the approximate central p -dimensional Wishart distribution, which can be easily derived by equating the first two moments of the quadratic form associated with the l th source of variation in model (3) to those of the central Wishart distribution. A detailed explanation of what is known as the multivariate Satterthwaite's approximation can be found in Vallejo and Ato (2006) in the context of univariate repeated measures. Applying the approach of these authors, the Wilks' lambda criterion for testing $H_0(t) : \mathbf{C}_t \mathbf{B} \mathbf{A} = \mathbf{0}$ against $H_1(t) : \mathbf{C}_t \mathbf{B} \mathbf{A} \neq \mathbf{0}$ can be expressed as the following ratio of two determinants: $\Lambda_t = |\mathbf{E}_t^*| / |(\mathbf{H}_t + \mathbf{E}_t^*)|$, where the hypothesis matrix ($\mathbf{H}_t$) and the error matrix ($\mathbf{E}_t^*$) are determined by

$$\mathbf{H}_t = (\mathbf{C}_t \hat{\mathbf{B}} \mathbf{A})' [\mathbf{C}_t (\mathbf{X}' \mathbf{X})^{-1} \mathbf{C}_t']^{-1} (\mathbf{C}_t \hat{\mathbf{B}} \mathbf{A}), \quad (5)$$

and

$$\begin{aligned} \mathbf{E}_t^* &= (f_{e(t)} / f_{h(t)}) \sum_{i=1}^a \sum_{j=1}^b c_i \hat{\Sigma}_{ij}, \quad i = 1, \dots, a; j \\ &= 1, \dots, b; t = A, B, \text{ or } AB, \end{aligned} \quad (6)$$

where $f_{e(t)}$ and $f_{h(t)}$ are the approximate degrees of freedom for $\mathbf{E}_t^*$ and $\mathbf{H}_t$, respectively, c_i is the i th diagonal element of the matrix $\mathbf{C}_t' (\mathbf{C}_t \mathbf{D} \mathbf{C}_t')^{-1} \mathbf{C}_t \mathbf{D}$, and $\mathbf{D}$ is a diagonal matrix with $1/n_{ij}$ as its ij th diagonal element. Using results from Vallejo and Ato (2006), the approximate number of degrees of freedom associated with the error and hypothesis, respectively, are

$$f_{e(t)} = \frac{p + p^2}{\sum_{i=1}^a \sum_{j=1}^b \frac{1}{n_{ij}-1} \left[tr^2 \left(c_i \hat{\Sigma}_{ij} \hat{\Sigma}^{-1} \right) + tr \left(c_i \hat{\Sigma}_{ij} \hat{\Sigma}^{-1} \right)^2 \right]} \quad (7)$$

and

$$f_{h(t)} = \frac{p + p^2}{\sum_{i=1}^a \sum_{j=1}^b G_i + tr^2 \left(\sum_{i=1}^a \sum_{j=1}^b c_i \hat{\Sigma}_{ij} \hat{\Sigma}^{-1} \right) + tr \left(\sum_{i=1}^a \sum_{j=1}^b c_i \hat{\Sigma}_{ij} \hat{\Sigma}^{-1} \right)^2}, \quad (8)$$

where $G = \left[tr^2 \left(\hat{\Sigma}_{ij} \hat{\Sigma}^{-1} \right) + tr \left(\hat{\Sigma}_{ij} \hat{\Sigma}^{-1} \right)^2 \right] (c_i - c_t)$, $\hat{\Sigma} = (c_1 \hat{\Sigma}_1 + \dots + c_{ab} \hat{\Sigma}_{ab})$, and $c_t = n_{ij}/N$.

Using Rao's F -approximation to Wilks' Lambda, $H_0(t) : \mathbf{C}_t \mathbf{B} \mathbf{A} = \mathbf{0}$ is rejected if

$$F_{MBF(t)} = \frac{1 - \Lambda_t^{1/s_t}}{\Lambda_t^{1/s_t}} \left(\frac{df_{n(t)}}{df_{d(t)}} \right) \geq F^{1-\alpha}(df_{n(t)}, df_{d(t)}), \quad (9)$$

where $df_{n(t)} = hf_{h(t)}$, $df_{d(t)} = [f_{e(t)} - (h - f_{h(t)} + 1)/2] s_t - (hf_{h(t)} - 2)/2$,

and $s_t = \left[\left(h^2 f_{h(t)}^2 - 4 \right) / \left(h^2 + \hat{f}_{h(t)}^2 - 5 \right) \right]^{\frac{1}{2}}$, with h equal to the dimension of $\mathbf{E}_t^*$. Here, $F^{1-\alpha}(df_{n(t)}, df_{d(t)})$ is the upper $1 - \alpha$ critical value of an F distribution.

Combining homoscedastic and heteroscedastic MANOVA results

Combination rules for homoscedastic MANOVA in multiply imputed datasets

Once the data have been multiply imputed and the standard MANOVA model has been used to analyze the completed datasets repeatedly, the problem lies in knowing how to combine the multiple values of Wilks' lambda test to yield a single inference. Using the theory developed by Raghunathan (2016) and Raghunathan and Dong (2011) to combine random variables with Snedecor's F distribution in the context of ANOVA models and the well-known relationship between Wilks' lambda distribution and the F distribution, it is possible to derive the formulas that facilitate the combining of the estimates obtained in the analysis of each complete dataset using the standard MANOVA model. We detail this approach below. We use M to refer the number of imputations and m to refer to each individual imputation.

As mentioned previously, after generating multiple imputed values for each missing value, a MANOVA analysis can be run within each completed dataset. Then inferences can be combined across datasets into a single pooled result. In our concrete case, this can be done using the estimates

$$\hat{Q}_{n(t)}^{(m)} \left[= \left(1 - \hat{\Lambda}_t^{1/s_t} \right)^{(m)} df_{n(t)}^{(m)} \right] \text{ and } \hat{Q}_{d(t)}^{(m)} \left[= \left(\hat{\Lambda}_t^{1/s_t} \right)^{(m)} df_{d(t)}^{(m)} \right]$$

of the quantities $Q_{n(t)}^{(m)}$ and $Q_{d(t)}^{(m)}$ from the m th dataset, which are associated with the numerator and denominator of the transformation of Wilks' lambda criterion to an $F_{(t)}$ statistic. With regard to the numerator of the F -ratio, the aggregation process necessary to obtain MI inference is determined by computing the final MI point estimate of the parameter being studied, $\bar{Q}_{n(t)}$, the within-imputation variance, $\bar{W}_{n(t)}$, and the between-imputation variance, $B_{n(t)}$. Using the standard theory of MI proposed by Rubin (1987), these quantities are given by the following expressions:

$$\bar{Q}_{n(t)} = \frac{1}{M} \sum_{m=1}^M \left(\hat{Q}_{n(t)}^{(m)} \right)^{-1} \quad (10)$$

$$\bar{W}_{n(t)} = \frac{1}{M} \sum_{m=1}^M \left(df_{n(t)}^{(m)} \hat{Q}_{n(t)}^{2(m)} \right)^{-1} \quad (11)$$

$$B_{n(t)} = \frac{1}{M-1} \sum_{m=1}^M \left(\hat{Q}_{n(t)}^{(m)} - \bar{Q}_{n(t)} \right)^2 \quad (12)$$

where $\bar{Q}_{n(t)}$ is simply the average of $Q_{n(t)}^{(m)}$ over the M imputed datasets, $\bar{W}_{n(t)}$ is the average of the squared $Q_{n(t)}^{(m)}$ over the analyses from the M imputed datasets, and $B_{n(t)}$ is the sample variance of $Q_{n(t)}^{(m)}$ over the M imputed datasets or the uncertainty that is due to missing data. The quantities referring to the denominator of the F -ratio are defined in a similar way to those used to describe the numerator, namely:

$$\bar{Q}_{d(t)} = \frac{1}{M} \sum_{m=1}^M \left(\hat{Q}_{d(t)}^{(m)} \right)^{-1} \quad (13)$$

$$\bar{W}_{d(t)} = \frac{1}{M} \sum_{m=1}^M \left(df_{d(t)}^{(m)} \hat{Q}_{d(t)}^{2(m)} \right)^{-1} \quad (14)$$

$$B_{d(t)} = \frac{1}{M-1} \sum_{m=1}^M \left(\hat{Q}_{d(t)}^{(m)} - \bar{Q}_{d(t)} \right)^2. \quad (15)$$

To test the null multivariate hypothesis $H_0(t) : \mathbf{C}_t \mathbf{B} \mathbf{A} = \mathbf{0}$, the F -ratio test is proposed, where F is defined by

$$F_{MI(t)} = \frac{\bar{Q}_{d(t)}}{\bar{Q}_{n(t)}}. \quad (16)$$

As the reference distribution for the proposed MI F -statistic, it is recommended to use an F distribution with the degrees of freedom $df_{cn(t)} = 2 \bar{Q}_{n(t)}^2 / T_{n(t)}$ and $df_{cd(t)} = 2 \bar{Q}_{d(t)}^2 / T_{d(t)}$, where $T_{n(t)} = 2 \bar{W}_{n(t)} + B_{n(t)} + B_{n(t)}/M$ and $T_{d(t)} = 2 \bar{W}_{d(t)} + B_{d(t)} + B_{d(t)}/M$ estimate the total MI variance of $Q_{n(t)}^{-1|}$ and $Q_{d(t)}^{-1|}$ given the observed data Y_{obs} , as discussed by Rubin (1987). The total variance is equal to twice the within-imputation variance plus the between-imputation variance weighted by a correction for a finite number of imputations $(1+1/M)$. When $M = 1$, the ratio $F_{MI(t)}$ and numerator and denominator degrees of freedom associated with the terms of the ratio (i. e., $df_{cn(t)}$, $df_{cd(t)}$) coincide with the quantities obtained using the standard MANOVA model when there are no missing data.

To derive the numerator degrees of freedom, it is assumed that $Q_{n(t)}^{-1|} | Y_{obs}$ is approximately distributed as $c_{n(t)} \chi_{df_{cn(t)}}^2$,

where $c_{n(t)}$ is a constant and $\chi^2_{df_{cn(t)}}$ is a central chi-square with $df_{cn(t)}$ degrees of freedom, and $c_{n(t)}$ and $df_{cn(t)}$ are determined by matching the first two moments of $Q_{n(t)}^{-1}|Y_{obs}$ to those of $c_{n(t)}\chi^2_{df_{cn(t)}}$, respectively. The numerator $df_{cn(t)}$ is obtained by simultaneously solving the following equations:

$$\begin{aligned}\bar{Q}_{n(t)} &= E\left(c_{n(t)}\chi^2_{df_{cn(t)}}\right) = c_{n(t)}df_{cn(t)} \\ T_{n(t)} &= V\left(c_{n(t)}\chi^2_{df_{cn(t)}}\right) = 2c_{n(t)}^2df_{cn(t)}\end{aligned}\quad (17)$$

i.e.,

$$df_{cn(t)} = \frac{2\bar{Q}_{n(t)}^2}{2\bar{W}_{n(t)} + B_{n(t)} + B_{n(t)}/M}\quad (18)$$

The approximate $df_{cd(t)}$ for the denominator is derived in the same way as the numerator. Computational details of the implementation of Eqs. 10–18 will be described in the section entitled “Empirical illustration using real data”.

Combination rules for heteroscedastic MANOVA in multiply imputed datasets

When there are missing data and the assumption of equality of covariance matrices is not tenable, a simple alternative for dealing with missing data is to use multiple imputation followed not by the homoscedastic MANOVA model, but by a heteroscedastic MANOVA model (e.g., the MBF procedure). The proposed combination rules for the MBF multiply imputed datasets are implemented in a manner analogous to the pooling procedure described above. That is, we computed the appropriate MBF analysis with each imputed complete dataset to obtain the M estimates, $\hat{Q}_{n(t)}^{(m)} = \left(1 - \hat{\Lambda}_t^{1/s_t}\right)^{(m)} df_{d(t)}^{(m)}$

and $\hat{Q}_{d(t)}^{(m)} = \left(\hat{\Lambda}_t^{1/s_t}\right)^{(m)} df_{n(t)}^{(m)}$, associated with the numerator

and denominator of the transformation of Wilks' lambda criterion to the $F_{MBF(t)}$ statistic given in Eq. 9. Then the aggregation process necessary to obtain MI inference was determined by computing the final MI estimates of the parameters being studied,

$\bar{Q}_{n(t)}$ and $\bar{Q}_{d(t)}$, the within-imputation variances, $\bar{W}_{n(t)}$ and $\bar{W}_{d(t)}$, and the between-imputation variances, $B_{n(t)}$ and $B_{d(t)}$, using the specific formulas of Eqs. 10–15. Next, from the final estimates we calculated the $F_{MI-MBF(t)}$ -statistic and from the within- and between-imputation variances the numerator and denominator degrees-of-freedom approximation, using the Eqs. 16 and 18.

Method

Design

In order to numerically assess the power and robustness of the standard MANOVA model when missing values in the outcomes are imputed (i.e., MI-MANOVA) and covariance matrices are equal, we carried out a Monte Carlo study using a two-factor multivariate design with three and four levels, respectively. The same design has been utilized to assess the operating characteristics of the MBF procedure when multiple imputation is used to deal with missing data (i.e., MI-MBF) and covariance matrices are unequal. Many independent variables studied by educational and psychological researchers have this number of conditions or levels (see Baker et al., 2017).

Study variables

For comparative purposes, in assessing the validity of inferences made from the LD and MI approaches, two scenarios are considered. One is based on homoscedastic MANOVA, and the other on heteroscedastic MANOVA. In both scenarios, our simulation study assumed that missingness was MCAR, as both LD and MI work under this mechanism, so LD can serve as a lower bound of the power in datasets with the specific percentage of missing data. In the context of multiple regression analysis of multiply imputed data, Van Ginkel (2019) used the same reasoning. In this paper, all multivariate missing values were imputed using a fully conditional specification (FCS; also known as sequential regression and multivariate imputation by chained equations) algorithm as implemented in the SAS Proc MI (SAS Institute, 2018). With FCS, the missing values are imputed by sampling iteratively from univariate conditional distributions, where each incomplete variable is an outcome at one stage of the algorithm and a predictor variable at all others (Van Buuren, 2007).

For the first scenario, three factors were manipulated in order to examine the robustness, in terms of control of type I error rates, of the proposed combination rules for MANOVA imputed datasets when the assumption of equality of covariance matrices was met. These factors were as follows: (a) number of dependent variables, (b) total sample size, and (c) missing data rate.

The value of p , the number of dependent variables, was set at either two or four. We initially planned to investigate $p = 6$ as well, however, preliminary simulations indicated that the validity of the complete-case analysis is questionable when the sample size is small, due to the fact that there are sixty-four possible missing patterns in six-outcome data. It is doubtful that this change substantially affected the results for the procedure of interest (i.e., MI-MANOVA).

Total sample sizes were set at 108 and 216. Only these two sizes were examined because we felt they would suffice to provide a comparison between the procedures for small and moderate sample sizes. The two unequal cell sizes were as follows: $n_{1j} = (6, 6, 7, 8)$, $n_{2j} = (6, 8, 10, 11)$, and $n_{3j} = (7, 10, 12, 17)$ for $N = 108$ and $n_{1j} = (12, 12, 14, 16)$, $n_{2j} = (12, 16, 20, 22)$, and $n_{3j} = (14, 20, 24, 34)$ for $N = 216$. Balanced designs were not included in the study. Within each of these two total sample size conditions, a moderate degree of cell size inequality was examined, as indexed by a coefficient of variation, Δn_{ij} , where $\Delta n_{ij} = (N/ab) \left[\sum_{ij} (n_{ij} - n)^2 / ab \right]^{1/2}$ and $n = \sum_{ij} n_{ij} / ab$. Thus, the coefficient of variation of the cell sizes was set at 0.345 both for $N = 108$ and $N = 216$.

We simulated MCAR data, with low and medium rates of missingness, by randomly deleting scores from the dependent variables. The deleting process typically produced a missing data rate of 10 to 20% for each variable. Enders (2003) suggests that a missing rate of 15 to 20% is common in educational and psychological studies.

For the second scenario, four factors were manipulated in order to examine the robustness, in terms of control of type I error rates, of the proposed combination rules for MBF imputed datasets when the assumption of equality of covariance matrices did not hold. These factors were as follows: (a) number of dependent variables, (b) total sample size, (c) missing data rate, and (d) relationship of cell sizes to dispersion matrices. The first three factors were applied as defined in Scenario 1. Regarding the relationship of cell sizes to dispersion matrices, both the positive and negative relationships between cell sizes and dispersion matrices were investigated. In the positive relationship, the larger cell sample sizes were associated with the larger dispersion matrices. In the negative relationship, the smaller cell sample sizes were associated with the larger

dispersion matrices. Table 1 presents the values of the population covariance matrices for $p = 2$ and for $p = 4$. The heterogeneous covariance matrices we used here were obtained as follows: $\Sigma_{ij} = (2 \times i \times j) / 2$. In particular, we studied a 3×4 multivariate design in which unequal cell covariance matrices were in the ratio of 1:2:3:4:2:4:6:8:3:6:9:12.

The constants in both scenarios were: (a) the shape of the distribution of the measured variable, (b) the degree of correlation between the dependent variables (i.e., $\rho = 0.60$), (c) the degree of sample size imbalance (i.e., $\Delta n_{ij} = 0.345$), (d) the missing data mechanism (i.e., MCAR), (e) the method used to create the M complete datasets (i.e., FCS), and (f) the number of imputations. Additionally, the degree of heterogeneity between the covariance matrices was kept constant in Scenario 2. Here, we will focus on examining the performance of developed combination rules for MANOVA and MBF imputed datasets when the data were generated from a multivariate normal distribution, after performing 10 imputations using a sequential regression imputation method as implemented through the SAS PROC MI module. Although the optimal number of imputations depends on the percentage of missing information, current practice favors carrying out between 5 and 20 imputations in the case of having a maximum of 20% of missing values, and up to 50 imputations in the case of having a higher percentage of unobserved data (Kenward and Carpenter, 2007).

In both scenarios, we also focused on the comparison of the powers of these approaches (i.e., MI-MANOVA and MI-MBF) for detecting main and interaction effects in two-way multivariate designs, but only for the methods whose validity does not depend on the underlying assumptions made. In this power study, the effects of the non-null hypothesis were investigated using the maximum range mean configuration. In order to detect the sensitivity of the row main effect, we

Table 1 Parameter values of the covariance matrices used to generate the data (top panel), configurations of the pattern means examined (central panel), and noncentrality parameters for power calculations (bottom panel)

Unstructured ($\rho = 0.6; p = 2$)				Unstructured ($\rho = 0.6; p = 4$)								
$\Sigma = Cov\left(e_{ijkl}, e_{ijkl'}\right) = \sigma_l^2 \text{ if } l = l \text{ and } \sigma_{ll}, \text{ if } l \neq l'$				$\Sigma = Cov\left(e_{ijkl}, e_{ijkl'}\right) = \sigma_l^2 \text{ if } l = l \text{ and } \sigma_{ll}, \text{ if } l \neq l'$								
$\sigma_1^2 = 1; \sigma_2^2 = 2; \sigma_{12} = 0.8485$				$\sigma_1^2 = 1; \sigma_2^2 = 2; \sigma_3^2 = 3; \sigma_4^2 = 4$ $\sigma_{12} = 0.8485; \sigma_{13} = 1.0392; \sigma_{14} = 1.2000$ $\sigma_{23} = 1.4697; \sigma_{24} = 1.6971; \sigma_{34} = 2.0785.$								
<div><div>Rows</div><div>$\begin{bmatrix} \mu & \mu & \mu & \mu \\ 0 & 0 & 0 & 0 \\ -\mu & -\mu & -\mu & -\mu \end{bmatrix}$</div></div>				<div><div>Columns</div><div>$\begin{bmatrix} \mu & 0 & 0 & -\mu \\ \mu & 0 & 0 & -\mu \\ \mu & 0 & 0 & -\mu \end{bmatrix}$</div></div>				<div><div>Interaction</div><div>$\begin{bmatrix} \mu & 0 & 0 & -\mu \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$</div></div>				
MANOVA ($p = 2$)				MANOVA ($p = 4$)			MBF ($p = 2$)			MBF ($p = 4$)		
N	λ_A	λ_B	λ_{AB}	λ_A	λ_B	λ_{AB}	λ_A	λ_B	λ_{AB}	λ_A	λ_B	λ_{AB}
108	6.41	7.62	10.42	8.56	10.34	13.52	6.62	7.82	10.60	8.66	10.36	13.91
216	11.56	13.58	18.43	15.57	18.63	24.50	11.61	13.94	18.64	15.66	18.92	24.83

Note. p = number of dependent variables; ρ = degree of correlation between dependent variables; l = the value of the l subscript ranges from 1 to p ; N = total sample size; λ = noncentrality parameter

included the configuration $(\mu, 0, -\mu)$ within each of the four levels of the main effect of B. To detect the sensitivity of the column main effect, we included the configuration $(\mu, 0, 0, -\mu)$ within each of the three levels of the main effect of A. There is no reason to think that much would be learned by manipulating other configurations of the pattern of means. When the interest was focused on detecting the power of the interaction effect between A and B, only the first level of A presented a means vector identical to that described; for the other two levels, the means vector was the null vector. For each configuration of the non-null effects involved in the design, we selected μ values that provided nominal power $(1 - \beta)$ of 0.50 when $N = 108$ and $\alpha = 0.05$. The magnitude of the power difference would be changed by using some other value, but the pattern of results would not likely be altered. This pattern of means was investigated for the MANOVA and MBF procedures based on the original or complete data when the correlation among the dependent variables was 0.60. For the MBF procedure, the nominal power for detecting the main and interaction effects was limited to the positive pairing condition. Details of the pattern of means, covariance matrices, and noncentrality parameters for power calculations are shown in Table 1. Noncentrality parameters were calculated using a method similar to the one suggested by Shieh (2003) for the multivariate general linear model.

Data generation and evaluation criteria

In order to investigate the behavior of the proposed approaches, multivariate normal data were generated using the following steps:

1. For each cell of the design, generate $\mathbf{E}_{ij}$, an $n_{ij} \times p$ matrix of pseudorandom normal variates with mean $E(\mathbf{E}_{ij}) = \mathbf{0}$ and $Cov(\mathbf{E}_{ij}) = \mathbf{I}_p$.
2. Create multivariate complete datasets $\mathbf{Y}_{ij} = \mathbf{B}_{ij} + \sum_{ij}^{1/2} \mathbf{E}_{ij}'$, where $\sum_{ij}^{1/2}$ is a square root matrix of dimension p satisfying the equality $\sum_{ij}^{1/2} \sum_{ij}^{1/2} = \sum_{ij}$ and $\mathbf{B}_{ij}$ is an $n_{ij} \times p$ matrix containing the population fixed effects.
3. Multiply $\mathbf{Y}_{ij}$ by d_{ij} , a constant selected to simulate the required degree of heterogeneity.

Five thousand replications of each of the 144 combinations of conditions (48 under Scenario 1 and 96 under Scenario 2) were performed using a .05 nominal alpha level. The proportion of false rejections of the null and alternative hypotheses was computed to evaluate the operating characteristics of the methods investigated. A SAS/IML macro was used for all calculations.

Results

Type I error rates

Tables 2 and 3 show empirical rates of type I error for tests of main and interaction effects when the effect size is null. Bradley's (1978) liberal criterion was adopted to evaluate the conditions under which the homoscedastic and heteroscedastic procedures were considered robust with respect to type I error control. According to this criterion, tests whose empirical error rate ($\hat{\alpha}$) lies in the interval $.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$, are considered robust. Therefore, for the nominal α level

Table 2 Percentages of type I error for the omnibus effects of a multivariate 3×4 factorial design using the MANOVA model for the original data, LD analysis, and the multiply imputed data

MCAR 10% ($p = 2$)				MCAR 10% ($p = 4$)			MCAR 20% ($p = 2$)			MCAR 20% ($p = 4$)		
N	A	B	AB	A	B	AB	A	B	AB	A	B	AB
MANOVA without missing data												
108	5.22	5.04	4.73	4.82	4.56	5.34	5.22	5.04	4.73	4.82	4.56	5.34
216	5.20	4.98	4.68	5.10	4.74	4.84	5.20	4.98	4.68	5.10	4.74	4.84
LD-MANOVA												
108	5.14	5.66	4.78	4.98	4.96	5.31	5.23	5.13	5.84	5.00	5.05	4.80
216	5.18	4.98	4.92	5.10	4.90	5.14	5.50	5.10	4.70	4.93	5.29	5.17
IM-MANOVA												
108	4.04	3.86	3.22	5.05	4.03	3.63	4.51	3.99	2.73	5.76	4.16	2.54
216	4.40	4.12	3.04	5.78	4.06	3.32	4.94	3.72	2.58	5.61	4.30	3.02

See the note to Table 1. MCAR = data are missing completely at random; A = main effect of factor A; B = main effect of factor B; AB = interaction effect between A and B; LD-MANOVA = MANOVA model based on listwise deletion; MI-MANOVA = MANOVA model based on the multiply imputed datasets

Table 3 Percentages of type I error for the omnibus effects of a multivariate 3×4 factorial design using the MBF approach for the original data, LD analysis, and multiply imputed data

MCAR 10% ($p = 2$)					MCAR 10% ($p = 4$)			MCAR 20% ($p = 2$)			MCAR 20% ($p = 4$)		
N	Pair	A	B	AB	A	B	AB	A	B	AB	A	B	AB
MBF without missing data													
108	+	5.30	4.94	5.62	5.44	4.94	5.04	5.30	4.94	5.62	5.44	4.94	5.04
216	+	6.06	5.72	6.54	5.32	4.78	6.10	6.06	5.72	6.54	5.32	4.78	6.10
108	–	5.10	4.68	5.46	4.28	4.73	4.88	5.10	4.68	5.46	4.28	4.73	4.88
216	–	6.00	5.46	6.18	5.54	5.50	5.78	6.00	5.46	6.18	5.54	5.50	5.78
LD-MBF													
108	+	5.04	4.78	5.18	4.46	4.32	3.84	5.24	4.70	5.38	4.06	3.32	3.60
216	+	6.12	5.84	6.14	5.28	5.36	6.48	6.10	5.84	5.94	4.92	5.54	5.92
108	–	5.22	4.46	4.76	3.30	3.58	3.90	5.09	3.92	4.92	3.08	3.06	2.86
216	–	5.74	5.66	6.34	5.86	4.78	4.86	5.12	5.28	6.12	4.94	4.78	5.24
MI-MBF													
108	+	4.98	4.50	4.58	4.40	3.22	4.06	4.58	3.99	4.14	6.34	3.36	4.38
216	+	5.00	4.18	4.12	4.26	3.40	4.50	6.26	4.72	4.18	5.16	3.06	3.32
108	–	5.94	4.24	5.26	4.32	3.84	5.14	6.70	4.02	4.64	6.32	4.64	5.34
216	–	5.32	4.91	5.64	5.48	4.12	5.60	6.42	4.58	4.94	7.08	4.66	6.02

See the notes to Tables 1 and 2. + Pair = positive pairing of cell sizes and covariance matrices; – Pair = negative pairing of cell sizes and covariance matrices; LD-MBF = MBF approach based on listwise deletion; MI-MBF = MBF approach based on the multiply imputed datasets

employed in this study ($\alpha = 5\%$), the interval used for defining the robustness of the tests was $2.5 \leq \hat{\alpha} \leq 7.5$. Correspondingly, a test procedure was considered too liberal if its estimated type I error rate was greater than 7.5 and too conservative if the error rate was less than 2.5. With regard to the method for identifying a nonrobust procedure, it should be noted that several standards have been used by researchers (see Bradley, 1978); therefore, different interpretations of the results are possible. Nevertheless, we found it beneficial to use the abovementioned criterion in order to facilitate comparison between our results and those obtained by other researchers.

Table 2 contains the empirical rates (percentages) of type I error for the global effects of a 3×4 multivariate design when both normality and homoscedasticity were present, and the missing data were generated under a non-monotone MCAR process. From the table, it can be seen that the performance of both listwise deletion of missing data followed by use of the standard MANOVA model (i.e., LD-MANOVA) and multiple imputation of missing data followed by the MANOVA model (i.e., MI-MANOVA) are close to the nominal level of 5%. In general, the differences between LD-MANOVA and MI-MANOVA are small and scarcely affected by the independent variables manipulated in our investigation. For LD-MANOVA, the percentage of type I error varies from 4.70 to 5.84, while for MI-MANOVA this percentage varies from 2.54 to 5.76. Needless to say, the rates of error for the standard MANOVA model without missing data (i.e., from the complete data) were also well controlled (i.e., type I error rates ranged from 4.56 to 5.34).

The empirical type I error rates of the MBF procedure for testing the main and interaction effects when the covariance matrices are unequal but normality holds are contained in Table 3. Again, our descriptive analysis of the results indicated that rates of type I error were well controlled when the missing data were generated under a non-monotone MCAR process. That is, all type I error rates of listwise deletion followed by the use of the MBF procedure (i.e., LD-MBF) and multiple imputation followed by the MBF procedure (i.e., MI-MBF) were always within the bounds of Bradley's (1978) liberal criterion regardless of the number of dependent variables, cell sizes, missing data rate, or type of pairing of group sizes and covariance matrices investigated. For the heteroscedastic technique, the error rates ranged between 2.86 and 6.48 for the LD-MBF procedure, and between 3.06 and 7.08 for the MI-MBF procedure. Similar to the results for the homoscedastic MANOVA, the error rates for the MBF approach without missing data were also well controlled (i.e., the type I error rates ranged from 4.28 to 6.54).

It is important to note that although the type I error rates were always within Bradley's (1978) bounds, the LD-MANOVA and MI-MANOVA approaches tended to produce error rates that were closer to theoretical expectation than those of the LD-MBF and MI-MBF procedures.

Power results

Making power comparisons between procedures is of interest only when they are able to control the type I error rate close to

a nominal level (i.e., 5%). As has been shown previously, the results presented in Tables 2 and 3 under null hypothesis indicate that all MANOVA and MBF rates of error were well controlled when the missing data were generated under a non-monotone MCAR process. Thus, the power values of these homoscedastic and heteroscedastic procedures, when the alternative hypothesis is true, can be directly compared in Tables 4 and 5.

The results for the omnibus effects under normality and equal covariance matrices are shown in Table 4. From the table, it can be seen that for the main effect of A, the standard MANOVA model without missing data performed slightly better than the MI-MANOVA, and much better than the LD-MANOVA approach. Averaged across all the manipulated conditions, the overall success was 66.04% for the standard MANOVA model without missing data, 43.46% for the LD-MANOVA, and 64.71% for the MI-MANOVA. A similar pattern was observed by examining the remaining overall design effects. More specifically, for the main effect of B, the overall success was 65.74% for the standard MANOVA model without missing data, 41.45% for the LD-MANOVA, and 63.33% for the MI-MANOVA. For the interaction effect, the overall success was 67.35% for the standard MANOVA model without missing data, 41.20% for the LD-MANOVA, and 60.90% for the MI-MANOVA. Consequently, averaged over the main and interaction effects, the overall difference between the standard MANOVA model without missing data and MI-MANOVA was 3.4 percentage points and the difference between MI-MANOVA and LD-MANOVA was 20.9 percentage points, regardless of the number of dependent variables, cell sizes, or rates of missingness. It should be noted that none of the power advantage could be due to differences

in type I error control, since all LD-MANOVA and MI-MANOVA rates of error were well controlled.

The empirical percentages of power of the MBF procedure for testing the global effects when the covariance matrices are unequal but normality holds are contained in Table 5. From the table, it can be seen that for the main effect of A, the MBF procedure without missing data performed slightly better than the MI-MBF, and much better than the LD-MBF approach. Averaged across all the manipulated conditions, the overall success was 59.65% for the MBF procedure without missing data, 35.37% for the LD-MBF, and 57.77% for the MI-MBF. A similar pattern was observed by examining the remaining omnibus effects of design. More specifically, with regard to the main effect of B, the overall success was 59.48% for the MBF procedure without missing data, 34.92% for the LD-MBF, and 56.02% for the MI-MBF. With regard to the interaction effect, the overall success was 66.14% for the MBF procedure without missing data, 36.63% for the LD-MBF, and 61.57% for the MI-MBF. Hence, it is clear that all analyses that have missing data without imputation (i.e., LD-MBF) led to very different results than the analysis performed on each complete dataset, as well as that based on data completed by imputation (MI-MBF). To sum up, the average difference between the MBF procedure without missing data and the MI-MBF procedure was 1.9 percentage points for the main effect of A, 3.5 percentage points for the main effect of B, and 4.6 percentage points for the interaction effect. At the same time, the average difference between the MI-MBF and LD-MBF procedures was 22.4 percentage points for the main effect of A, 21.1 percentage points for the main effect of B, and 24.9 percentage points for the interaction effect.

Table 4 Percentages of power for the omnibus effects of a multivariate 3×4 factorial design using the MANOVA model for the original data, LD analysis, and the multiply imputed data

MCAR 10% ($p = 2$)				MCAR 10% ($p = 4$)			MCAR 20% ($p = 2$)			MCAR 20% ($p = 4$)		
N	A	B	AB	A	B	AB	A	B	AB	A	B	AB
MANOVA without missing data												
108	51.25	49.17	50.81	51.28	51.20	51.24	51.25	49.17	50.81	51.28	51.20	51.24
216	79.24	79.32	82.12	82.38	83.26	84.82	79.24	79.32	82.12	82.38	83.26	84.82
LD-MANOVA												
108	41.26	39.17	39.23	30.49	31.05	29.74	31.11	29.18	30.37	17.12	17.91	16.03
216	69.01	67.89	69.85	58.35	57.97	57.53	55.25	54.85	55.19	45.07	33.61	31.62
IM-MANOVA												
108	50.01	46.77	45.35	50.80	48.92	45.17	49.56	46.47	41.96	51.29	50.71	44.05
216	77.86	76.82	78.06	80.94	80.94	80.60	77.20	75.40	73.91	80.01	80.61	78.09

See the notes to Tables 1 and 2. Bold values denote conditions where the power of the MI-MANOVA exceeded that of the MANOVA model based on the complete data analysis

Table 5 Percentages of power for the omnibus effects of a multivariate 3×4 factorial design using the MBF approach for the original data, LD analysis, and the multiply imputed data

MCAR 10% ($p = 2$)					MCAR 10% ($p = 4$)			MCAR 20% ($p = 2$)			MCAR 20% ($p = 4$)		
N	Pair	A	B	AB	A	B	AB	A	B	AB	A	B	AB
MBF without missing data													
108	+	49.34	48.37	48.31	50.97	50.61	48.91	49.34	48.37	48.31	50.97	50.61	48.91
216	+	78.44	79.72	85.86	83.00	83.64	87.38	78.44	79.72	85.86	83.00	83.64	87.38
108	–	41.45	37.15	43.75	41.37	40.11	43.97	41.45	37.15	43.75	41.37	40.11	43.97
216	–	65.79	65.75	84.46	69.89	70.47	86.44	65.79	65.75	84.46	69.89	70.47	86.44
LD-MBF													
108	+	38.06	37.27	34.92	26.69	24.97	23.73	27.48	26.27	24.76	13.86	16.68	15.14
216	+	67.57	68.55	57.35	58.30	57.26	57.94	54.40	54.92	54.64	32.65	42.24	41.42
108	–	26.79	26.31	30.65	17.51	16.31	20.17	19.05	18.24	20.16	8.87	10.30	12.55
216	–	52.47	52.23	55.15	41.36	40.02	52.74	38.88	39.22	49.62	21.98	27.84	35.12
MI-MBF													
108	+	46.07	43.85	42.15	47.93	45.35	42.23	43.97	42.33	37.25	49.03	43.23	40.23
216	+	75.68	76.20	85.86	80.42	79.78	82.08	73.79	71.83	74.41	79.72	77.74	78.37
108	–	38.67	35.85	42.89	36.17	34.75	42.85	36.69	36.79	43.05	40.97	39.77	43.45
216	–	65.65	65.41	80.38	70.03	67.77	84.46	66.71	65.23	81.16	72.85	70.47	84.34

See the notes to Tables 1, 2, and 3. Bold values denote conditions where the power of the MI-MBF exceeded that of the MBF approach based on the complete data analysis

Empirical illustration using real data

The purpose of this section is to provide a comparison of the MBF without missing data, the LD-MBF, and the MI-MBF tests based upon a subset of the data described in Cueli et al. (2019). Missing data were generated under a non-monotone MCAR process, deleting 10% of all values in the dataset. The chief aim of this study was to determine the influence of initial levels of mathematics competency in relation to performance gains after an intervention with integrated dynamic representation (IDR; González-Castro et al., 2014) in elementary school students. The participants were 288 students (aged 6–8 years) who were classified according to their levels of mathematics competency (low–medium–high). Later, the students were assigned to the two primary intervention groups, the experimental group (students who received the IDR intervention) and the control group (students who followed the traditional instructional methods). The starting hypothesis of the study was that the students with the lowest initial levels of mathematics competency would benefit more from the intervention than those with medium and/or high competency levels. The formal competencies were assessed using four specific subscales (i.e., conventionalisms, number facts, formal calculation, and formal concepts) from the Test of Early Mathematics Abilities (Ginsburg and Baroody, 2003).

Table 6 shows the results obtained after applying the MBF without missing data, LD-MBF, and MI-MBF procedures to the data of Cueli et al. (2019) to check the individual effects of

the types of treatment and levels of mathematics competency, as well as the effect of the two factors in combination, when the four formal dependent variables were considered simultaneously. In Table 6, the F -values, P -values and the associated degrees of freedom of all the tests were computed using the F -approximation statistic for the Wilks' lambda, which is extensively used by data analysts.

A comparison of the results obtained with the three methods evaluated indicates that there are statically significant differences between the types of treatment and among the levels of mathematics competency (see Table 6). Of particular interest is the fact that the interaction effect between type of treatment and mathematical competence is statistically significant for both the MBF without missing data and MI-MBF methods. Note that both P -values here (0.0078 for MBF without missing data and 0.0152 for MI-MBF based on 20 imputations) are appreciably smaller than the P -value of 0.2325 obtained with the LD-MBF method. Although there are differences in the results, the conclusions drawn from using the MI-MBF procedure have more weight than those produced using the LD-MBF method. We found that the MI-MBF procedure was more efficient, as can be seen from the greater power in Table 5 and the smaller P -values in Table 6.

Also, we can see from the results reported in Table 6 that the approach adopted here for the handling of missing values in multivariate data analysis (i.e., MI-MBF), generally speaking, provides conclusions about the main and interaction effects that are consistent with those obtained using the

Table 6 Test results for the influence of mathematical competencies on the program effectiveness

Method	Types of treatment				Levels of competency				Types × Levels			
	df_n	df_d	F	Pr >F	df_n	df_d	F	Pr >F	df_n	df_d	F	Pr >F
5 Imputations												
MBF	4.00	237.20	81.42	.0001	7.58	453.88	3.51	.0008	7.58	453.88	2.69	.0078
LD-MBF	4.00	153.61	42.58	.0001	7.68	297.32	2.86	.0051	7.68	297.32	1.33	.2325
MI-MBF	3.91	201.21	66.76	.0001	7.32	444.41	3.39	.0013	6.96	445.75	2.36	.0229
10 Imputations												
MBF	4.00	237.20	81.43	.0001	7.58	453.88	3.51	.0008	7.58	453.88	2.69	.0078
LD-MBF	4.00	153.61	42.58	.0001	7.68	297.32	2.86	.0051	7.68	297.32	1.33	.2325
MI-MBF	3.92	216.47	68.36	.0001	7.24	446.02	3.49	.0010	6.76	444.49	2.43	.0201
20 Imputations												
MBF	4.00	237.20	81.43	.0001	7.58	453.88	3.51	.0008	7.58	453.88	2.69	.0078
LD-MBF	4.00	153.61	42.58	.0001	7.68	297.32	2.86	.0051	7.68	297.32	1.33	.2325
MI-MBF	3.99	221.71	69.73	.0001	7.28	449.47	3.53	.0009	6.89	449.48	2.53	.0152

See the note to Table 4. df_n = numerator degrees of freedom (df); df_d = denominator df. One hundred and ten data points were removed from the original study according to a MCAR mechanism. MBF = MBF approach based on the complete data analysis

procedure that serves as a benchmark (i.e., the MBF procedure without missing data). Although there are differences in F -values, P -values, and degrees of freedom among these procedures, we can see that they are rather small. Furthermore, it should be noted that the degree of discrepancy between these methods decreases by increasing the number of imputations of missing values. Thus, multivariate analyses of the real datasets once again confirmed that MI-MBF performs nearly as well as the MBF procedure without missing data but is notably better than LD-MBF. Consequently, the superiority of the MBF without missing data and MI-MBF procedures compared with LD-MBF is consistent with the results obtained in our Monte Carlo simulations.

We now illustrate how Eqs. 8–16, which were presented in the subsection entitled “Combining MBF inferences from multiple imputations”, may be applied to the previous study to combine the estimates obtained in the analysis of each complete dataset with the MBF approach. Owing to the limitations of space, our interest focuses only on the null hypothesis of zero interaction effect with five multiple imputations. The values obtained after applying the MBF approach to each of the imputed datasets are $\Lambda_1 = 0.9325$, $\Lambda_2 = 0.9261$, $\Lambda_3 = 0.9367$, $\Lambda_4 = 0.9314$, and $\Lambda_5 = 0.9109$ for Wilks’ Λ , and $s_1 = 1.9211$, $s_2 = 1.9157$, $s_3 = 1.9154$, $s_4 = 1.9072$, and $s_5 = 1.9122$ for s (see Eq. 7). On the other hand, the degrees of freedom associated with Rao’s F -approximation to Λ are $df_{n1} = 7.6136$, $df_{n2} = 7.5881$, $df_{n3} = 7.5863$, $df_{n4} = 7.5473$, and $df_{n5} = 7.5771$ for the numerator of the F -ratio, and $df_{d1} = 458.7256$, $df_{d2} = 453.5790$, $df_{d3} = 449.2838$, $df_{d4} = 446.6193$, and $df_{d5} = 452.5821$ for the denominator of the F -ratio.

That said, testing for the interaction effect with imputed data ($M = 5$) involves the following steps:

1. Calculate the MI estimates of parameters for the numerator and denominator of the F -ratio using Eqs. 8 and 11. Substituting values in these algebraic expressions yields $Q_{n(AB)} = [(1/((1-0.9325^{1/s_1}) \times 458.7256)) + (1/((1-0.9261^{1/s_2}) \times 453.5790)) + (1/((1-0.9367^{1/s_3}) \times 449.2838)) + (1/((1-0.9314^{1/s_4}) \times 446.6193)) + (1/((1-0.9109^{1/s_5}) \times 452.5821))] / 5 = 0.0582385$, and $Q_{d(AB)} = [(1/(0.9325^{1/s_1} \times 7.6136)) + (1/(0.9261^{1/s_2} \times 7.5881)) + (1/(0.9367^{1/s_3} \times 7.5863)) + (1/(0.9314^{1/s_4} \times 7.5473)) + (1/(0.9109^{1/s_5} \times 7.5771))] / 5 = 0.1371957$.

2. Calculate the within-imputation variance for the numerator and denominator of the F -ratio using Eqs. 9 and 12. Substituting values in these algebraic expressions yields $W_{n(AB)} = [1/(((1-0.9325^{1/s_1}) \times 458.7256)^2 \times 7.6136) + 1/(((1-0.9261^{1/s_2}) \times 453.5790)^2 \times 7.5881) + 1/(((1-0.9367^{1/s_3}) \times 449.2838)^2 \times 7.5863) + 1/(((1-0.9314^{1/s_4}) \times 446.6193)^2 \times 7.5473) + 1/(((1-0.9109^{1/s_5}) \times 452.5821)^2 \times 7.5771)] / 5 = 0.0004533$, and $W_{d(AB)} = [1/(((0.9325^{1/s_1}) \times 7.6136)^2 \times 458.7256) + 1/(((0.9261^{1/s_2}) \times 7.5881)^2 \times 453.5790) + 1/(((0.9367^{1/s_3}) \times 7.5863)^2 \times 449.2838) + 1/(((0.9314^{1/s_4}) \times 7.5473)^2 \times 446.6193) + 1/(((0.9109^{1/s_5}) \times 7.5771)^2 \times 452.5821)] / 5 = 0.0000416$.

3. Calculate the between-imputation variance for the numerator and denominator of the F -ratio using Eqs. 10 and 13. Substituting values in these algebraic expressions yields $B_{n(AB)} = [((1/((1-0.9325^{1/s_1}) \times 458.7256)) - 0.0582385)^2 + ((1/((1-0.9261^{1/s_2}) \times 453.5790)) - 0.0582385)^2 + ((1/((1-0.9367^{1/s_3}) \times 449.2838)) - 0.0582385)^2 + ((1/((1-0.9314^{1/s_4}) \times 446.6193)) - 0.0582385)^2 + ((1/((1-0.9109^{1/s_5}) \times 452.5821)) - 0.0582385)^2] / (5 - 1) = 0.0000568$, and

$$B_{d(AB)} = [(((1/(0.9325^{1/s^1}) \times 7.6136)) - 0.1371957)^2 + ((1/(0.9261^{1/s^2}) \times 7.5881)) - 0.1371957)^2 + ((1/(0.9367^{1/s^3}) \times 7.5863)) - 0.1371957)^2 + ((1/(0.9314^{1/s^4}) \times 7.5473)) - 0.1371957)^2 + ((1/(0.9109^{1/s^5}) \times 7.5771)) - 0.1371957)^2] / (5-1) = 9.8601E-7.$$

4. Calculate the F -ratio test proposed using Eq. 14. Substituting values in this algebraic expression yields $F_{MI-MBF(AB)} = 0.1371957/0.0582385 = 2.3557579$.

5. Calculate the numerator and denominator degrees of freedom of this F using Eq. 16. Substituting values in this algebraic expression yields $df_{n(AB)} = (2 \times 0.0582385^2) / (2 \times 0.0004533 + (1+1/5) \times 0.0000568) = 6.9583614$, and $df_{d(AB)} = (2 \times 0.1371957^2) / (2 \times 0.0000416 + (1+1/5) \times 9.8601E-7) = 445.75253$.

In order to save space, we do not report these real datasets, but they are available from the first author on request, and the SAS codes used to perform the previous calculations are provided in the [Appendix](#).

Discussion

In this paper, we considered a standard homoscedastic MANOVA model and proposed an extension of this test in which the missing data problem is tackled using multiple imputation of the missing values, and each imputed dataset is analyzed using the standard MANOVA for complete data. For pooling the results of the standard MANOVA approach, combination rules have been developed by adapting the method of Raghunathan (2016) and Raghunathan and Dong (2011) in the context of ANOVA, and that of Vallejo et al. (2018) in the context of repeated measures. We ran a Monte-Carlo simulation to evaluate the performance of the proposed approach. This was done by comparing the power and robustness of the standard MANOVA without missing data, the listwise deletion of missing data followed by the standard MANOVA model (i.e., LD-MANOVA), and the multiple imputation of missing data followed by the MANOVA model (i.e., MI-MANOVA). Furthermore, we revisited the MBF approach and proposed an extension of this procedure to deal with the missing data problem and potential heterogeneity of covariance matrices. Again, to examine the merits of this approach, we compared the results of the MBF procedure without missing data, listwise deletion of missing data followed by the MBF procedure (i.e., LD-MBF), and multiple imputation of missing data followed by the MBF test (i.e., MI-MBF).

A major result from this study is that the investigated methods can adequately control the type I error rate at the nominal level in the presence of the MCAR missing data mechanism when both normality and homogeneity of covariance matrices were present. That is, the false rejection rate was controlled in all conditions investigated for the MANOVA without missing data, LD-MANOVA, and MI-MANOVA.

Virtually identical results were obtained for the MBF procedure without missing data, LD-MBF, and MI-MBF when the covariance matrices were unequal but normality held. Consequently, any apparent power advantage of one procedure over another cannot be due to differences in type I error rates.

With respect to power, the situation is substantially different. Under normality and equality of covariance matrices, the simulation results indicated that the MANOVA without missing data had the highest power for the main and interaction effects, as expected. The power of the proposed MI-MANOVA approach was only slightly lower. Notably lower power was observed for the LD-MANOVA approach. Averaged across all the manipulated conditions, we found that the advantage of the standard MANOVA approach without missing data over the MI-MANOVA approach was less than 3 percentage points. The simulation study also indicated that, under the conditions we simulated here, the MI-MANOVA model was substantially more powerful than the LD-MANOVA model. The average difference between the two procedures was approximately 13.1 percentage points for $p = 2$ and approximately 28.8 percentage points for $p = 4$. For departures from homogeneity of covariance matrices, the same pattern of results was obtained. Specifically, the MI-MBF procedure performed only slightly worse than the MBF procedure without missing data in detecting non-null effects (average difference = 3.3 percentage points). However, the MI-MBF test was substantially more powerful than the LD-MBF test. The average difference between the two procedures was approximately 16.4 percentage points for $p = 2$ and approximately 29.9 percentage points for $p = 4$. When the missing data were generated by a MAR process, the differences were even greater. This set of results is available from the first author on request. In conclusion, the MI-MBF analysis appears to be a reasonable approach if the missingness only depends on the observed part of the data (i.e., if MAR holds).

In summary, when the researcher can ignore the reasons for missing data in testing multivariate hypothesis, multiple imputation is almost always a preferred choice over listwise deletion. This recommendation is based on the superior power one will achieve by using the multiple imputation method, which suffers only modest loss of power with respect to the complete data analysis, in comparison to the listwise deletion method. That is, although the LD-MANOVA and LD-MBF procedures are able to provide very effective type I error control, they always give less power than the MI-MANOVA and MI-MBF procedures. Hence, on the basis of our results, we recommend using listwise deletion with caution unless the sample size is very large and/or the missing data rate very small.

It is important to note that these findings are specific to the parameter sets employed and factors examined in our simulation study. Other alternative parameter sets or other factors could give different results. Thus, before proceeding with the generalization of the two-stage approach developed in

the current work, in future research it would be very useful to examine its performance under some of the conditions which here were maintained constant as they were not the focus of this article (e.g., the number of imputations, the imputation method used to create complete datasets, or the degree of sample size imbalance). The effects of some of these variables are currently unknown. Although this means that the recommendations should fit the specific conditions we manipulated, we believe they sufficiently probed the effects of the examined factors, and they also permit generalizations across a broad range of conditions likely to be encountered by behavioral science researchers (e.g., bigger sample sizes, larger degrees of covariance matrices inequality, or different rates of missingness at each level of the row main effect).

Limitations and future research

The present paper suggests that the rules developed for pooling the results obtained from datasets in which the missing data have been imputed perform reasonably well and may be of practical use for the MANOVA and MBF procedures. Despite this, both the MI-MANOVA and MI-MBF procedures have some limitations that future research should take into account. A limitation associated with the actual study is related to the number of imputations performed. As Rubin (1987) and others (e.g., Schafer, 1997) have indicated, if the between-imputation variance is large, a relatively high number of imputations may be required to prevent different researchers from obtaining different numbers from the same data using the same method (i.e., to facilitate the reproducibility of the final results). When using multiple imputations, a question arises immediately: how many imputations are needed for a repeat analysis of the same data to produce similar results? Although a number of proposals have been made to address this problem (Kenward, & Carpenter, 2007; White, Royston, & Wood, 2011; Van Buuren 2012; Graham, Olchowski, & Gilreath, 2007), a good rule of thumb is that the number of imputations should be at least equal to the percentage of incomplete cases (Von Hippel, 2009). In the current research, this rule is only partially respected, and based on the empirical study that has been conducted, we believe that by increasing the number of imputations the statistical power can be improved. At the same time, the probability of reproducing essentially the same results in a repeat analysis of the same data also increases.

In connection with the above, another question of central interest in the MI procedure is the specification of the imputation model. Several MI techniques have been proposed to impute missing continuous data with an arbitrary pattern of missingness, including the joint modeling (JM) approach and the fully

conditional specification (FCS) approach. JM imputes data by drawing missing values simultaneously for all incomplete variables using a multivariate distribution, whereas FCS imputes data on a variable-by-variable basis using a set of univariate conditional distributions. Although the performance of these MI approaches has been repeatedly compared in performing regression analyses (e.g., McNeish, 2017; Mistler & Enders, 2017; Kropko et al., 2014; Lee and Carlin, 2010; Van Buuren, 2007), there has not been a comprehensive evaluation of the relative performance of these methods in the homoscedastic and heteroscedastic MANOVA context. Thus, the perspective of JM in comparison with FCS is partially lacking in the literature.

A further limitation of this study is that we only investigated the performance of the proposed solution when the cell covariance matrices were equal (MI-MANOVA) and unequal (MI-MBF) and the degree of sample size imbalance moderate. It would be useful to confirm whether the two-stage approach developed in the current work continues to perform well with a larger cell size inequality.

Additionally, we did not evaluate the consequences of violating the population normality assumption. The effect of non-normality on the type I and type II error rates will not be of much consequence in the case of near-normal populations. However, the presence of a fair degree of skewness and/or kurtosis, as is not uncommon in educational, health, and psychological studies (see, e.g., Bono et al., 2017; Cain et al., 2017), leads to a more conservative alpha level and, thus, to more demanding sample size requirements. Also, the current study does not include covariates in the analytic model. However, it is often the case that there exist covariates that are related to the dependent variables. The inclusion of the covariates in the model leads to a decrease in the determinant of the error covariance matrix and thus it increases the power of the design. Future research could also focus on examining the performance of these procedures in analyzing multivariate analysis of covariance models.

Finally, another limitation to be noted is that none of the proposed procedures has been implemented in the major statistical packages. However, it should be stressed that a SAS/IML program to implement these approaches is available upon request from the first author. Despite these limitations, researchers should be comfortable using the proposed combination rules for the handling of missing values in multivariate data analysis, especially under conditions that are not optimal for parametric likelihood-based methods (e.g., the multivariate mixed model).

Acknowledgements We are grateful to the Associate Editor and reviewers for their constructive suggestions on an earlier draft of this manuscript. This research was supported by grant no. PGC2018-101574-B-I00 (AEI/FEDER, UE) from the Spanish Ministry of Science, Innovation, and Universities.

Appendix

SAS Program for pooling the results from MBF analyses on multiply imputed datasets

***The SAS program reproduced the same results as the significance test for the interaction effect with five imputed datasets, which have been described previously;**

```
proc iml;
lmda=0.9325260; lmdb=0.9261456; lmdc=0.9367438; lmdd=0.9314166; lmde=0.9109146;
sga=1.9210784; sgb=1.9157644; sgc=1.9153852; sgdd=1.9072227; sge=1.9122328;
dfna=7.6135694; dfnb=7.5881189; dfnc=7.5863050; dfnd=7.5473545; dfne=7.5712431;
dfda=458.72560; dfdb=453.57903; dfdc=449.28377; dfdd=446.61927; dfde=452.58206;
m=5;

*Calculate the MI estimates of parameters for the numerator and denominator of the F-ratio using Equations 8 and 11, and the within-imputation variance for numerator and denominator of the F-ratio using Equations 9 and 12;
start combine1(lmdj,sgj,dfnj,dfdj,qnj,qdj,wnj,wdj)global(m);
  qnj=sum((((1-lmdj##(1/sgj))*dfdj))##-1)/m)//qnj);
  qdj=sum((((lmdj##(1/sgj))*dfnj))##-1)/m)//qdj);
  wnj=sum((((1-lmdj##(1/sgj))*dfdj))##2)*dfnj##-1)/m)//wnj);
  wdj=sum((((lmdj##(1/sgj))*dfnj))##2)*dfdj##-1)/m)//wdj);
finish;
call combine1(lmda,sga,dfna,dfda,qna,qda,wna,wda);
call combine1(lmdb,sgb,dfnb,dfdb,qna,qda,wna,wda);
call combine1(lmdc,sgc,dfnc,dfdc,qna,qda,wna,wda);
call combine1(lmdd,sgd,dfnd,dfdd,qna,qda,wna,wda);
call combine1(lmde,sge,dfne,dfde,qna,qda,wna,wda);

*Calculate the between-imputation variance for the numerator and denominator of the F-ratio using Equations 10 and 13, the ratio F-test proposed using Equation 14, the numerator and denominator degrees of freedom of this F using Equation 16, and the probability of rejecting the null hypothesis of zero interaction effect;
start
combine2(lmdj,sgj,dfnj,dfdj,bnj,bdj,frj,dfrnj,dfrdj,pvrj)global(m,qna,
qda,wna,wda);
  bnj=sum((((1/((1-lmdj##(1/sgj))*dfdj))-qna))##2)/(m-1))//bnj);
  bdj=sum((((1/((lmdj##(1/sgj))*dfnj))-qda))##2)/(m-1))//bdj);
  frj=round((qda/qna),.001);
  dfrnj=round(((2*qna##2)/(2*wna+(1+1/m)*bnj)),.001);
  dfrdj=round(((2*qda##2)/(2*wda+(1+1/m)*bdj)),.001);
  pvrj=round((1-probf(frj,dfrnj,dfrdj)),.001);
finish;
call combine2(lmda,sga,dfna,dfda,bna,bda,fra,dfrna,dfrda,pvra);
call combine2(lmdb,sgb,dfnb,dfdb,bna,bda,fra,dfrna,dfrda,pvra);
call combine2(lmdc,sgc,dfnc,dfdc,bna,bda,fra,dfrna,dfrda,pvra);
call combine2(lmdd,sgd,dfnd,dfdd,bna,bda,fra,dfrna,dfrda,pvra);
call combine2(lmde,sge,dfne,dfde,bna,bda,fra,dfrna,dfrda,pvra);
print fra dfrna dfrda pvra;
```

References

- Baker, T.B., Smith, S. S., Bolt, D.M., Loh, W.Y., Mermelstein, R., Fiore, M.C., ... Collins, L.M. (2017). Implementing clinical research using factorial designs: A primer. *Behavior Therapy*, 48, 567–580.
- Blanca, M.J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, 29, 552–557.
- Bono, R., Blanca, M.J., Arnau, J., & Gómez-Benito, J. (2017). Non-normal Distributions Commonly Used in Health, Education, and Social Sciences: A Systematic Review. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.01602>
- Bradley, J. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144–152.
- Brunner, E., Dette, H., & Munk, A. (1997). Box-type approximations in factorial designs. *Journal of the American Statistical Association*, 92, 1494–1503.
- Cain, M.K., Zhang, Z., & Yuan, K.H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, 49, 1716–1735.
- Cueli, M., Areces, D., García, T., Rodríguez, C., Vallejo, G., & González-Castro, P. (2019). Influence of initial mathematical competencies on the effectiveness of a classroom-based intervention. *British Journal of Educational Psychology*, 89, 288–306.
- Enders, C.K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods*, 8, 322–337.
- Garage, J., Mathew, T., & Weerahandi, S. (2004). Generalized p -values and generalized confidence regions for the multivariate Behrens–Fisher problem and MANOVA. *Journal of Multivariate Analysis*, 88, 177–189.
- Ginsburg, H.P., & Baroody, A.J. (2003). *Test of early mathematics ability* (3rd). Austin, TX: PRO- ED.
- González-Castro, P., Cueli, M., Cabeza, L., Alvarez-García, D., & Rodríguez, C. (2014). Improving basic math skills through integrated dynamic representation strategies. *Psicothema*, 26, 378–384.
- Graham, J.W., Olchowski, A.E., & Gilreath, T.D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Preventive Science*, 8, 206–13.
- Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Pooling ANOVA results from multiply imputed datasets: A simulation study. *Methodology*, 12, 75–88.
- Harrar, S.W., & Bathke, A.C. (2012). A modified two-factor multivariate analysis of variance: asymptotics and small sample approximations. *Annals of the Institute of Statistical Mathematics*, 64, 135–165.
- He, F., Mazumdar, S., Tang, G., Bhatia, T., Anderson, S.J., Dew, M.A., ... Reynolds, C.F. (2017). Non-parametric MANOVA approaches for non-normal multivariate outcomes with missing values. *Communications in Statistics-Theory and Methods*, 46, 7188–7200.
- Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, 67, 85–92.
- Kenward, M., & Carpenter, J. (2007). Multiple imputation: Current perspectives. *Statistical Methods in Medical Research*, 16, 199–218.
- Konietschke, F., Bathke, A.C., Harrar, S.W., & Pauly, M. (2015). Parametric and Nonparametric Bootstrap tests for General MANOVA. *Journal of Multivariate Analysis*, 140, 291–301.
- Krishnamoorthy, K., & Lu, F. (2010). A parametric bootstrap solution to the MANOVA under heteroscedasticity. *Journal of Statistical Computation & Simulation*, 80, 873–887.
- Kropko, J., Goodrich, B., Gelman, A., & Hill, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Political Analysis*, 22, 497–519.
- Lee, K.J., & Carlin, J.B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, 171, 624–632.
- 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 statistic and an F-reference distribution. *Journal of the American Statistical Association*, 86, 1065–73.
- Mardia, K.V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika*, 58, 105–121.
- McNeish, D. (2017). Missing data methods for arbitrary missingness with small samples. *Journal of Applied Statistics*, 44, 24–39.
- Meng, X.L., & Rubin, D.B. (1992). Performing likelihood ratio tests with multiply imputed data sets. *Biometrika*, 79, 103–111.
- Mertler, C.A., & Vannatta, R.A. (2017). *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation* (6th). New York: Routledge.
- Mistler, S.A., & Enders, C.K. (2017). A Comparison of joint model and fully conditional specification imputation for multilevel missing data. *Journal of Educational and Behavioral Statistics*, 42, 432–466.
- Nel, D.G. (1997). Tests for equality of parameter matrices in two multivariate linear models. *Journal of Multivariate Analysis*, 61, 29–37.
- Olson, C.L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 69, 894–908.
- Raghunathan, T. (2016). *Missing Data Analysis in Practice*. Boca Raton, FL: Chapman and Hall/CRC.
- Raghunathan, T., & Dong, Q. (2011). Analysis of variance from multiply imputed data sets. Unpublished manuscript, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan. Retrieved from <http://www-personal.umich.edu/~teraghu/Raghunathan-Dong.pdf>.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Rubin, D.B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366–374.
- SAS Institute, Inc (2018). *SAS/STAT® 15.1 user's guide*. Cary, NC: SAS Institute, Inc.
- Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- Shieh, S. (2003). A comparative study of power and sample size calculations for multivariate general linear models. *Multivariate Behavioral Research*, 38, 285–307.
- Vallejo, G., & Ato, M. (2006). Modified Brown-Forsythe procedure for testing interaction effects in split-plot designs. *Multivariate Behavioral Research*, 41, 549–578.
- Vallejo, G., & Ato, M. (2012). Robust tests for multivariate factorial designs under heteroscedasticity. *Behavior Research Methods*, 44, 471–489.
- Vallejo, G., Arnau, J., & Ato, M. (2007). Comparative robustness of recent methods for the analysis of multivariate repeated measures designs. *Educational & Psychological Measurement*, 67, 410–432.
- Vallejo, G., Fernández, M.P., Livacic-Rojas, P.E., & Tuero-Herrero, E. (2018). Data analysis of incomplete repeated measures using a multivariate extension of the Brown-Forsythe procedure. *Psicothema*, 30, 434–441.
- Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16, 219–42.
- Van Buuren, S., (2012). *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC.

- 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, 514–529.
- Van Ginkel, J.R., & Kroonenberg, P.M. (2014). Analysis of variance of multiply imputed data. *Multivariate Behavioral Research*, 49, 78–91.
- Van Ginkel J.R., Linting M., Rippe R.C.A. & Van der Voort A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, 3, 297–308.
- Von Hippel, P.T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, 39, 265–91.
- White, I.R., Royston, P., & Wood, A.M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30, 377–99.
- Xu, L.W. (2015). Parametric bootstrap approaches for two-way MANOVA with unequal cell sizes and unequal cell covariance matrices. *Journal of Multivariate Analysis*, 133, 291–303.
- Zhang, J.T. (2011). Two-way MANOVA with unequal cell sizes and unequal cell covariance matrices. *Technometrics*, 53, 426–439.
- Zhang, J.T. & Xiao, S. (2012). A note on the modified two-way MANOVA tests. *Statistics and Probability Letters*, 82, 519–527.
- Zhang, J.T., Zhou, B., Guo, J., & Liu, X. (2016). A modified Bartlett test for heteroscedastic two-way MANOVA. *Journal of Advanced Statistics*, 1, 94–108.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.