Econometrics and Statistics 25 (2023) 51–65

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

Combining rules for F- and Beta-statistics from multiply-imputed data

Ashok Chaurasia*

School of Public Health Sciences, Faculty of Health, University of Waterloo, Waterloo, Ontario N2L3G1, Canada

ARTICLE INFO

Article history:
Received 16 October 2020
Revised 30 August 2021
Accepted 31 August 2021
Available online 17 September 2021

Keywords:
combining F- and Beta-statistics
combining R2
F-tests
missing data
multiple imputation
linear models

ABSTRACT

Missing values in data impede the task of inference for population parameters of interest. Multiple Imputation (MI) is a popular method for handling missing data since it accounts for the uncertainty of missing values. Inference in MI involves combining point and variance estimates from each imputed dataset via Rubin's rules. A sufficient condition for these rules is that the estimator is approximately (multivariate) normally distributed. However, these traditional combining rules get computationally cumbersome for multicomponent parameters of interest, and unreliable at high rates of missingness (due to an unstable variance matrix). New combining rules for univariate F- and Beta-statistics from multiply-imputed data are proposed for decisions about multicomponent parameters. The proposed combining rules have the advantage of being computationally convenient since they only involve univariate F- and Beta-statistics, while providing the same inferential reliability as the traditional multivariate combining rules. Simulation study is conducted to demonstrate that the proposed method has good statistical properties of maintaining low type I and type II error rates at relatively large proportions of missingness. The general applicability of the proposed method is demonstrated within a lead exposure study to assess the association between lead exposure and neurological motor function.

1. Introduction

Statistical inference is the task of making statements for the unknown population parameter of interest via its point and variance estimates obtained from a large and representative sample. However, this task is impeded when the sampled data have missing values. For example, consider the Landrigan et al. (1975) study where interest lies in assessing the association between blood-lead levels and neurological motor function. Specifically, one of the objectives is to assess the strength of the association between neurological motor function and level of lead exposure in children, while the second objective is to assess the same strength of association after controlling for the confounding effects of subject's age and sex. The first objective is assessed via the coefficient of determination (R^2) and the second is addressed via a multiparameter partial F-test. However, the data are plagued with missing values in the range of 20-30% for the models of interest, with a sample size of 124 subjects. Given this substantial amount of missingness, Multiple Imputation (MI) could be beneficial for inference. MI is one of the modern methods for handling missing data that has gained traction in applied research (MI; Rubin, 1987). In MI, under certain assumptions about the missing mechanism and pattern, multiple imputed values are simulated to serve

* Corresponding author.
E-mail address: a4chaura@uwaterloo.ca

https://doi.org/10.1016/j.ecosta.2021.08.013

as surrogates for the missing cells, thereby creating multiple imputed datasets as surrogates for the original incomplete data. Each imputed dataset is then analysed and their results (point and variance estimates) are combined using Rubin's traditional combining rules for inference.

Rubin's traditional combining rules rely on (i) sufficient condition of normality of the statistic of interest, and (ii) the stability of the total variance estimate. These conditions are questionable when the estimate for the parameter of interest is non-normal or multicomponent, and the data are plagued with missing values. For example, if the parameter of interest is multicomponent (i.e., $\boldsymbol{\theta} \in \mathbb{R}^p$), then the combining stage of MI becomes computationally cumbersome and can yield an unstable variance matrix (Rubin, 1987). This computational burden comes from the singularity of the variance matrix at high percentages of missingness. Consider the lead exposure study, where the parameters of interest are non-normal multiple/partial R^2 and mutiparameter partial F-tests in linear regression. For the measure of R^2, existing methods either lack theoretical justification (Harel, 2009; Chaurasia and Harel, 2015), or are unreliable (van Ginkel, 2019, see p.525). In terms of conducting multiparameter partial F-tests, literature is lacking since the traditional MI-methods only provide inference akin to type III tests, i.e., tests of each *single* effect given *all* other effects are included in the model.

In summary, for non-normal statistics of interest, existing methods in MI are either unreliable (e.g. Harel, 2009; Chaurasia and Harel, 2015; van Ginkel, 2019), or computationally cumbersome (e.g. Meng and Rubin, 1992; Raghunathan and Dong, 2011; Grund et al., 2016). Therefore, in MI it would be convenient if computation with multivariate statistics could be reduced to computation with univariate statistics when making inference for multicomponent parameters, since this would (i) reduce the computation cost for inference for a p-dimensional entity from $\mathcal{O}(p^3)$ to $\mathcal{O}(1)$, and (ii) eliminate reliance on unverifiable assumptions (like *equal and proportional loss of information*) to superficially stabilize the variance matrix. Such a combining technique would usher MI into research areas to address objectives that were previously either non-existent, or inaccessible with traditional MI methods.

This article proposes a computationally convenient, statistically reliable, and easily adaptable methodology which combines univariate F-tests rather than multivariate point and variance estimates from each of the imputed datasets. Section 2 provides the necessary preliminaries – the relation between F- and Beta-distributed random variables, transformation of F- and Beta-distributed random variables to approximate normal distribution, and motivation for the proposed transformations. Section 3 provides a review of MI combining rules for univariate and multivariate statistics, followed by the proposed rules for combining F- and Beta-statistics. The inferential ability of the proposed methods is assessed via an extensive simulation study and compared to the inferential ability of Rubin's traditional multivariate method in Section 4. Section 5 lists a few examples of types of statistical tests that can be conducted with multiply-imputed data using the proposed methodology. In Section 6, the proposed method is applied to a lead exposure study to assess the the association between blood-lead levels and neurological motor functions. A summary of findings and closing remarks are given in Section 7.

2. Preliminaries

2.1. Relationship between F- and Beta-distributed random variables

Consider random variable $F \sim \mathrm{F}(p, q, \lambda)$ with p numerator degrees of freedom, q denominator degrees of freedom, and λ as the non-centrality parameter. Then, the transformation $B = \frac{pF}{q+pF}$ is Beta-distributed, i.e., $B \sim \mathrm{Beta}(a, b, \lambda)$, with the two shape parameters $a = \frac{p}{2}$, $b = \frac{q}{2}$. Similarly, for $B \sim \mathrm{Beta}(a, b, \lambda)$ with a and b as the two shape parameters, the transformation $F = \frac{q}{p}\left(\frac{B}{1-B}\right)$ is F-distributed, i.e., $F \sim \mathrm{F}(p, q, \lambda)$, with $p = 2a$ and $q = 2b$. Equivalently, the relationship between random variables F and B can be expressed via their respective cumulative distribution functions $\mathfrak{F}(\cdot)$ and $\mathfrak{B}(\cdot)$ (or $\mathfrak{B}_\star(\cdot)$), as follows

$$\mathfrak{F}(u) = \Pr(F < u; p, q, \lambda) = \underbrace{\Pr(B < w; a, b, \lambda)}_{=\mathfrak{B}(w)} = \underbrace{\Pr(B_\star < w_\star; b, a, \lambda)}_{=\mathfrak{B}_\star(w_\star)},$$

where $u = \frac{q}{p}\left(\frac{w}{1-w}\right)$, $w = \frac{pu}{q+pu}$, $w_\star = 1 - w$, and $B_\star = 1 - B \sim \mathrm{Beta}(b, a, \lambda)$.

2.2. Transformations

Fisher's-Z for Correlation Coefficient (ZR)

For bivariate normal data $(\boldsymbol{x}, \boldsymbol{y})$ of size n, Fisher (1915) proposed the following normal approximation for the observed sample correlation coefficient estimate r_{xy}:

$$\mathbb{T}(r_{xy}) \approx \Phi\left(\sigma_{\text{ZR}}^{-1}(x_{\text{ZR}} - \mu_{\text{ZR}})\right), \tag{1}$$

where $\mathbb{T}(\cdot)$ is the sampling cumulative distribution function for the sample correlation coefficient, $\Phi(\cdot)$ is the standard normal cumulative distribution function, $x_{\text{ZR}} = \operatorname{arctanh}(r_{xy})$, $\mu_{\text{ZR}} = \operatorname{arctanh}(\rho)$, and $\sigma_{\text{ZR}} = 1/\sqrt{n-3}$, with ρ as the population correlation coefficient between $\boldsymbol{x}$ and $\boldsymbol{y}$.

Fisher's-Z for F (ZF)

Aroian (1941, p.439) proposed a normal approximation for $F \sim \mathrm{F}(p, q, \lambda)$ such that

$$\mathfrak{F}(F) \approx \Phi\left(\sigma_{\text{ZF}}^{-1}(x_{\text{ZF}} - \mu_{\text{ZF}})\right), \tag{2}$$

where $x_{\text{ZF}} = 0.5\log(F)$, $\mu_{\text{ZF}} = 0.5\left(q^{-1} - p^{-1}\right)$, and $\sigma_{\text{ZF}} = \sqrt{0.5(q^{-1} + p^{-1})}$.

Severo-Zelen (SZ)

Severo and Zelen (1960) proposed a normal approximation for $F \sim \mathrm{F}(p, q, \lambda)$ such that

$$\mathfrak{F}(F) \approx \Phi(x_{\text{SZ}}), \tag{3}$$

where $x_{\text{SZ}} = \dfrac{(1-h)\sqrt[3]{aF} - (1-g)}{\sqrt{g + (aF)^{\frac{2}{3}}h}}$, $g = \dfrac{2(p+2\lambda)}{9(p+\lambda)^2}$, $h = \dfrac{2}{9q}$, and $a = \dfrac{p}{p+\lambda}$.

Wise (W)

For $B \sim \text{Beta}(a, b)$ where $a \geq b$, Wise (1960) proposed a normal approximation such that

$$\mathfrak{B}(B) \approx \Phi\left(\sigma_{\text{W}}^{-1}(x_{\text{W}} - \mu_{\text{W}})\right), \tag{4}$$

where $x_{\text{W}} = (-\log(B))^{\frac{1}{3}}$, $\mu_{\text{W}} = -\dfrac{A}{D}$ and $\sigma_{\text{W}} = \dfrac{1}{D}$, with $A = -3\sqrt{b}\left(1 - \dfrac{1}{9b}\right)$, $N = a + \dfrac{b}{2} - \dfrac{1}{2}$, and $D = 3\left[N\sqrt{b}\left(1 - \dfrac{(b-1)(b+\frac{1}{3})}{12N^2}\right)\right]^{\frac{1}{3}}$. For $B_* \sim \text{Beta}(a_*, b_*)$ where $a_* < b_*$, replace B with $1 - B_*$, a with b_*, and b with a_* (respectively) in Eq. (4).

Hodgson (H)

Hodgson (1968, see approximation C', p.10) proposed a transformation of R^2 using sample size (n) and number of explanatory variables in the model (k). However, since R^2 is Beta-distributed, simple algebraic equivalence between degrees of freedom and shape parameters of this distribution gives the following equivalent transformation. For $B \sim \text{Beta}(a, b)$ the normal approximation to test the null hypothesis of $\rho = \rho_0$ is

$$\mathfrak{B}_\star(B_\star) \approx \Phi\left(\sigma_{\text{H}}^{-1}(x_{\text{H}} - \mu_{\text{H}})\right), \tag{5}$$

where $x_{\text{H}} = \left[\dfrac{(a+b)(1-B_\star)}{b(1-\rho_0^2)}\right]^{\frac{1}{2}}$, $\mu_{\text{H}} = 1$, and $\sigma_{\text{H}} = \dfrac{1}{2}\sqrt{\dfrac{a + 2b\rho_0^2}{b(a+b)}}$.

2.3. Motivation

Motivation for the proposed method of using normal transformations of F- and Beta-distributed random variables is presented via two demonstrations – simple linear regression (SLR) and multiple linear regression (MLR). The SLR demonstration emphasizes proper application, while the MLR demonstration highlights the effectiveness of the aforementioned transformations in statistical decision making. In the SLR demonstration, interest lies in inference for the correlation coefficient between outcome $\boldsymbol{y}$ and predictor $\boldsymbol{x}$, denoted as $\rho_{xy} \in (-1, 1)$. Using Fisher's-Z transformation (Eq. 1) with the Pearson's correlation coefficient estimate $r_{xy} = \hat{\rho}_{xy} = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}}$ is a proper use of the transformation. However, application of Fisher's-Z transformation with the estimator $r = +\sqrt{R^2}$, where R^2 is the coefficient of determination, is wrong because the relationship between Pearson's correlation coefficient and the coefficient of determination is not one-to-one. This is being emphasized because existing literature in MI commits this fallacy (e.g. Harel, 2009; Chaurasia and Harel, 2015). Figure 1 shows that the correct application of ZR with r_{xy} (solid curve) has nominal Type I error rates – under null hypothesis ($\rho = 0$) the rate of rejecting this hypothesis is around 0.05, while for the alternative hypothesis ($\rho \neq 0$) the power approaches 1. In contrast, for ZR with the invalid estimate r (dashed curve) the Type I error rate is around 0.10, which is twice the expected rate of 0.05; this demonstrates the inferential unreliability of estimator $r = +\sqrt{R^2}$. For this reason Schafer (1997, p.109) proposed using ZR in MI only to estimate the correlation coefficient between a pair of imputed variables. The incorrect application of Fisher's-Z transformation has huge implications in MLR.

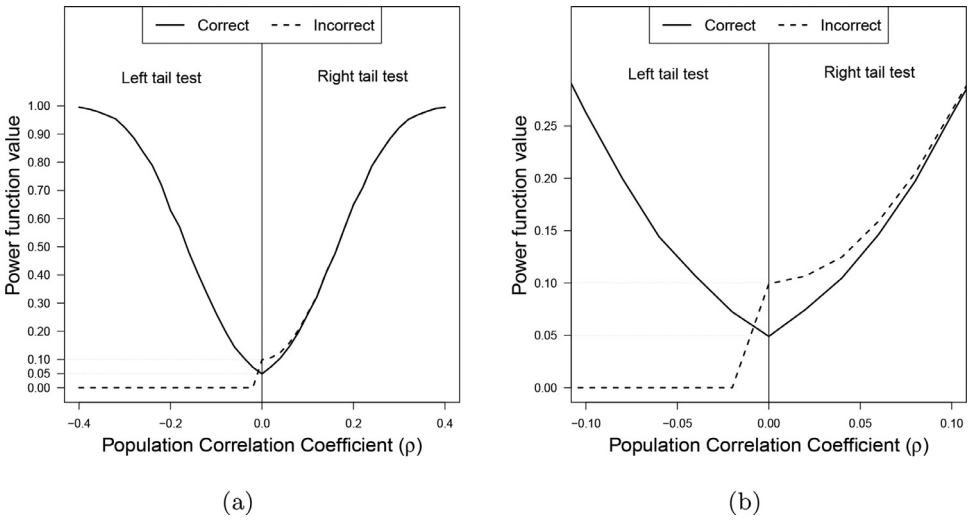

Fig. 1. Power function values under correct and incorrect usage of Fisher's-Z transformation of correlation coefficient in SLR for (a) $\rho \in [-0.4, -0.4]$, and (b) $\rho \in [-0.1, -0.1]$.

For the MLR demonstration, Figure 2 shows the power function for each of the aforementioned transformations for data with no missing values. For this illustration, $G = 1000$ random values were generated such that $F \sim \mathrm{F}(p = 7, q = 94, \lambda)$, for $\lambda = 0, 1, \ldots, 30$, to represent F-statistics from MLR with 7 predictor variables and sample size of 102. Equivalently, there are 1000 R^2 values that are Beta-distributed such that $R^2 = B \sim \mathrm{Beta}(a = 3.5, b = 47, \lambda)$. The aforementioned transformations were applied, their corresponding decisions (to reject the null hypothesis at significance level of $\alpha = 0.05$) recorded, and then averaged (over G) to obtain their respective power function values at each of the λ values. Figure 2a confirms the unreliability of the incorrect application of ZR with $r = \sqrt{R^2}$ (- -•- -) since its type I error rate (i.e., power at $\lambda = 0$) is around 0.80, instead of the expected value of 0.05. This disqualifies it as an α-level test for inference using F-tests (or Beta-tests) and therefore it does not appear in Figure 2b. In contrast, other transformations (when properly applied) serve as α-level tests, with ordering from most to least powerful test as Hodgson > SZ ≃ Wise > ZF (note, Hodgson (- -×- -) is closest to the F-True (▬), while SZ (- -⊙- -) and Wise (- -+- -) are similar). Given this demonstration in data with no missing values, existing methods with multiply imputed data that incorrectly apply ZR transformation with $r = \sqrt{R^2}$ (e.g., Harel, 2009; Chaurasia and Harel, 2015) should be avoided due to their higher than expected type I error rates.

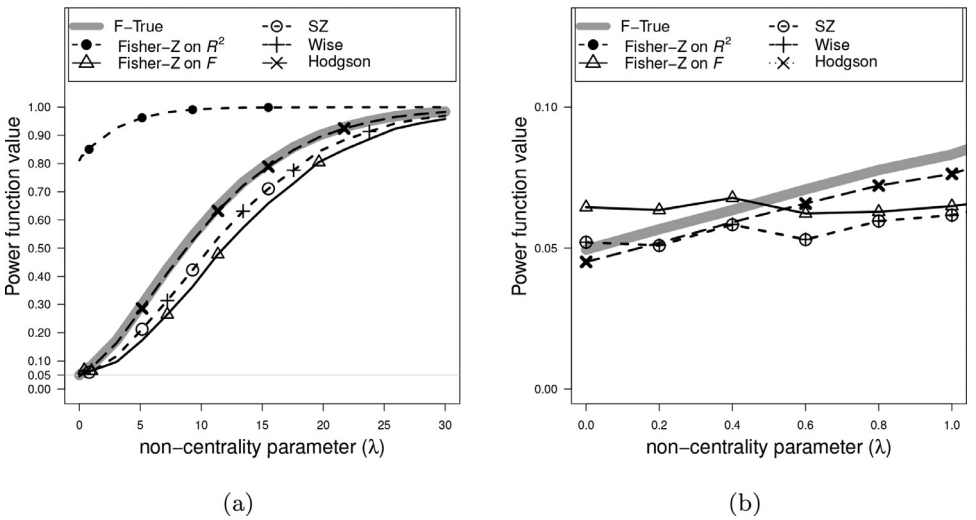

Fig. 2. Power function values for transformations of coefficient of determination in MLR, such that $\mathrm{F}(p = 7, q = 94, \lambda)$ for (a) $\lambda \in [0, 30]$, and (b) $\lambda \in [0, 1]$.

3. Proposed methodology

3.1. Multiple Imputation combining rules for univariate normal statistics

Suppose incomplete data $\mathfrak{D}$ is imputed m times – $\mathcal{D}_1, \cdots, \mathcal{D}_m$, and the parameter of interest is $\theta \in \mathbb{R}^1$. Let the pair $\left(Q_j, U_j\right)$ denote the point and variance estimates for θ from $\mathcal{D}_j$, $j = 1, \ldots, m$. If $(Q_j - \theta) \dot{\sim} \mathcal{N}(0, U_j)$, then Rubin's combining rules (Rubin, 1987) for inference for θ are

$$\text{(Point estimate) } \bar{Q} = m^{-1} \sum_{j=1}^{m} Q_j, \tag{6}$$

$$\text{(Within-data variance) } \bar{U} = m^{-1} \sum_{j=1}^{m} U_j, \tag{7}$$

$$\text{(Between-data variance) } \bar{V} = (m-1)^{-1} \sum_{j=1}^{m} (Q_j - \bar{Q})^2, \text{ and} \tag{8}$$

$$\text{(Total variance) } T = \bar{U} + (1 + m^{-1})\bar{V}. \tag{9}$$

The univariate test statistic for the null hypothesis $H_0 : \theta = \theta_0$ is $t = T^{-\frac{1}{2}}(\bar{Q} - \theta_0)$, which is t-distributed with ν (denominator) degrees of freedom. For inference via confidence intervals (CI), the $(1-\alpha) \times 100\%$ CI for θ is $\bar{Q} \pm t_{\frac{\alpha}{2},\nu}\sqrt{T}$, where $t_{\frac{\alpha}{2},\nu}$ is the $(50 \times \alpha)^{\text{th}}$ percentile of the t-distribution with ν degrees of freedom. Equivalently, the above hypothesis test for θ can be conducted via the F-statistic $\mathcal{F}_{Rubin} = t^2$, which is F-distributed with numerator and denominator degrees of freedom 1 and ν, respectively. Since the original data are incomplete, the denominator degrees of freedom (d.d.o.f) ν is unknown and must be estimated. Existing literature on the estimators of ν includes (but is not limited to) Rubin (1987); Li et al. (1991); Barnard and Rubin (1999), and Reiter (2007). The estimates for ν as proposed by Rubin (1987) and Li et al. (1991) tend to exceed the d.d.o.f that would be available if the data were fully observed. Hence, estimates for ν by Barnard and Rubin (1999), and Reiter (2007) were proposed since they are guaranteed to be less that the d.d.o.f that would be available if the data was fully observed. The formulae for these estimates are presented in Section A of supplementary material.

3.2. Multiple Imputation combining rules for multivariate normal statistics

In the case of multicomponent parameter $\boldsymbol{\theta} \in \mathbb{R}^p$, Rubin's traditional combining rules involve vector estimate $\mathbf{Q}$ $(\in \mathbb{R}^p)$ and the $p \times p$ variance-covariance matrix $\mathbf{T}$, where this matrix can be singular (at higher rates of missingness), and hence unstable for constructing test statistics (Rubin, 1987; Schafer, 1997; Allison, 2002). For this reason, Rubin proposed the alternative total variance estimate $\mathbf{T}^* = (1+g)\bar{\mathbf{U}}$, where $g = (1 + m^{-1})\ trace(\mathbf{B}\bar{\mathbf{U}}^{-1})/p$ is the relative increase in variance in estimating $\boldsymbol{\theta}$ under the *proportional and equal loss of information* (due to missingness) across the components of $\boldsymbol{\theta}$ (Rubin, 1987, see p.96 for details). This unverifiable assumption, along with instability of $\mathbf{T}$, becomes problematic with increasing rates of missingness and p. Therefore, the next Section proposes methodology that avoids the aforementioned problems associated with combining multivariate normal statistics by taking advantage of transformations from Section 2.2.

3.3. Combining F- and Beta-statistics from imputed data

In this section, it is proposed that the inference for a multidimensional parameter of interest $\boldsymbol{\theta}$ from imputed data be assessed by combining univariate F-statistics $f_1, \ldots, f_m$ as an alternative to Rubin's combining rules for multivariate statistics. Specifically, it is proposed that transformations (from Section 2.2) of F-statistics, $f_1, \ldots, f_m$, be combined using Rubin's univariate combining rules. These transformations (to approximately Normal distributions) meet the normality condition for applicability of Rubin's combining rules. Note that with Rubin's univariate combining rules, there is no need to superficially stabilize the variance T or assume *proportional and equal loss of information*.

For illustration of the proposed methodology, consider F-statistics, $f_1, \ldots, f_m$, with common parameters p and q from imputed datasets each of size n. Suppose *Fisher's-Z for F* is the transformation of interest. Then, the steps for the new combining rules are as follows:

1. Transform F-statistics to approximate normal under ZF: $Q_j^{(ZF)} = 0.5 \log(f_j)$ [from Eq. (2)].
2. Obtain θ_0 for univariate combining rules under the transformation ZF: $\theta_0^{(ZF)} = 0.5\left(q^{-1} - p^{-1}\right)$ [from $\mu_{_{ZF}}$ of Eq. (2)].

3. Obtain the variance estimate under the transformation ZF: $U_j^{(\mathrm{ZF})} = 0.5(q^{-1} + p^{-1})$ [from σ^2_{ZF} of Eq. (2)].
4. Use estimates from Steps 1–3 in univariate combining rules of Section 3.1 to obtain the set

$$\mathcal{S}_{\mathrm{ZF}} = \left\{ \bar{Q}_{(\mathrm{ZF})},\ \bar{U}_{(\mathrm{ZF})},\ T_{(\mathrm{ZF})},\ \mathcal{F}_{(\mathrm{ZF})} = \frac{(\bar{Q}_{(\mathrm{ZF})} - \theta_0^{(\mathrm{ZF})})^2}{T_{(\mathrm{ZF})}},\ \nu^{(\mathrm{ZF})} \right\}.$$

In this set, the overall MI statistic $\mathcal{F}_{(\mathrm{ZF})}$ with degrees of freedom 1 and $\nu^{(\mathrm{ZF})}$ (estimated using formulae from Section A of supplementary material) is used for inference for multicomponent $\boldsymbol{\theta}$. Similarly, we have sets $\mathcal{S}_{\mathrm{H}}$, $\mathcal{S}_{\mathrm{SZ}}$, $\mathcal{S}_{\mathrm{W}}$ for transformations of Hodgson, Severo-Zelen, and Wise, respectively. In these sets, one can also back transform $\bar{Q}_{[\cdot]}$ to obtained the MI-combined F- or Beta- statistic as follows:

- Under Fisher's-Z for F from Section 2.2, the combined F-statistic from MI is $F_{(\mathrm{ZF})} = \exp\{2\bar{Q}_{(\mathrm{ZF})}\}$.
- Under Severo-Zelen with null hypothesis $\lambda = 0$ from Section 2.2, the combined F-statistic from MI is $F_{(\mathrm{SZ})} = a^{-1}\eta^3$, where η is the principal root from the following quadratic equation solution set

$$\frac{(1-h)(1-g) \pm \bar{Q}^2_{(\mathrm{SZ})}\sqrt{g(1-h)^2 + h(1-g)^2 - \bar{Q}^2_{(\mathrm{SZ})}gh}}{(1-h)^2 - \bar{Q}^2_{(\mathrm{SZ})}h}.$$

- Under Wise from Section 2.2, the combined Beta-statistic from MI is $B_{*_{\mathrm{W}}} = \exp\left\{-\bar{Q}^3_{(\mathrm{W})}\right\}$.
- Under Hodgson for null hypothesis $\rho_0 = 0$ from Section 2.2, the combined Beta-statistic from MI is $B_{*_{\mathrm{H}}} = \frac{n-k-1}{k\left(\left[1-\left(\frac{n-k-1}{n-1}\right)\bar{Q}^2_{(\mathrm{H})}\right]^{-1} - 1\right)}$.

In this proposed methodology, the combining rules only involve one-dimensional statistics, and hence are computationally convenient, faster (compared to computation with vectors and inverse of matrices), and reliable, as they do not depend on unverifiable assumptions (like *proportional and equal loss of information*).

4. Monte Carlo Study

4.1. Full and Incomplete data simulation

Consider the MLR framework $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is the fully observed $n \times 1$ random outcome vector, $\mathbf{X}$ is the fully observed (fixed) $n \times p$ matrix of predictor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown (fixed) regression coefficients, and $\boldsymbol{\varepsilon}$ is the unobserved random noise vector from $\mathbf{N}_n(0, \sigma^2\mathbf{I})$ with unknown σ^2 and identity matrix $\mathbf{I}$. Suppose the data matrix $\mathfrak{D} = [\boldsymbol{y}\ \mathbf{X}]$. Then, for a given sample size n and p predictor variables, $\mathfrak{D}$ is generated from a $(p+1)$-variate normal distribution $\mathbf{N}_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \mathbf{0}$ is the mean vector and $\boldsymbol{\Sigma}$ is the variance-covariance matrix; specific values of $\boldsymbol{\Sigma}$ under the null and alternative hypothesis, denoted as Σ_0 and Σ_1 (respectively) for $p = \{3,\ 5\}$, are as follows:

$$\left(\boldsymbol{\Sigma}_0^{[p]}\right)_{ij} = \begin{cases} 1 & \text{if } i = j; \\ 0.05 & \text{if } i = 1, j = 2, p = 3; \\ 0.03 & \text{if } i = 1, j = 2, p = 5; \\ 0 & \text{if } i \neq j, i \neq 1, j \neq 2 \end{cases} \quad \left(\boldsymbol{\Sigma}_1^{[p]}\right)_{ij} = \begin{cases} 10 & \text{if } i = j, p = 3; \\ 6 & \text{if } i = 1, j = 2, p = 3; \\ 1 & \text{if } i = j, p = 5; \\ 0.65 & \text{if } i = 1, j = 2, p = 5; \\ 0 & \text{if } i \neq j, i \neq 1, j \neq 2. \end{cases} \tag{10}$$

Sample sizes of $n = \{20, 50, 100\}$ were considered in the simulation.

For a given percentage of missingness $\delta \in (0,\ 1)$, values in $\mathfrak{D}$ are made *missing at random* (MAR) as follows. Suppose i denotes the i^{th} case or observation of $\mathfrak{D}$ and u_δ represents the $(1-\delta)100^{th}$ percentile of $\boldsymbol{x}_3$. Then, for the cases where $x_{3i} \leq u_\delta$, y_i is randomly missing for half of the cases and x_{1i} is missing for the remaining cases. The following values for percentage of missingness were considered in our simulations: δ (in %) = 5 to 45 by 5. The corresponding incomplete data is denoted as $\mathfrak{D}_{\text{inc}}$.

The imputation phase/model assumes (i) data are MAR, (ii) a normal posterior predictive imputation model for $\boldsymbol{y}$, (iii) a predictive means matching model for $\boldsymbol{x}_1$, and (iv) includes main effects for all predictor variables in $\mathfrak{D}_{\text{inc}}$. This imputation scheme was performed m times to yield imputed datasets $\mathcal{D}_1, \ldots, \mathcal{D}_m$, with $m = 10$ to 100 by 10. Given this range of values for number of imputations, only results for $m = 100$ are presented/discussed since lower values of m are generally not recommended (e.g. Von Hippel, 2020; Graham et al., 2007; Von Hippel, 2005). Specifically, existing literature states

that the number of imputations is dependent on data/model features such as rate and pattern/mechanism of missingness, (imputation/analysis) model complexity, and parameter of interest.

4.2. Analysis and inference with multiple imputed data

In the analysis phase, the global F-test was conducted with $\mathfrak{D}$ and $\mathfrak{D}_{\text{inc}}$. The same test was conducted with imputed datasets $\mathcal{D}_1, \ldots, \mathcal{D}_m$ using proposed methods from Section 3 and their results are compared to those from Rubin's traditional multivariate combining method. Significance level of $\alpha = 0.05$ was used in all analyses. The process of simulating data, inducing missingness, imputation, analysis, and combining results was repeated 1000 times. A supplementary analysis was also conducted where Fisher's Z is incorrectly applied with $r = +\left|\sqrt{R^2}\right|$ to demonstrate its unreliability. Results of this supplementary analysis are presented and discussed in Section B of supplementary material.

4.3. Comments on simulation configurations

First, sample sizes of $n = \{20, 50, 100\}$ were chosen to study the resiliency of proposed methods in data of relatively small to large sample size (compared to p). Second, for $p = 3$, in the alternative hypothesis $(\boldsymbol{\Sigma}_1)_{12} = Cov(\boldsymbol{y}, \boldsymbol{x}_1) = 6$ (equivalent to correlation $\rho^{(1)}_{y,x_1} = 0.6$), while under the null hypothesis $\rho^{(0)}_{y,x_1} = (\boldsymbol{\Sigma}_0)_{12} = 0.05$ is smaller by a factor of 12. Similarly, $\rho^{(1)}_{y,x_1} = 0.65$ for $p = 5$, while under the null hypothesis $\rho^{(0)}_{y,x_1} = 0.03$ is smaller by a factor of 21.67. These correlation values suggest that under the null hypothesis, none of the predictors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_5$ are associated with $\boldsymbol{y}$, while under the alternative hypothesis only $\boldsymbol{x}_1$ is associated with $\boldsymbol{y}$. Third, the pattern of induced missingness in the data is severe (with severity depending on the value of δ) because it causes an entire portion of the data to be missing (specifically, the lower portion of $\mathfrak{D}$ with respect to ordered $\boldsymbol{x}_3$). Fourth, the assumptions made in the imputation phase are common in practice when dealing with incomplete data. Thus, these choices represent severe conditions under which the performance of the proposed method is assessed, and then compared to the performance of Rubin's combining rules for multicomponent parameters.

4.4. Monte Carlo Results

The observed Type I errors rates for $n = 20$ and $p = \{3, 5\}$ are visually summarized in Figure 3; these are based on values presented in Tables 1 and 2 when using ν_{BR}. For $n = 20$ and $p = 3$ in Figure 3a, the Type I error rates are low and within the range 0.021–0.108. However, only methods of F_{mult}, F^*_{mult} and the proposed method via Hodgson's transformation (when using ν_{BR}) qualify as α-level tests. Observed Type I errors are determined to be dissimilar from the nominal level of 0.05 by using the interval (0.036, 0.064), which was computed via confidence interval for proportion of 0.05 and number of simulation of 1000 as the size. The Type I errors rates for $n = 20$ and $p = 5$, as shown in Figure 3b, were similar. A few exceptions or boundary cases exist in these figures/tables for the three recommended methods, which is attributed to the sample size of $n = 20$ not being large enough relative to $p = \{3, 5\}$ at the various of levels of missingness considered in our simulations. We can see these exceptions or boundary cases disappear in larger sample sizes of $n = \{50, 100\}$ with $p = \{3, 5\}$, as shown in Tables C.1, C.2, E.1, E.2, and Figures D.1 and F.1 of supplementary material.

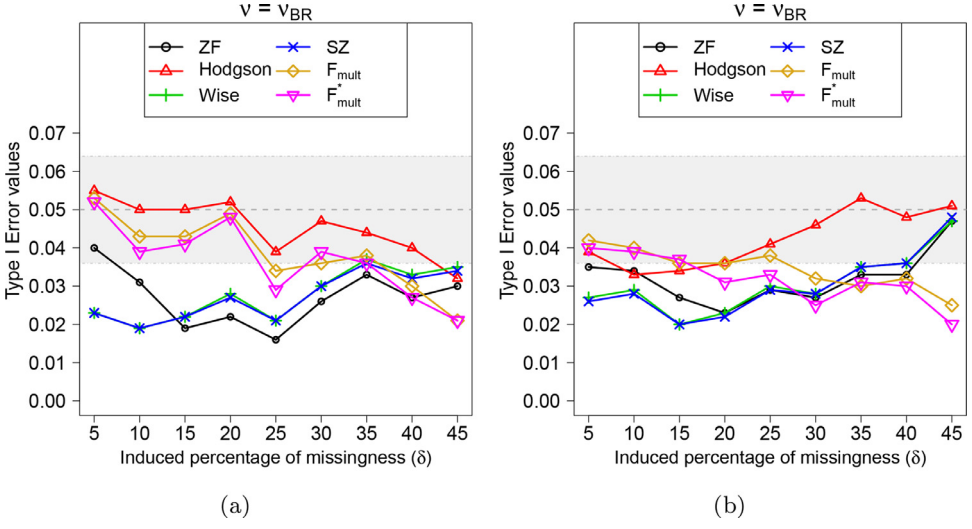

Fig. 3. Comparison of Type I error rates of proposed and traditional methods (F_{mult} and F^*_{mult}) via ν_{BR} for configurations (a) $n = 20$, $p = 3$, and (b) $n = 20$, $p = 5$. The shaded region is the region of indifference from the expected Type I error of 0.05.

Table 1
Type I error rates under transformations from Section 2.2 (using the five denominator degrees of freedom estimates) and their comparison to traditional Rubin's (multivariate) methods in MLR for $n = 20$, $p = 3$, and $m = 100$.

	Aroian's (ZF)						Hodgson's (H)				
δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}	δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}
5	0.065	0.065	0.065	0.040	0.040	5	0.072	0.072	0.072	0.055	0.058
10	0.060	0.060	0.060	0.031	–	10	0.066	0.066	0.066	0.050	0.052
15	0.040	0.040	0.040	0.019	0.021	15	0.070	0.070	0.070	0.050	–
20	0.044	0.044	0.044	0.022	0.024	20	0.072	0.072	0.072	0.052	–
25	0.034	0.034	0.034	0.016	0.017	25	0.067	0.067	0.067	0.039	–
30	0.050	0.050	0.050	0.026	0.027	30	0.078	0.078	0.078	0.047	–
35	0.053	0.053	0.053	0.033	0.034	35	0.070	0.070	0.070	0.044	–
40	0.050	0.050	0.050	0.027	0.029	40	0.076	0.076	0.076	0.040	–
45	0.053	0.053	0.053	0.030	0.032	45	0.071	0.071	0.071	0.032	–
	Wise's (W)						Severo-Zelen's (SZ)				
δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}	δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}
5	0.043	0.043	0.043	0.023	0.023	5	0.042	0.042	0.042	0.023	0.023
10	0.033	0.033	0.033	0.019	0.019	10	0.033	0.033	0.033	0.019	0.019
15	0.032	0.032	0.032	0.022	0.022	15	0.032	0.032	0.032	0.022	0.022
20	0.047	0.047	0.047	0.028	0.029	20	0.047	0.047	0.047	0.027	0.029
25	0.037	0.037	0.037	0.021	0.022	25	0.037	0.037	0.037	0.021	0.021
30	0.052	0.052	0.052	0.030	0.032	30	0.051	0.051	0.051	0.030	0.032
35	0.061	0.061	0.061	0.037	0.038	35	0.061	0.061	0.061	0.036	0.038
40	0.054	0.054	0.054	0.033	0.037	40	0.052	0.052	0.052	0.032	0.037
45	0.059	0.059	0.059	0.035	0.038	45	0.059	0.059	0.059	0.034	0.038
	F_{mult}						F^*_{mult}				
δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}	δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}
5	0.105	0.105	0.105	0.053	0.053	5	0.108	0.108	0.108	0.052	0.052
10	0.104	0.104	0.104	0.043	0.043	10	0.099	0.099	0.099	0.039	0.041
15	0.101	0.101	0.101	0.043	0.043	15	0.096	0.096	0.096	0.041	0.042
20	0.105	0.105	0.105	0.049	0.055	20	0.094	0.094	0.094	0.048	0.057
25	0.094	0.094	0.094	0.034	0.042	25	0.093	0.093	0.093	0.029	0.038
30	0.098	0.099	0.099	0.036	0.044	30	0.092	0.092	0.092	0.039	0.047
35	0.096	0.097	0.097	0.038	0.050	35	0.088	0.088	0.090	0.036	0.041
40	0.088	0.088	0.088	0.030	–	40	0.081	0.083	0.083	0.027	–
45	0.080	0.080	0.080	0.021	–	45	0.072	0.072	0.072	0.021	–

Type I error for fully observed data $\mathfrak{D}$ is 0.067.
– denotes that ν_{Rt} was negative in simulations.

Table 2
Type I error rates under transformations from Section 2.2 (using the five denominator degrees of freedom estimates) and their comparison to traditional Rubin's (multivariate) methods in MLR for $n = 20$, $p = 5$, and $m = 100$.

	Aroian's (ZF)						Hodgson's (H)				
δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}	δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}
5	0.057	0.057	0.057	0.035	0.035	5	0.045	0.045	0.045	0.039	0.039
10	0.057	0.057	0.057	0.034	0.034	10	0.050	0.050	0.050	0.033	0.035
15	0.043	0.043	0.043	0.027	0.027	15	0.047	0.047	0.047	0.034	0.035
20	0.046	0.046	0.046	0.023	0.024	20	0.049	0.049	0.049	0.036	–
25	0.050	0.050	0.050	0.029	0.033	25	0.062	0.062	0.062	0.041	0.045
30	0.051	0.051	0.051	0.027	0.027	30	0.073	0.073	0.073	0.046	0.048
35	0.058	0.058	0.058	0.033	0.036	35	0.081	0.081	0.081	0.053	–
40	0.073	0.073	0.073	0.033	0.038	40	0.096	0.096	0.096	0.048	–
45	0.078	0.078	0.078	0.047	0.052	45	0.107	0.107	0.107	0.051	–
	Wise's (W)						Severo-Zelen's (SZ)				
δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}	δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}
5	0.043	0.043	0.043	0.027	0.027	5	0.043	0.043	0.043	0.026	0.027
10	0.044	0.044	0.044	0.029	0.029	10	0.044	0.044	0.044	0.028	0.028
15	0.038	0.038	0.038	0.020	0.020	15	0.038	0.038	0.038	0.020	0.020
20	0.042	0.042	0.042	0.023	0.024	20	0.040	0.040	0.040	0.022	0.023
25	0.045	0.045	0.045	0.030	0.031	25	0.044	0.044	0.044	0.029	0.030
30	0.051	0.051	0.051	0.028	0.030	30	0.051	0.051	0.051	0.028	0.028
35	0.060	0.060	0.060	0.035	0.042	35	0.058	0.058	0.058	0.035	0.038
40	0.073	0.073	0.073	0.036	0.041	40	0.073	0.073	0.073	0.036	0.040
45	0.080	0.080	0.080	0.047	0.054	45	0.079	0.079	0.079	0.048	0.053
	F_{mult}						F^*_{mult}				
δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}	δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}
5	0.128	0.128	0.128	0.042	0.042	5	0.130	0.130	0.130	0.040	0.043
10	0.127	0.127	0.127	0.040	0.042	10	0.122	0.122	0.122	0.039	0.040
15	0.113	0.113	0.113	0.036	0.038	15	0.107	0.107	0.107	0.037	0.039
20	0.118	0.118	0.118	0.036	0.039	20	0.114	0.114	0.114	0.031	0.036
25	0.124	0.125	0.125	0.038	0.042	25	0.111	0.112	0.112	0.033	0.041
30	0.116	0.117	0.117	0.032	0.039	30	0.103	0.104	0.104	0.025	0.035
35	0.123	0.123	0.123	0.030	0.041	35	0.122	0.122	0.122	0.031	0.035
40	0.121	0.122	0.123	0.032	–	40	0.110	0.110	0.110	0.030	–
45	0.113	0.114	0.114	0.025	–	45	0.099	0.102	0.103	0.020	–

Type I error for fully observed data $\mathfrak{D}$ is 0.049.
– denotes that ν_{Rt} was negative in simulations.

Main conclusions from the above tables of Type I errors are as follows:

- For scalar or multicomponent parameters of interest, the estimate for ν proposed by Barnard and Rubin (1999) is recommended over the estimates provided by Rubin (1987), Li et al. (1991) and Reiter (2007). In our simulations, for small sample sizes and large percentage of missingness (at least 10%), the estimate for ν by Reiter (2007) had a tendency to be negative.
- The proposed combining rules using Hodgson's transformation has the expected Type I error rate of 0.05, similar to Rubin's traditional multivariate methods (F_{mult} and F^*_{mult}). Hence, these qualify as α-level tests. The transformations of ZF, Wise, and SZ are not recommended because they do not qualify as α-level tests, as illustrated in Figures 3 (D.1 and F.1 of supplementary material).
- Unlike Rubin's traditional multivariate methods (F_{mult} and F^*_{mult}), the proposed method based on Hodgson's transformation has the expected Type I error rate of 0.05 without the (i) computational burden of calculation with matrices (and their inverses, which may not exists at high rates of missingness), and (ii) reliance on unverifiable assumptions like *proportional and equal loss of information* across the components of a multivariate parameter of interest.

Since the proposed method based on Hodgson's transformation had expected Type I error rate of 0.05 and similar to the traditional cumbersome multivariate methods (F_{mult} and F^*_{mult}), these tests can next be compared in terms of their power values to determine superiority or equivalence.

Under the alternative hypothesis (where at least one of the predictor variables is associated with $\boldsymbol{y}$), the observed power values for $n = 20$ and $p = \{3, 5\}$ are visually summarized in Figure 4; these are based on values presented in Tables 3 and 4 for ν_{BR}. This figure shows that the proposed method (based on Hodgson's transformation) had similar power values to the traditional multivariate methods of F_{mult} and F^*_{mult} when using ν_{BR}. However, the proposed method attains these equivalent power values based on univariate statistics (only), and thus does not need the assumption of *equal and proportional loss of information* to (superficially) stabilize the variance.

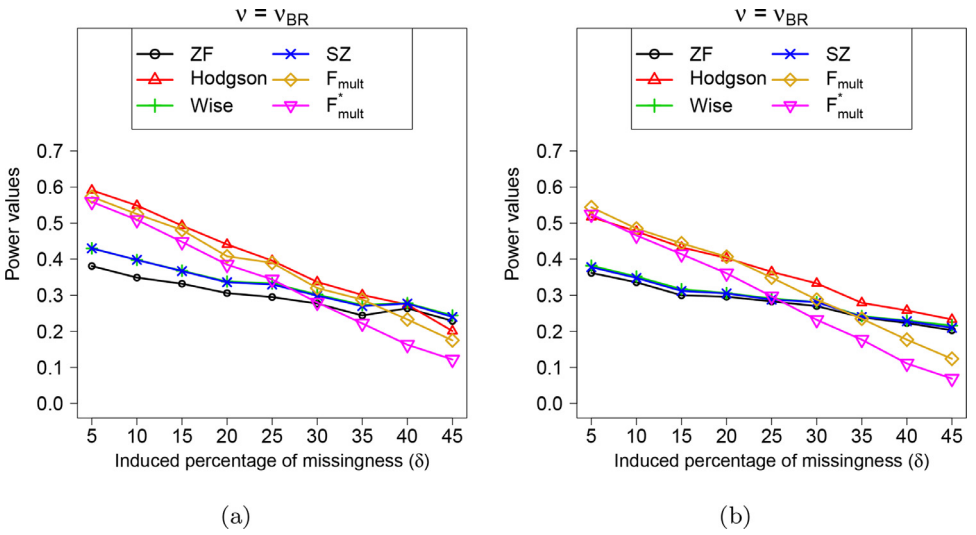

Fig. 4. Comparison of Power values of proposed and traditional methods (F_{mult} and F^{*}_{mult}) via ν_{BR} for configurations (a) $n = 20$, $p = 3$, and (b) $n = 20$, $p = 5$.

Table 3
Power values under transformations from Section 2.2 (using the five denominator degrees of freedom estimates) and their comparison to traditional Rubin's (multivariate) methods in MLR for $n = 20$, $p = 3$, and $m = 100$.

	Aroian's (ZF)						Hodgson's (H)				
δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}	δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}
5	0.472	0.472	0.472	0.381	0.381	5	0.637	0.637	0.637	0.591	0.594
10	0.458	0.458	0.458	0.349	0.354	10	0.598	0.598	0.598	0.549	0.558
15	0.425	0.425	0.425	0.332	0.339	15	0.570	0.570	0.570	0.493	0.512
20	0.395	0.395	0.395	0.306	0.312	20	0.518	0.518	0.518	0.441	–
25	0.391	0.391	0.391	0.295	0.309	25	0.497	0.497	0.496	0.396	–
30	0.363	0.363	0.363	0.277	0.287	30	0.447	0.447	0.447	0.337	0.348
35	0.340	0.340	0.340	0.244	0.255	35	0.422	0.422	0.422	0.301	–
40	0.339	0.339	0.339	0.264	0.267	40	0.404	0.404	0.404	0.274	–
45	0.324	0.324	0.324	0.229	0.245	45	0.386	0.386	0.385	0.201	–
	Wise's (W)						Severo-Zelen's (SZ)				
δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}	δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}
5	0.514	0.514	0.514	0.430	0.430	5	0.513	0.513	0.513	0.430	0.430
10	0.498	0.498	0.498	0.398	0.406	10	0.498	0.498	0.498	0.398	0.405
15	0.473	0.473	0.473	0.368	0.378	15	0.469	0.469	0.469	0.367	0.375
20	0.433	0.433	0.433	0.338	0.341	20	0.431	0.431	0.431	0.336	0.341
25	0.423	0.423	0.423	0.333	0.347	25	0.420	0.420	0.420	0.330	0.346
30	0.381	0.381	0.381	0.302	0.318	30	0.378	0.378	0.378	0.299	0.317
35	0.367	0.367	0.367	0.273	0.286	35	0.365	0.365	0.365	0.270	0.286
40	0.350	0.350	0.350	0.278	0.287	40	0.348	0.348	0.348	0.277	0.285
45	0.347	0.347	0.347	0.244	0.263	45	0.344	0.344	0.344	0.240	0.259
	F_{mult}						F^{*}_{mult}				
δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}	δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}
5	0.711	0.711	0.711	0.573	0.578	5	0.707	0.707	0.707	0.559	0.567
10	0.683	0.683	0.683	0.525	0.536	10	0.674	0.674	0.674	0.509	0.526
15	0.642	0.642	0.642	0.481	0.498	15	0.632	0.632	0.632	0.448	0.464
20	0.604	0.604	0.604	0.408	0.438	20	0.580	0.580	0.580	0.385	0.401
25	0.571	0.571	0.571	0.390	0.426	25	0.544	0.544	0.544	0.344	0.375
30	0.512	0.512	0.512	0.320	0.346	30	0.477	0.477	0.477	0.280	0.321
35	0.488	0.490	0.491	0.288	0.323	35	0.425	0.425	0.426	0.222	0.267
40	0.447	0.452	0.453	0.233	–	40	0.384	0.384	0.384	0.163	–
45	0.417	0.417	0.418	0.175	–	45	0.332	0.332	0.332	0.122	–

Power value for fully observed data $\mathfrak{D}$ is 0.663.
– denotes that ν_{Rt} was negative in simulations.

Table 4
Power values under transformations from Section 2.2 (using the five denominator degrees of freedom estimates) and their comparison to traditional Rubin's (multivariate) methods in MLR for $n = 20$, $p = 5$, and $m = 100$.

	Aroian's (ZF)						Hodgson's (H)				
δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}	δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}
5	0.464	0.464	0.464	0.362	0.364	5	0.586	0.586	0.586	0.517	0.524
10	0.428	0.428	0.428	0.336	0.337	10	0.555	0.555	0.555	0.477	0.482
15	0.406	0.406	0.406	0.300	0.310	15	0.519	0.519	0.519	0.433	0.449
20	0.402	0.402	0.402	0.296	0.303	20	0.490	0.490	0.490	0.403	0.419
25	0.383	0.383	0.383	0.283	0.291	25	0.474	0.474	0.474	0.365	0.387
30	0.361	0.361	0.361	0.270	0.283	30	0.449	0.449	0.449	0.333	–
35	0.337	0.337	0.337	0.239	0.242	35	0.405	0.405	0.405	0.279	–
40	0.339	0.339	0.339	0.223	0.240	40	0.413	0.413	0.413	0.258	–
45	0.336	0.336	0.336	0.203	0.228	45	0.395	0.395	0.395	0.233	–
	Wise's (W)						Severo-Zelen's (SZ)				
δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}	δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}
5	0.476	0.476	0.476	0.382	0.382	5	0.473	0.473	0.473	0.378	0.379
10	0.442	0.442	0.442	0.352	0.352	10	0.437	0.437	0.437	0.348	0.351
15	0.421	0.421	0.421	0.317	0.326	15	0.416	0.416	0.416	0.312	0.320
20	0.414	0.414	0.414	0.306	0.313	20	0.410	0.410	0.409	0.305	0.309
25	0.391	0.391	0.391	0.291	0.302	25	0.388	0.388	0.388	0.288	0.299
30	0.368	0.368	0.368	0.281	0.286	30	0.365	0.365	0.365	0.281	0.285
35	0.342	0.342	0.342	0.242	0.251	35	0.338	0.338	0.338	0.240	0.249
40	0.349	0.349	0.349	0.230	0.253	40	0.344	0.344	0.344	0.228	0.248
45	0.344	0.344	0.344	0.216	0.238	45	0.340	0.340	0.340	0.210	0.235
	F_{mult}						F^*_{mult}				
δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}	δ (%)	ν_{Rubin1}	ν_{Rubin2}	ν_{Li}	ν_{BR}	ν_{Rt}
5	0.756	0.756	0.756	0.544	0.550	5	0.753	0.753	0.753	0.524	0.534
10	0.716	0.716	0.716	0.485	0.501	10	0.705	0.705	0.705	0.466	0.478
15	0.686	0.686	0.686	0.444	0.462	15	0.666	0.666	0.666	0.414	0.432
20	0.643	0.644	0.644	0.407	0.438	20	0.612	0.612	0.613	0.361	0.394
25	0.605	0.605	0.605	0.349	0.389	25	0.565	0.565	0.565	0.297	0.341
30	0.548	0.550	0.550	0.288	0.342	30	0.502	0.502	0.502	0.232	0.272
35	0.508	0.509	0.510	0.235	0.288	35	0.456	0.456	0.458	0.177	0.218
40	0.475	0.478	0.479	0.177	–	40	0.411	0.411	0.412	0.111	–
45	0.425	0.430	0.432	0.124	–	45	0.347	0.347	0.347	0.069	–

Power value for fully observed data $\mathfrak{D}$ is 0.626.
– denotes that ν_{Rt} was negative in simulations.

The similarity of observed Type I errors and power values imply that the proposed method based on Hodgson's transformation is inferentially equivalent to Rubin's traditional cumbersome multivariate methods, without the computational burden and reliance on unverifiable assumptions – existence of an inverse variance matrix, and proportional and equal loss of information. Similar findings were observed in other configurations with power values increasing to 1 as sample size increases (see Tables C.3, C.4, E.3, E.4, and Figures D.2 and F.2 in supplementary material).

5. Areas of application

The simulation results in Section 4 demonstrate the inferential equivalence between proposed combining rules using Hodgson's transformation and Rubin's traditional multivariate combining rules for multicomponent parameters of interest. With the proposed methodology, several types of statistical tests (where the data test statistic is F- or Beta-distributed random variable) now become feasible in the MI framework. Some examples include (multiparameter) partial F-tests, ANOVA (one-, two-, or *m*-way) for significance of factors/effects, test for equality of variances in normally distributed populations (via F-test, Levene's test, or Brown-Forsythe test), multiple/partial correlation coefficients or coefficients of determination, Scheffe's interval or test, and tests of hypothesis in restricted linear regression. This range of applicability with the proposed methodology should appeal to practitioners because the steps involved are simple and straightforward – (i) collect F-statistic values from analysis of each imputed dataset, (ii) transform them using Hodgson's transformation, and (iii) combine them via univariate combining rules.

The above range of applicability is not available under the traditional combining rules. To illustrate this, we consider the Landrigan et al. (1975) study. In this study, objectives include (i) assessing the strength of the association between neurological motor function and low or high levels of lead exposure in children (via R^2), and (ii) assessing the same strength of association after controlling for subject's age and sex (via multiparameter partial F-test). Current MI literature for R^2 lacks theoretical justification and is unreliable (Harel, 2009; Chaurasia and Harel, 2015; van Ginkel, 2019), while the multiparam-

eter partial F-tests is lacking in MI because the traditional MI-methods only allow one to determine whether a single effect is significant given all other effects are included in the model. Section 6 provides a real data setting with the Landrigan et al. (1975) study, where the proposed methodology helps address the aforementioned objectives of associations between neurological motor function and lead exposure in children.

6. Data application

Background

In December 1971, the City-County Health Department in El Paso, Texas, discovered that an ore smelter in El Paso was discharging large quantities of lead and other metallic wastes into the air between the years 1969 and 1971. In an effort to assess the association between lead exposure and neuro-psychological function, Landrigan et al. (1975) investigated a group of children that lived within 6.6 km of the aforementioned lead-emitting smelter in El Paso, Texas. The authors recorded various biomedical measures (such as blood-lead levels in μg per 100 ml. in 1972 and 1973), neurological measures (motor functions like finger-wrist tapping test), and psychological measures, along with other demographical measures (such as age, and sex). The children in this study where divided into two groups – *Lead-absorption* and *Control*. The *Lead-absorption* group of 46 children were those with blood-lead levels $\geq 40\mu$g/mL in 1972 and the control group consisted of 78 children who had blood-lead levels $< 40\mu$g/mL in 1972 and 1973.

For more details of the survey population and evaluations measures, we direct readers to the original work of Landrigan et al. (1975). For information on the study/variable summary and various study objectives, we direct readers to Rosner (2015). Lastly, the related data files can be obtained from biostat.mc.vanderbilt.edu/wiki/Main/DataSets.

Variables and Models of interest

The variables of interest in our application are similar to Landrigan et al. (1975) – (i) the number of finger-wrist taps in the dominant hand (a measure of neurological function; denoted as `maxfwt`), (ii) main predictor/exposure variables of blood-lead levels in 1972 and 1973 (denoted as `ld72` and `ld73`, respectively), and (iii) control variables `age` (in months), and (binary) `sex`.

The analysis models of interest include a (i) Simple Linear Regression (SLR) model to study the strength of association between `maxfwt` and `age` under parallelism assumption for the *Lead-absorption* and *Control* group, and (ii) Multiple Linear Regression (MLR) model to study the strength of association between `maxfwt` and blood-lead exposures `ld72` and `ld73`, after controlling for the confounders of `age` and `sex`. Both these models will be investigated under complete case analysis (CCA) and with imputed datasets after the original incomplete data has gone through multiple data imputation. We used a significance level of 5% in all analyses.

Missing values

Among 124 cases, the incomplete counts (percentages) for variables of interest are as follow: 25 cases (20.16%) with missing values in `maxfwt`, and 40 cases (32.26%) with missing values in Wechsler full-scale IQ score (`WWPS` which is another measure of neurological function). When using CCA for the SLR and MLR models of interest, the analyses discarded 25 incomplete cases (20.16%) due to missing values in `maxfwt`. Such a situation warrants the use of MI.

Imputation model and assumptions

Since the proposed method uses MI, it is assumed that data are MAR, and imputation models are fully and conditionally specified (for each variable with missing values), with imputation based on the semi-parametric approach of Predictive Mean Matching (PMM) for each of the variables with missing values. All variables in the original data were imputed, and variables other than `maxfwt`, `ld72`, `ld73`, `age` and `sex`, were used as auxiliary variables for PMM. The PMM method was chosen since it ensures that imputed values are restricted to the observed (and hence plausible) values, and can preserve non-linear relationships unlike the traditional regression imputation methods, which assume a joint multivariate normal distribution (Buuren and Groothuis-Oudshoorn, 2010; Horton and Lipsitz, 2001; Van-Buuren and Oudshoorn, 1999). For the number of imputations we consider $m = 1000$.

Analyses results: SLR

In this model, interest lies in assessing if the *Lead-absorption* and *Control* groups differed in finger-wrist tapping scores (`maxfwt`) when adjusting for age under parallelism. We assumed parallelism because it is a simpler model and Landrigan et al. (1975) found the slopes to be similar for the *Lead-absorption* and *Control* groups. For comparison, we fitted the SLR model under CCA (which discards 20.16% of the data due to missing `maxfwt`) and with imputed data from MI. In our MI analysis, we estimated the regression coefficients as an average of estimates from multiple imputed datasets, and used

Table 5
Simple Linear Regression (SLR) results comparing the association between `Age` and neurological function `maxfwt` for the *Lead-absorption* and *Control* groups from CCA and MI.

	CCA-SLR			MI-SLR		
Parameter	Estimate	SE	pvalue	Estimate	SE	pvalue
Absorption (Intercept)	22.043	3.492	8.48×10^{-9}	23.711	3.517	2.18×10^{-9}
(diff. to) Control	4.847	2.034	0.019	4.594	2.009	0.025
Age (in years)	0.221	0.027	1.02×10^{-12}	0.211	0.027	2.40×10^{-12}
	$r^2_{\text{Absorption}} = 31.42\%$			$r^2_{\text{Absorption}} = 45.00\%$		
	$r^2_{\text{Control}} = 47.60\%$			$r^2_{\text{Control}} = 60.23\%$		

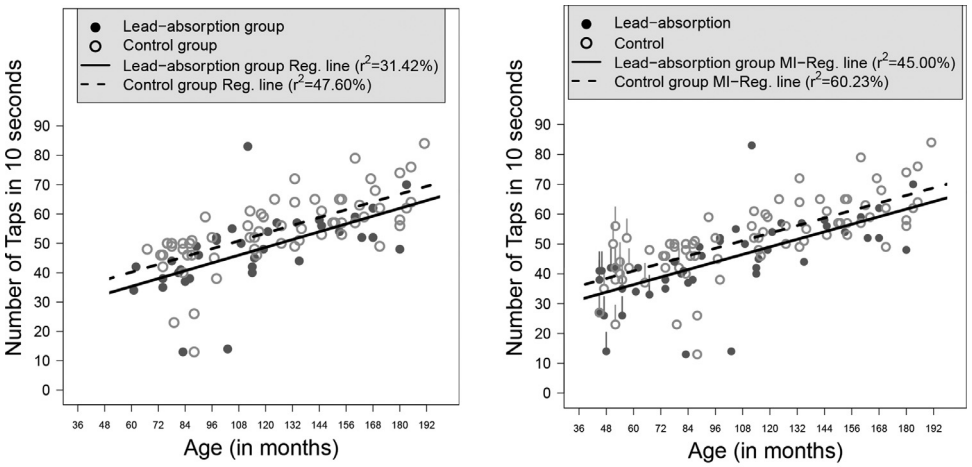

Fig. 5. Comparing SLR results of Finger-wrist tapping score (`maxfwt`) regressed on `age` from CCA as done in Landrigan et al. (1975, see regression figure on p.710) to our proposed method via Hodgson's transformation. The push-pin symbols in (b) denote one set of imputations in the respective groups.

our proposed method under Hodgson's transformation (with ν_{BR}) for the computation of coefficient of determination (R^2). Findings from this analysis, are summarized in Table 5 and Figure 5.

Results in Table 5 show that under CCA, the *Lead-absorption* group (on average) has significantly lower finger-wrist taping scores (by 4.847 units) after accounting for effect of `age`. Under MI, the *Lead-absorption* group again shows significantly lower finger-wrist taping scores (by 4.594 units). The coefficient of determination estimate from MI is higher in value by (approximately) 15% in both the *Lead-absorption* and *Control* groups, respectively. This suggest that CCA results are likely to underestimate the coefficient of determination and hence the linear ability of `age` in explaining the variation in neurological function (when measured via `maxfwt`).

Analyses results: MLR

In this model, interest lies in assessing the strength of association between finger-wrist tapping scores (`maxfwt`) and blood-lead levels (`ld72` and `ld73`) after controlling for the effects of `age` and `sex`. In other words, the hypothesis of interest is

$$H_0 : \boldsymbol{\beta}_{\texttt{ld72,ld73|age,sex}} = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_{\texttt{ld72,ld73|age,sex}} \neq \mathbf{0}.$$

For the above conditional test, we used the partial F-test in CCA, while in MI the partial F-tests from multiply-imputed data were combined using the proposed method under Hodgson's transformation (with ν_{BR}). Additionally, we estimated (under CCA and MI) the additional variation explained by `ld72` and `ld73` when the model already contains `age` and `sex`. CCA results yield a significant partial F-statistic ($F_{\texttt{ld72,ld73|age,sex}} = 4.106$ with 2 and 94 degrees of freedom; pvalue=0.019), and $R^2_{\texttt{ld72,ld73|age,sex}} = 8.03\%$. This CCA result suggests that blood-lead levels are associated with neurological function (when measured via `maxfwt`) such that they significantly explained additional 8.03% of the variation observed in the outcome `maxfwt` after accounting for the impact of `age` and `sex`. When assessing the same objective through MI under our proposed combining rule, the partial F-statistic was significant ($F_{\texttt{ld72,ld73|age,sex}} = 2.720$ with 2 and 56.998 degrees of freedom; pvalue=0.022), and $R^2_{\texttt{ld72,ld73|age,sex}} = 8.71\%$. This MI result (based on the proposed combining rules) suggests that blood-lead levels are associated with neurological function (when measured via `maxfwt`) such that they significantly explained additional 8.71% of the variation observed in the outcome `maxfwt` after accounting for the impact of `age` and `sex`. Though

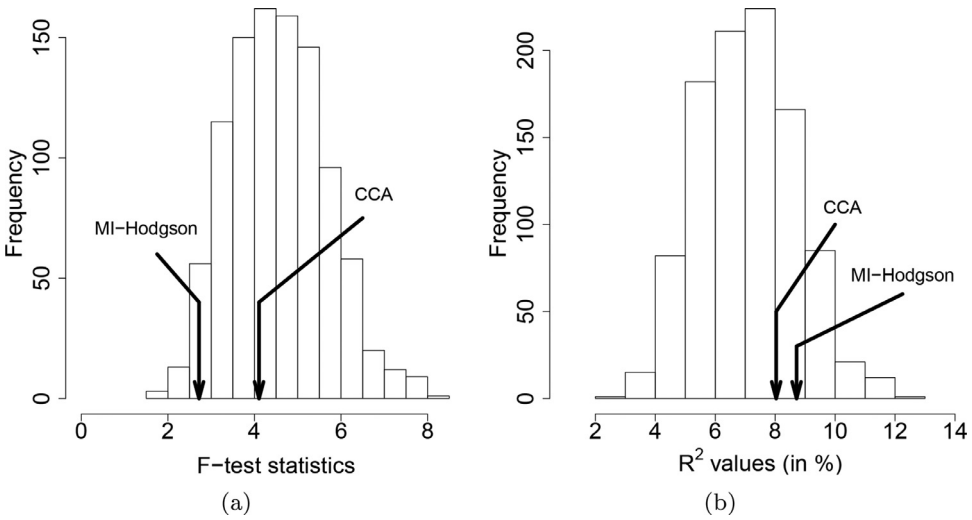

Fig. 6. Histograms describing the distribution of (a) MLR partial F-statistic values and (b) MLR partial R^2 values, from analysis of 1000 imputed datasets, and comparing the MI-combined partial F-test statistic and R^2 value to those from CCA.

the conclusion from CCA and MI are the same, it should be noted that the results from CCA are likely to be biased and underestimate the true strength of association of blood-lead levels with neurological function. These findings are summarized in Figure 6, which demonstrates the ability of MI to account for the variability of missing values in the statistics of interest under the newly proposed combining rules for F- and Beta- statistics.

7. Discussion

F-statistics play an important role in various statistical tests for data with or without missing values. In the case of incomplete data, multiple imputation is one way of handling missingness which accounts for the additional uncertainty due to missing values. However, computation for inference in traditional MI in Section 7 Discussion on page 14 methods get exponentially cumbersome for multicomponent parameters of interest, because they involve calculations with vectors and matrices. This article provides a univariate equivalent for the traditional multivariate MI methods without the computational cost (of calculation with matrices) and reliance on unverifiable assumptions such as *proportional and equal loss of information* (due to missingness across the components of the multivariate parameter of interest). Among the proposed methods, the combining rules for the (univariate) F- and *Beta*-statistics via Hodgson's transformation had nominal Type I error rates and power values similar to the traditional multivariate MI methods. Hence, our proposed new combining rule (via Hodgson's transformation) now serves as a viable tool to conduct various tests (see examples in Section 5) with imputed data. Given the proposed method's simplicity and reliable inferential properties, it has general applicability wherever the inference involves univariate F- and Beta-statistics, without any of the burdens and assumptions of the traditional multivariate combining rules of MI.

Acknowlgedgments

The author wishes to express his gratitude to the associate editor and two anonymous referees for their constructive feedback that helped improve the original version of this manuscript. This research work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2018-05479. All analyses and interpretations expressed herein are those of the author and do not necessarily reflect or represent the position of the author's institution or grant agency/council.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecosta.2021.08.013

References

Allison, P.D., 2002. Missing Data. Sage Publication Inc., Newbury Park, CA.
Aroian, L.A., 1941. A study of ra fisher's z distribution and the related f distribution. The Annals of Mathematical Statistics 12 (4), 429–448.
Barnard, J., Rubin, D.B., 1999. Small-sample degrees of freedom with multiple imputation. Biometrika 86, 948–955.
Buuren, S.v., Groothuis-Oudshoorn, K., 2010. mice: Multivariate imputation by chained equations in r. Journal of statistical software 1–68.

Chaurasia, A., Harel, O., 2015. Partial f-tests with multiply imputed data in the linear regression framework via coefficient of determination. Statistics in medicine 34 (3), 432–443.
Fisher, R.A., 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10 (4), 507–521.
van Ginkel, J.R., 2019. Significance tests and estimates for r2 for multiple regression in multiply imputed datasets: A cautionary note on earlier findings, and alternative solutions. Multivariate behavioral research 54 (4), 514–529.
Graham, J.W., Olchowski, A.E., Gilreath, T.D., 2007. How many imputations are really needed? some practical clarifications of multiple imputation theory. Prevention science 8 (3), 206–213.
Grund, S., Lüdtke, O., Robitzsch, A., 2016. Pooling anova results from multiply imputed datasets: A simulation study. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 12, 75–88.
Harel, O., 2009. The estimation of $\mathbf{R}^2$ and adjusted $\mathbf{R}^2$ in incomplete data sets using multiple imputation. Journal of Applied Statistics 36 (10), 1109–1118.
Hodgson, V., 1968. Some Normalizing Transformations of the Multiple Correlation Coefficient. Technical Report. Florida State University Tallahassee Department of Statistics.
Horton, N.J., Lipsitz, S.R., 2001. Multiple imputation in practice: comparison of software packages for regression models with missing variables. The American Statistician 55 (3), 244–254.
Landrigan, P.J., Baloh, R.W., Barthel, W.F., Whitworth, R.H., Staehling, N.W., Rosenblum, B.F., 1975. Neuropsychological dysfunction in children with chronic low-level lead absorption. The Lancet 305 (7909), 708–712.
Li, K.H., Raghunathan, T.E., Rubin, D.B., 1991. Large sample significance levels from multiply imputed data using moment-based statistics and an f reference distribution. Journal of the American Statistical Association 86, 1065–1073.
Meng, X.-L., Rubin, D.B., 1992. Performing likelihood ratio tests with multiply-imputed data sets. Biometrika 79, 103–111.
Raghunathan, T., Dong, Q., 2011. Analysis of variance from multiply imputed data sets. Ann Arbor: University of Michigan.
Reiter, J.P., 2007. Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. Biometrika 94, 502–508.
Rosner, B., 2015. Fundamentals of Biostatistics, 8 Nelson Education.
Rubin, D., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York.
Schafer, J., 1997. Analysis of Incomplete Multivariate Data. Chapman and Hall, London.
Severo, N.C., Zelen, M., 1960. Normal approximation to the chi-square and non-central f probability functions. Biometrika 47 (3/4), 411–416.
Van-Buuren, S., Oudshoorn, C., 1999. Flexible multivariate imputation by mice. TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054, Leiden.
Von Hippel, P.T., 2005. Teacher's corner: how many imputations are needed? a comment on hershberger and fisher (2003). Structural Equation Modeling 12 (2), 334–335.
Von Hippel, P.T., 2020. How many imputations do you need? a two-stage calculation using a quadratic rule. Sociological Methods & Research 49 (3), 699–718.
Wise, M., 1960. On normalizing the incomplete beta-function for fitting to dose-response curves. Biometrika 47 (1-2), 173–175.