A comparison of different measures of the proportion of explained variance in multiply imputed data sets

Ginkel, J.R. van; Karch, J.D.

Received: 14 September 2023 | Accepted: 21 March 2024

DOI: 10.1111/bmsp.12344

the british psychological society
promoting excellence in psychology

ARTICLE

A comparison of different measures of the proportion of explained variance in multiply imputed data sets

Joost R. van Ginkel | **Julian D. Karch**

Methodology and Statistics, Leiden University, Leiden, The Netherlands

Correspondence
Joost R. van Ginkel, Methodology and Statistics, Leiden University, PO Box 9500, 2300 RB Leiden, The Netherlands.
Email: jginkel@fsw.leidenuniv.nl

Abstract

The proportion of explained variance is an important statistic in multiple regression for determining how well the outcome variable is predicted by the predictors. Earlier research on 20 different estimators for the proportion of explained variance, including the exact Olkin–Pratt estimator and the Ezekiel estimator, showed that the exact Olkin–Pratt estimator produced unbiased estimates, and was recommended as a default estimator. In the current study, the same 20 estimators were studied in incomplete data, with missing data being treated using multiple imputation. In earlier research on the proportion of explained variance in multiply imputed data sets, an estimator called $\widehat{R}^2_{SP}$ was shown to be the preferred pooled estimator for regular R^2. For each of the 20 estimators in the current study, two pooled estimators were proposed: one where the estimator was the average across imputed data sets, and one where $\widehat{R}^2_{SP}$ was used as input for the calculation of the specific estimator. Simulations showed that estimates based on $\widehat{R}^2_{SP}$ performed best regarding bias and accuracy, and that the Ezekiel estimator was generally the least biased. However, none of the estimators were unbiased at all times, including the exact Olkin–Pratt estimator based on $\widehat{R}^2_{SP}$.

KEYWORDS

missing data, Monte Carlo simulation, multiple imputation, multiple linear regression, proportion of explained variance

1 | INTRODUCTION

An important part of interpreting the results of multiple regression analysis is looking at how well the outcome variable Y is predicted by the predictor variables $X_1, \ldots, X_p$. In a complete population, a measure that expresses this is the proportion of explained variance, denoted ρ^2. In the population, ρ^2 may be computed in the following way. Suppose that σ_Y^2 is the total variance of Y and that σ_ϵ^2 is the error variance of the regression model. The formula for ρ^2 is

$$\rho^2 = 1 - \frac{\sigma_\epsilon^2}{\sigma_Y^2}. \tag{1}$$

Various sample estimators for the proportion of explained variance exist. The most widely known and most widely used one is the sample proportion explained variance, denoted R^2. Because R^2 is an intrinsically biased estimator of ρ^2, several alternative estimators have been proposed in the statistical literature. Of these estimators, adjusted R^2, also known as the Ezekiel estimator, is the best known. Two less well-known variants of the Ezekiel estimator have been proposed, namely the Smith and Wherry estimators (Raju et al., 1997; Yin & Fan, 2001).

Besides R^2 and the different variants of adjusted R^2, various other estimators have been proposed to reduce or even completely remove the intrinsic bias of R^2, namely the maximum likelihood estimator (Alf & Graf, 2002), and different variants of the Olkin–Pratt estimator (Karch, 2020; Olkin & Pratt, 1958; Raju et al., 1997; Yin & Fan, 2001). A more extensive discussion of all of the above-mentioned estimators will follow in Section 2.

1.1 | Proportion of explained variance and missing data

When data sets are incomplete because not all respondents responded to all the variables, missing data need to be handled prior to carrying out a multiple regression analysis. The default way to deal with missing data in many statistical software packages is the removal of all cases with at least one missing value from the analysis. This method is called listwise deletion. Although listwise deletion is easy to apply, it comes with two major problems. Firstly, useful information is thrown away, resulting in a loss of power. Secondly, in order for listwise deletion to lead to unbiased results, the missing data on specific variables should not depend on other observed variables, on unobserved variables, or on the (hypothetical) values of the missing data themselves. This assumption is called the missing completely at random assumption (MCAR; Little & Rubin, 2002, p. 10). This assumption is, however, often not realistic.

Two methods that resolve the above-mentioned problems of listwise deletion are full information maximum likelihood (FIML) and multiple imputation (MI). In FIML, the likelihood function is calculated for all the observed values in the data, without deleting any cases. Additionally, in FIML the MCAR assumption may be loosened in order for results still to be unbiased. Missing at random (MAR; Little & Rubin, 2002, p. 10) is a missingness mechanism where the probability of data being missing depends on observed variables but not on unobserved variables, or on the unobserved values of the missing data themselves. Under MAR, FIML will give unbiased results.

In MI, on the other hand (Rubin, 1987; Van Buuren, 2012; Van Buuren et al., 2006), the missing data are estimated M times using a statistical model that is suitable for the data in question. This results in M complete versions of the original incomplete data set. Next, the M data sets are analysed using the statistical analysis of interest. Finally, the M analyses or statistics are combined into one overall result. Like FIML, MI will give unbiased results under MAR as well.

As a side note, when the probability of data being missing depends on unobserved variables or the hypothetical values of the missing data themselves, missing data are said to be missing not at random

(MNAR; Little & Rubin, 2002, p. 10). Under MNAR, none of the above-mentioned methods for dealing with missing data are guaranteed to give unbiased results. Methods for dealing with MNAR are beyond the scope of the current paper. The interested reader is referred to Fay (1986), Galimard et al. (2016), Heckman (1976), and Moustaki and Knott (2000).

In practice, FIML and MI will often give similar results (e.g., Collins et al., 2001). Consequently, reasons to prefer one method over the other are mostly practical. If a researcher wants to apply only statistical methods that rely on maximum likelihood estimation, FIML is preferred over MI for two main reasons. Firstly, it is less work than MI, since FIML is the default method in many statistical software packages that rely on maximum likelihood estimation. Secondly, when FIML is applied to the same data set repeatedly, results will always be identical, whereas repeatedly applying MI to the same data set will result in different imputed values each time, and consequently, slightly different results of statistical analyses. If, on the other hand, some of the intended analyses can be applied using FIML while others cannot, then MI may be preferred over FIML for all analyses, to keep results among different analyses comparable.

In the current study the context is a situation where the researcher prefers MI over FIML, and wants to obtain an estimate of ρ^2 in multiply imputed data. Additionally, it should be noted that although it is technically possible to apply FIML to regression analysis using structural equation modelling, this is, in our experience, not frequently done by applied researchers.

1.2 | Combination strategies for R^2 in multiply imputed data

For many statistical analyses specific formulas have been developed for combining the M results of the M multiply imputed data sets, which we will refer to as *combination strategies*. For R^2, several combination strategies have been discussed in the literature. For example, Harel (2009) proposed a procedure where, first, a Fisher z transformation of the $\sqrt{R^2_m}$ of each imputed data set m is calculated; next, the average of the M Fisher z transformations is taken; and finally, the average is transformed back to its original scale. The resulting pooled estimator is denoted $\mathfrak{R}^2$. Also, Van Ginkel (2019) proposed averaging the R^2_ms across the M imputed data sets directly, denoted $\overline{R^2}$.

Van Ginkel (2020) argued that both $\mathfrak{R}^2$ and $\overline{R^2}$ have some theoretical flaws. Firstly, the Fisher z transformation is only suitable for correlations that vary between -1 and $+1$, while the square root of R^2 can only vary between 0 and 1. If one would like to use a transformation for the square root of R^2, it would be better to use a logit transformation instead.

However, performing a logit transformation will not resolve another problem that both $\mathfrak{R}^2$ and $\overline{R^2}$ have. Van Ginkel (2020) argued that a predictor X_j contributing almost nothing to ρ^2 in the population will have a disproportionally large contribution to $\mathfrak{R}^2$ and $\overline{R^2}$ in the sample. This is because in some imputed data sets the regression coefficient of X_j may be slightly negative while in others it may be slightly positive, due to sampling fluctuation. In each imputed data set, however, the contribution of X_j to R^2_m will be slightly positive because the sign of the regression coefficient does not matter for the total contribution of X_j to R^2_m. Consequently, both negative and positive regression coefficients of X_j across imputed data sets will have a slightly positive contribution to $\mathfrak{R}^2$ and $\overline{R^2}$, while these contributions should actually almost perfectly cancel each other out in a final pooled estimator for ρ^2. There is no transformation that can make up for this disproportionate contribution to an estimator for ρ^2.

To overcome the above problems, Van Ginkel (2020) proposed two alternative combination strategies for R^2 in multiply imputed data. He denoted the resulting estimators $\widehat{R}^2_{\text{PS}}$ and $\widehat{R}^2_{\text{SP}}$. In a simulation study, Van Ginkel (2020) showed that $\widehat{R}^2_{\text{PS}}$ and $\widehat{R}^2_{\text{SP}}$ were closer to the R^2 that would be obtained if no data were missing than $\mathfrak{R}^2$ and $\overline{R^2}$. In the remainder of this paper, $\mathfrak{R}^2$ is not further considered because of its theoretical flaws, but the suboptimal $\overline{R^2}$ is still used for reasons explained later.

In Van Ginkel's (2020) study it was found that $\widehat{R}^2_{PS}$ and $\widehat{R}^2_{SP}$ hardly differed in terms of bias and efficiency. However, he argued that $\widehat{R}^2_{SP}$ had a better theoretical justification, and showed that relations between R^2 and other statistics that are present when data are complete, are mostly maintained for $\widehat{R}^2_{SP}$ and not for $\widehat{R}^2_{PS}$. For this reason, he recommended using $\widehat{R}^2_{SP}$ as a pooled version of R^2 in multiply imputed data sets. For the current paper, we will follow this recommendation and continue with $\widehat{R}^2_{SP}$.

1.3 | Aim of the current study

Although combination strategies for R^2 in multiply imputed data have been proposed (Van Ginkel, 2020), the same is not the case for the alternative estimators of ρ^2 mentioned so far (e.g., the Ezekiel estimator variants and the Olkin–Pratt estimator variants), at least not to our knowledge. Consequently, researchers who prefer using these estimators in a multiply imputed data set currently have no way to do that. For these researchers it is important that pooled versions of these estimators in multiply imputed data become available as well.

Furthermore, once available, it is important to study their statistical properties to see whether the pooled estimators behave similarly to their complete-data counterparts. Thus, the goal of this study is twofold: to propose combination strategies for all of the above-mentioned alternative estimators for ρ^2; and to study the statistical properties of these estimators. Karch (2020) studied the bias and mean squared error (MSE) in all of the estimators for ρ^2 in complete data discussed so far. The current simulation study follows up on the study by Karch (2020) by studying all the estimators for ρ^2 from his study, but now in incomplete data where the missing data have been treated using MI.

In Section 2, the formulas for the different estimators of ρ^2 in complete data are given. Then pooled versions of the estimators in multiply imputed data are presented, along with their reasoning behind them. After discussing the estimators and pooled estimators, the setup of the simulation study is outlined. In Section 3, the results of the study are discussed. Finally, in Section 4, a conclusion will be drawn on whether properties of the pooled estimators of ρ^2 are preserved in multiply imputed data, and a general recommendation on whether to use averaging across imputed data sets or using $\widehat{R}^2_{SP}$ as input for the estimation of the specific estimator is given.

2 | METHOD

2.1 | Estimators of the proportion of explained variance in multiple regression

The most widely used estimator for ρ^2, R^2, may be computed in several ways, all of which are equivalent. One way that is in accordance with how ρ^2 is computed in Equation (1) is to replace σ^2_Y and σ^2_ϵ with their (biased) maximum likelihood (ML) estimates in Equation (1). Using the total sum of squares of variable Y, denoted SS_T, and the total sum of squares of the error, SS_e, the ML estimates of σ^2_Y and σ^2_ϵ are $\widehat{\sigma}^2_Y = SS_T/N$ and $\widehat{\sigma}^2_\epsilon = SS_e/N$, respectively. Substituting these estimators into Equation (1) gives

$$R^2 = 1 - \frac{SS_e}{SS_T}. \tag{2}$$

One property of R^2 is that it is a positively biased estimator of ρ^2 (Fisher, 1928). The Ezekiel estimator, better known as adjusted R^2, attempts to reduce this bias by using the unbiased estimators of σ^2_Y and σ^2_ϵ, namely $S^2_Y = SS_T/(N-1)$ and $S^2_e = SS_e/\left(N-p-1\right)$, respectively. The resulting estimator is

$$R^2_{\mathrm{E}} = 1 - \frac{N-1}{N-p-1}\left(1-R^2\right). \tag{3}$$

Two variants of R^2_{E} have been proposed, which are all aimed at reducing the intrinsic bias that R^2 has. One of these is the Smith estimator, which is computed as

$$R^2_{\mathrm{S}} = 1 - \frac{N}{N-p}\left(1-R^2\right). \tag{4}$$

The other is the Wherry estimator, computed as

$$R^2_{\mathrm{W}} = 1 - \frac{N-1}{N-p}\left(1-R^2\right). \tag{5}$$

For a more extensive explanation of both estimators, we refer to Raju et al. (1997) and Yin and Fan (2001).

A disadvantage of the above corrections is that they can sometimes lead to negative values, which is conceptually nonsensical (a negative proportion of explained variance is not possible). Alf and Graf (2002) suggested using the maximum likelihood estimator of ρ^2 under the assumption that both the predictors and the outcome variable are multivariate normally distributed. We denote this estimator R^2_{ML}. An advantage of R^2_{ML} is that it is on average closer to ρ^2 than R^2, and does not give negative values. R^2_{ML} is not easily expressed in closed form, so for technical details we only refer to Alf and Graf (2002).

Although all of the above estimators attempt to reduce bias in R^2 in different ways, none of them are exactly unbiased. Olkin and Pratt (1958) introduced an estimator of ρ^2 that is unbiased under the assumption of multivariate normality of the predictors and the outcome variable, called the Olkin–Pratt estimator. This estimator uses the hypergeometric function ${}^2F_1 = \left(a, b; c; z\right)$. The Olkin–Pratt estimator is given by

$$R^2_{\mathrm{OP}} = 1 - \frac{N-3}{N-p-1}\left(1-R^2\right){}^2F_1\left(1, 1; \frac{N-p+1}{2}; 1-R^2\right) \tag{6}$$

In the past it was considered difficult to implement R^2_{OP} in statistical software packages because of the inclusion of the hypergeometric function (Shieh, 2008). However, Karch (2020) showed that there were closed-form solutions to the hypergeometric function for all possible inputs required for the Olkin–Pratt estimator. Using this closed-form solution, the Olkin–Pratt estimator can easily be computed.

Since it was once considered difficult to compute the Olkin–Pratt estimator, some easier-to-compute approximations were introduced. One family of approximations uses only the first $K+1$ addends of the infinite series of the hypergeometric function. Suppose t_k is the kth addend of the hypergeometric series. This family of approximations can be expressed as

$$R^2_{\mathrm{OP}K} = 1 - \frac{N-3}{N-p-1}\left(1-R^2\right)\sum_{k=0}^{K} t_k. \tag{7}$$

Besides the above family of approximations, there are two other approximations, which are discussed by Raju et al. (1997) and Yin and Fan (2001). These variants use the R^2_{OP1} approximation as a starting point and correct for the omitted addends. The first variant is one by Pratt:

$$R^2_{\mathrm{P}} = 1 - \frac{N-3}{N-p-1}\left(1-R^2\right)\left(1 + \frac{2\left(1-R^2\right)}{N-p-2.3}\right). \tag{8}$$

The second variant is one by Claudy:

$$R^2_C = 1 - \frac{N-4}{N-p-1}\left(1-R^2\right)\left(1+\frac{2\left(1-R^2\right)}{N-p+1}\right) \tag{9}$$

(both formulas were first given in Claudy, 1978).

Like adjusted R^2, all of the above variants of the Olkin–Pratt estimator can be less than 0. To make up for negative values, the estimators can be modified by setting their values to 0 when the initial value drops below 0. Shieh (2008) called these estimators positive-part estimators.

2.2 | Combination strategies for alternative estimators of ρ^2 in multiply imputed data

In this section, we first explain the two proposed combination strategies by Van Ginkel (2020) that recovered R^2 best. Van Ginkel's proposed combination strategies make use of the fact that in complete data R^2 may also be written as

$$R^2 = \sum_{j=1}^{p} r_{X_jY}\widehat{\beta}_j, \tag{10}$$

where r_{X_jY} is the sample correlation between predictor X_j and Y, and $\widehat{\beta}_j$ is the standardised regression coefficient of predictor X_j.

Van Ginkel proposed two pooled versions of r_{X_jY}, and two pooled versions of $\widehat{\beta}_j$ which can be substituted into Equation (10) to get two pooled versions of R^2 in multiply imputed data. The first way to compute a pooled version of r_{X_jY} and of $\widehat{\beta}_j$ is as follows. Let $\widehat{b}_{j,m}$ be the sample estimate of the unstandardised regression coefficient of $X_{j,m}$ and $\widehat{\beta}_{j,m}$ be the corresponding standardised coefficient, let $r_{X_jY,m}$ be the sample correlation between $X_{j,m}$ and Y_m, let $s^2_{X_j,m}$ be the sample variance of $X_{j,m}$, and let $s^2_{Y,m}$ be the variance of Y_m, in imputed data set m ($m = 1,\ldots, M$).

By first pooling the regression coefficient of X_j and the standard deviations of X_j and Y, and using these pooled quantities for calculating a pooled standardised regression coefficient, we get:

$$\begin{aligned}
\overline{\widehat{b}}_j &= \frac{1}{M}\sum_{m=1}^{M}\widehat{b}_{j,m},\\
\widetilde{s}_{X_j} &= \sqrt{\frac{1}{M}\sum_{m=1}^{M}s^2_{X_j,m}},\\
\widetilde{s}_{Y} &= \sqrt{\frac{1}{M}\sum_{m=1}^{M}s^2_{Y,m}},\\
\overline{\widehat{\beta}}_{j,\mathrm{PS}} &= \frac{\widetilde{s}_{X_j}}{\widetilde{s}_{Y}}\overline{\widehat{b}}_j.
\end{aligned} \tag{11}$$

Here, the PS in the subscript stands for *pooling before standardisation* because the pooling of the necessary quantities takes place first, and then the standardisation.

Next, the standardised regression coefficient that is obtained when X_j is regressed on Y in a simple regression is used as a measure for the pooled correlation, and is denoted by $r_{X_jY,\mathrm{PS}}$. Using the $\overline{\widehat{\beta}}_{j,\mathrm{PS}}$ and $r_{X_jY,\mathrm{PS}}$, the first pooled version of R^2 proposed by Van Ginkel (2020) is

$$\widehat{R}^2_{\text{PS}} = \sum_{j=1}^{p} r_{X_j Y,\text{PS}} \overline{\widehat{\beta}}_{j,\text{PS}}. \tag{12}$$

The second way to compute a pooled version of $r_{X_j Y}$ and of $\widehat{\beta}_j$ is by averaging both quantities across the M imputed data sets:

$$\overline{\widehat{\beta}}_{j,\text{SP}} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\beta}_{j,m}, \tag{13}$$

$$\overline{r}_{X_j Y} = \frac{1}{M} \sum_{m=1}^{M} r_{X_j Y,m}. \tag{14}$$

The subscript SP in $\overline{\widehat{\beta}}_{j,\text{SP}}$ stands for *standardisation before pooling* because the regression coefficients are standardised per imputed data set first, and next the pooling takes place by averaging these standardised regression coefficients across the imputed data sets. Using the $\widehat{\beta}_{j,\text{PS}}$ and $r_{X_j Y,\text{PS}}$, we can obtain the second pooled version of R^2 by Van Ginkel (2020):

$$\widehat{R}^2_{\text{SP}} = \sum_{j=1}^{k} \overline{r}_{X_j Y} \overline{\widehat{\beta}}_{j,\text{SP}}. \tag{15}$$

For the other estimators for ρ^2 we propose two general combination strategies. The first strategy to pool the various estimators for ρ^2 is to replace R^2 in Equations (3–9) with $\widehat{R}^2_{\text{SP}}$, leading to the following pooled estimators for ρ^2:

$$\widehat{R}^2_{\text{E,SP}} = 1 - \frac{N-1}{N-p-1}\left(1 - \widehat{R}^2_{\text{SP}}\right), \tag{16}$$

$$\widehat{R}^2_{\text{S,SP}} = 1 - \frac{N}{N-p}\left(1 - \widehat{R}^2_{\text{SP}}\right), \tag{17}$$

$$\widehat{R}^2_{\text{W,SP}} = 1 - \frac{N-1}{N-p}\left(1 - \widehat{R}^2_{\text{SP}}\right), \tag{18}$$

$$\widehat{R}^2_{\text{OP,SP}} = 1 - \frac{N-3}{N-p-1}\left(1 - \widehat{R}^2_{\text{SP}}\right)^2 {}_2F_1\left(1, 1; \frac{N-p+1}{2}; 1 - \widehat{R}^2_{\text{SP}}\right), \tag{19}$$

$$\widehat{R}^2_{\text{OP}K,\text{SP}} = 1 - \frac{N-3}{N-p-1}\left(1 - \widehat{R}^2_{\text{SP}}\right) \sum_{k=0}^{K} t_{k,\text{SP}} \tag{20}$$

(where $t_{k,\text{SP}}$ is the kth addend of the hypergeometric series with $\widehat{R}^2_{\text{SP}}$ as input rather than R^2),

$$\widehat{R}^2_{\text{P,SP}} = 1 - \frac{N-3}{N-p-1}\left(1 - \widehat{R}^2_{\text{SP}}\right)\left(1 + \frac{2\left(1 - \widehat{R}^2_{\text{SP}}\right)}{N-p-2.3}\right), \tag{21}$$

$$\widehat{R}^2_{C,SP} = 1 - \frac{N-4}{N-p-1}\left(1-\widehat{R}^2_{SP}\right)\left(1+\frac{2\left(1-\widehat{R}^2_{SP}\right)}{N-p+1}\right). \tag{22}$$

The ML estimator based on $\widehat{R}^2_{SP}$ cannot easily be expressed in a formula, but is referred to as $\widehat{R}^2_{ML,SP}$. The above pooled estimators that may take on values below 0 can be made positive-part estimators by setting them to 0 whenever they are negative.

The second combination strategy is to calculate the specific estimator for each imputed data set separately, and average its M values across the M imputed data sets. This strategy is based on Van Ginkel's (2019) solution, which simply averages the R^2_m across the imputed data sets. Using averaging as the basis for a combination strategy, this leads to the following pooled versions of the above-discussed estimators for ρ^2 in multiply imputed data:

$$\overline{R^2_E} = 1 - \frac{N-1}{M(N-p-1)}\sum_{m=1}^{M}\left(1-R^2_m\right), \tag{23}$$

$$\overline{R^2_S} = 1 - \frac{N}{M(N-p)}\sum_{m=1}^{M}\left(1-R^2_m\right), \tag{24}$$

$$\overline{R^2_W} = 1 - \frac{N-1}{M(N-p)}\sum_{m=1}^{M}\left(1-R^2_m\right), \tag{25}$$

$$\overline{R^2_{ML}} = \frac{1}{M}\sum_{m=1}^{M} R^2_{ML,m}, \tag{26}$$

$$\overline{R^2_{OP}} = 1 - \frac{N-3}{M(N-p-1)}\sum_{m=1}^{M}\left(1-R^2_m\right)^2 F_1\left(1,1;\frac{N-p+1}{2};1-R^2_m\right), \tag{27}$$

$$\overline{R^2_{OPK}} = 1 - \frac{N-3}{M(N-p-1)}\sum_{m=1}^{M}\left(1-R^2_m\right)\sum_{k=0}^{K} t_{k,m}, \tag{28}$$

$$\overline{R^2_P} = 1 - \frac{N-3}{M(N-p-1)}\sum_{m=1}^{M}\left(1-R^2_m\right)\left(1+\frac{2\left(1-R^2_m\right)}{N-p-2.3}\right), \tag{29}$$

$$\overline{R^2_C} = 1 - \frac{N-4}{M(N-p-1)}\sum_{m=1}^{M}\left(1-R^2_m\right)\left(1+\frac{2\left(1-R^2_m\right)}{N-p+1}\right). \tag{30}$$

Like the pooled estimators based on $\widehat{R}^2_{SP}$, the estimators based on averaging that may take on values below 0 can be made positive-part estimators by setting them to 0 whenever they are negative.

2.3 | Critical discussion of both combination strategies

The combination strategies for the estimators of ρ^2 using $\widehat{R}^2_{SP}$ as input (Equations 16–22, and $\widehat{R}^2_{ML,SP}$) lack a solid statistical theory. However, since $\widehat{R}^2_{SP}$ has been shown to be an estimator that comes very close to the R^2 that would be obtained without missing data, it seems reasonable to assume that using $\widehat{R}^2_{SP}$ as an input will lead to better estimates of ρ^2 than simply averaging (Equations 23–30) will give.

The estimators based on averaging (Equations 23–30) are not based on a solid statistical theory either, other than that in Rubin's (1987) general MI framework an overall point estimate of an estimator is also obtained by averaging the M estimates across imputed data sets. However, some of the estimators (the different variants of the Olkin–Pratt estimator) are assumed to be unbiased in complete data under specific assumptions. Now suppose that the imputation process does not introduce any bias to an unbiased estimator in imputed data set m ($m = 1,\ldots, M$). Then the average of the M unbiased estimators will also be unbiased. The same is not necessarily true for the estimators of Equations (19) and (20), as the estimator $\widehat{R}^2_{SP}$ is a complex function of several pooled quantities from the regression analysis.

However, it has not been said that the imputation process will not introduce any bias to an unbiased estimator of ρ^2 in imputed data set m. Rubin's idea of averaging parameter estimates across imputed data sets relied on the assumption that in complete data the estimator in question has a normal sampling distribution and a confidence interval based on a z- or t-distribution. The pooled estimators in Equations (23–30) do not fit into that framework, so the question is to what extent they preserve the statistical properties of their complete-data counterparts. It is, for example, not clear whether the Olkin–Pratt estimator will still be unbiased when it is computed for each of the M imputed data sets and the average is used as an overall estimator.

Additionally, as already argued, predictors with regression coefficients close to 0 make a disproportionally large contribution to $\overline{R^2}$. Similar problems may occur for the estimators of Equations (23–30). This is why the estimator $\widehat{R}^2_{SP}$ was introduced by Van Ginkel (2020) in the first place – to overcome this problem.

In short, because both general pooling strategies lack a solid statistical theory, it is difficult to predict which of the two methods will produce the least bias in the estimators of ρ^2. A simulation study will have to show which of these two strategies is the preferred one. In the next subsection, the design of this simulation study will be discussed.

2.4 | Design of the simulation study

2.4.1 | Fixed design characteristics

The general model for simulating data was the following regression model:

$$Y = b_0 + \sum_{j=1}^{p} b_j X_j + \varepsilon \tag{31}$$

with

$$X \sim MVN\left(\mathbf{0}_p \boldsymbol{\Sigma}_p\right),$$

$$\boldsymbol{\Sigma}_p = \begin{bmatrix} 1 & & & \\ .3 & \ddots & & \\ \vdots & \ddots & \ddots & \\ .3 & \cdots & .3 & 1 \end{bmatrix},$$

$$\varepsilon \sim N(0, 10),$$

$$b_0 = 100,$$

$$b_j = \frac{10\left(\frac{1}{1-\rho^2} - 1\right)}{\sqrt{p + .3p(p-1)}}.$$

Using this model, data were simulated under various values of ρ^2, numbers of predictors p, and sample sizes N (all to be discussed later on). For each specific combination of ρ^2, p, and N, $D = 1000$ replicated data sets were drawn. We did not vary the covariance matrix of the predictors since this does not influence the distribution of R^2 and thus the performance of the estimators. From the complete data sets, incomplete data were created by removing various percentages of data points from the data, using two different missingness mechanisms (both to be discussed in Section 2.4.2).

Finally, once incomplete data sets were created they were multiply imputed. MI was done using the using the *mice* package (Van Buuren & Groothuis-Oudshoorn, 2011) in R (R Core Team, 2021). The method of imputation was fully conditional specification (FCS; Van Buuren, 2012, pp. 108–116; Van Buuren et al., 2006) where for each variable the imputation model was a normal linear regression model with the other variables as predictors. In the *mice* package this option can be found in the `mice.impute.norm` function. Twenty iterations were used for the FCS algorithm, and the number of imputations was $M = 25$.

2.4.2 | Independent variables

Sample size

The sample sizes studied were $N = 50$, 100, 250, 500. These values cover typical samples sizes as obtained in psychology (Marszalek et al., 2011). We did not investigate very low sample sizes as this occasionally led to situations where some predictors were removed from the imputation model because of logged events (Van Buuren, 2012, p. 130).

Population proportion of explained variance

Values of ρ^2 that were studied were $\rho^2 = 0$, .2, .5, .75, and .90. Although Shieh (2008) studied more values of ρ^2 than are studied here, the range of values for ρ^2 is the same as in his study.

Number of predictors

The number of predictors was varied to be $p = 2$, 4, and 6. Although Karch (2020) also studied a situation of 10 predictors, this was infeasible for higher percentages of missingness, as this resulted in logged events in small sample sizes.

Missingness mechanism

Incomplete data were created according to two missingness mechanisms, namely MCAR and MAR. Here we focus on the MCAR results. The MAR results are only briefly discussed in the Appendix, where we also describe our procedure for generating MAR data.

The reason for focusing on MCAR in the main text is that MAR generally does not introduce any (additional) bias to statistics when missing data are handled using MI (Schafer, 1997, pp. 23–26). Consequently, it is largely irrelevant which of the two mechanisms, MCAR or MAR, is used for simulating missing data when bias is the outcome of interest. There are exceptions to this. For example, Seaman et al. (2012) showed that a particular MI procedure produced unbiased results only under MCAR. However, since we do not have any reason to presume a priori that MAR will introduce additional bias to the estimators in our study, we will not discuss MAR in the main text, in order not to make the discussion of the results any more extensive than it already is.

The reason for not considering MNAR is that there are an infinite number of ways in which data can be MNAR, so if MNAR is simulated in one particular way, that does not say anything about how the results will generalize to other MNAR mechanisms. Thus, to get some impression of the effect of MNAR, it would make sense to include at least a few different MNAR mechanisms. However, this would make the study design substantially larger, while also allowing it to drift too far from the goal of the study.

Percentage of missingness

Percentages of missing data of 6.25%, 12.5%, and 25% were studied, based on Van Ginkel (2019, 2020).

Estimator of ρ^2

A total of 20 estimators of ρ^2 were studied, namely R^2, R^2_{E}, R^2_{S}, R^2_{W}, R^2_{ML}, R^2_{OP}, R^2_{OP1}, R^2_{OP2}, R^2_{OP5}, R^2_{P}, R^2_{C}, $R^2_{\mathrm{E+}}$, $R^2_{\mathrm{S+}}$, $R^2_{\mathrm{W+}}$, $R^2_{\mathrm{OP+}}$, $R^2_{\mathrm{OP1+}}$, $R^2_{\mathrm{OP2+}}$, $R^2_{\mathrm{OP5+}}$, $R^2_{\mathrm{P+}}$, and $R^2_{\mathrm{C+}}$, where the "+" sign stands for the positive-part variant. The choice of $K = 1, 2, 5$ in $R^2_{\mathrm{OP}K}$ was based on Karch (2020).

Combination strategy for ρ^2

Two combination strategies for estimators of ρ^2 were studied: averaging across M imputed data sets, and using $\widehat{\mathrm{R}}^2_{\mathrm{SP}}$ in the formula for the specific estimator. The resulting pooled statistics based on inserting $\widehat{\mathrm{R}}^2_{\mathrm{SP}}$, are $\widehat{\mathrm{R}}^2_{\mathrm{SP}}$ itself, $\widehat{\mathrm{R}}^2_{\mathrm{E,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{S,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{W,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{ML,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{OP,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{OP1,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{OP2,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{OP5,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{P,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{C,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{E+,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{S+,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{W+,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{OP+,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{OP1+,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{OP2+,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{OP5+,SP}}$, $\widehat{\mathrm{R}}^2_{\mathrm{P+,SP}}$, and $\widehat{\mathrm{R}}^2_{\mathrm{C+,SP}}$; the pooled statistics based on averaging are $\overline{R^2}$, $\overline{R^2_{\mathrm{E}}}$, $\overline{R^2_{\mathrm{S}}}$, $\overline{R^2_{\mathrm{W}}}$, $\overline{R^2_{\mathrm{ML}}}$, $\overline{R^2_{\mathrm{OP}}}$, $\overline{R^2_{\mathrm{OP1}}}$, $\overline{R^2_{\mathrm{OP2}}}$, $\overline{R^2_{\mathrm{OP5}}}$, $\overline{R^2_{\mathrm{P}}}$, $\overline{R^2_{\mathrm{C}}}$, $\overline{R^2_{\mathrm{E+}}}$, $\overline{R^2_{\mathrm{S+}}}$, $\overline{R^2_{\mathrm{W+}}}$, $\overline{R^2_{\mathrm{OP+}}}$, $\overline{R^2_{\mathrm{OP1+}}}$, $\overline{R^2_{\mathrm{OP2+}}}$, $\overline{R^2_{\mathrm{OP5+}}}$, $\overline{R^2_{\mathrm{P+}}}$, and $\overline{R^2_{\mathrm{C+}}}$. Note that $\overline{R^2}$ is not recommended on the basis of the results by Van Ginkel (2020). However, including $\overline{R^2}$ makes it possible to study a possible interaction between an estimator of ρ^2 and a combination strategy for ρ^2.

The above factors resulted in a design with 4 (sample size) × 5 (population proportion of explained variance) × 3 (number of predictors) × 3 (percentage of missingness) × 20 (estimator) × 2 (combination strategy) = 7200 cells. Within each combination of sample size, population proportion of explained variance, and number of predictors, different seed values were used for the generation of the replicated data sets. Consequently, these factors were independent factors. Within each of these replicated data sets, different percentages of missing data were simulated, and the two variants of each of the 20 estimators were calculated. Consequently, percentage of missingness, estimator and combination strategy were dependent factors.

2.4.3 | Dependent variables

Difference between the estimator of ρ^2 *and* ρ^2

Suppose $\widehat{\rho}^2_d$ is an estimate of ρ^2 of replicated data set d ($d = 1, \ldots, D$) in a specific design cell. One dependent variable was the difference between $\widehat{\rho}^2_d$ and ρ^2. This difference averaged across D replications defines the estimated bias in ρ^2 for the specific design cell:

$$\widehat{B}(\rho^2) = \frac{1}{D}\sum_{d=1}^{D}\left(\widehat{\rho}^2_d - \rho^2\right). \tag{32}$$

Squared difference between the estimator of ρ^2 *and* ρ^2

Although bias quantifies how close on average an estimator is to ρ^2, it does not tell us how accurately ρ^2 is estimated across replications within a specific design cell. To get an impression of the accuracy of the estimators, another dependent variable was studied, namely the squared difference between $\widehat{\rho}^2_d$ and ρ^2. When this squared difference is averaged across D replications, this defines the estimated mean squared error of ρ^2 for a specific design cell:

$$\widehat{\mathrm{MSE}}\left(\rho^2\right) = \frac{1}{D}\sum_{d=1}^{D}\left(\widehat{\rho}_d^2 - \rho^2\right)^2. \tag{33}$$

2.5 | Statistical analyses

2.5.1 | One- and two-sample *t*-tests for initial selection of estimators

Because of the large design, a selection of results to discuss had to be made. In making this selection, a number of steps were taken. Firstly, to select only the estimators with the smallest bias, Cohen's *d* of a one-sample *t*-test (reference value 0) was determined for each design cell. Next, for each of the 20 estimators of ρ^2, Cohen's *d* was averaged across all design cells. Only the estimators with $|\overline{d}| < .20$ (a small effect according to Cohen, 1988) were considered to have a sufficiently small bias to further explore them in a subsequent analysis with $\widehat{\rho}_d^2 - \rho^2$ as the dependent variable.

For the selection of methods with the smallest MSE, a reference point of 0 does not make any sense as MSE cannot be 0. Instead, in order to make a selection of methods with the smallest MSE to submit to further analysis with $\left(\widehat{\rho}_d^2 - \rho^2\right)^2$ as the dependent variable, a different reference point was chosen: if the combination strategy was averaging, then the MSE of $\overline{R^2}$ was used as a reference value; if the combination strategy was based on standardisation before pooling, then $\widehat{R}^2_{\mathrm{SP}}$ was used as a reference value. The reasoning behind these reference points is that R^2 is considered a lower benchmark, so preferably you would want an estimator of ρ^2 to have a substantially smaller MSE than that of R^2. Using these reference points, for each design cell the MSE of the specific estimator was compared with the MSE of $\overline{R^2}$ / $\widehat{R}^2_{\mathrm{SP}}$ using a two-sample *t*-test. The Cohen's *d* values of each of these *t*-tests were averaged for each estimator across design cells. Estimators with $|\overline{d}| > .80$ (\ large effect according to Cohen, 1988) were considered to sufficiently deviate from the lower benchmark to submit them to further analyses.

It should be noted that the selection criteria of $|\overline{d}| < .20$ for bias and $|\overline{d}| > .80$ for MSE are somewhat arbitrary. However, the goal of using these criteria was to filter out as many estimators for the subsequent analyses, such that it would be easier to report.

2.5.2 | Effect sizes of ANOVA

The methods that met the earlier described criteria were submitted to two ANOVAs, one with $\widehat{\rho}_d^2 - \rho^2$ as the dependent variable, and one with $\left(\widehat{\rho}_d^2 - \rho^2\right)^2$ as the dependent variable. In both ANOVAs the independent variables were sample size, value of ρ^2, number of predictors, estimator, and combination strategy. Since many significant main effects and higher-order interaction effects were expected, neither the significant *F*-tests nor the corresponding means that were tested were reported. Instead, partial η^2 was computed for each effect in the ANOVAs. Next, the bias, MSE, and their standard deviations were reported in tables, for all combinations of factors that were part of main or interaction effects with partial $\eta^2 > .13$ (a large effect according to Cohen, 1988), aggregated across the remaining factors. The numbers in the tables were visually inspected and interpreted.

2.5.3 | Two-sample *t*-tests testing bias of imputed data against complete data

For each design cell, it was tested whether the bias of the imputed data differed significantly from the bias of the corresponding data sets without missing data. For each estimator, and for all combinations of factors that were part of main or interaction effects with partial $\eta^2 > .13$ (a large effect according to

TABLE 1 Cohen's d of the bias and MSE, and bias MSE of each estimator of ρ^2, averaged across all design cells, in ascending order with respect to the value of d for bias.

Estimator	Average Cohen's d of the bias across design cells	Average bias across design cells $\times 10^3$	Average Cohen's d of the MSE across design cells	Average MSE across design cells $\times 10^3$
$\boldsymbol{R^2_E}$	**.104**	**6.53**	**.208**	**4.33**
$\boldsymbol{R^2_S}$	**.108**	**6.82**	**.211**	**4.34**
$\boldsymbol{R^2_{OP}}$	**.147**	**9.19**	**.198**	**4.44**
$\boldsymbol{R^2_{OP5}}$	**.147**	**9.19**	**.198**	**4.44**
$\boldsymbol{R^2_P}$	**.147**	**9.16**	**.199**	**4.44**
$\boldsymbol{R^2_{OP2}}$	**.147**	**9.21**	**.199**	**4.44**
$\boldsymbol{R^2_{OP1}}$	**.152**	**9.50**	**.203**	**4.45**
$\boldsymbol{R^2_{E+}}$	**.161**	**7.69**	**.227**	**4.30**
$\boldsymbol{R^2_{S+}}$	**.164**	**7.95**	**.229**	**4.31**
$\boldsymbol{R^2_{ML}}$	**.194**	**10.03**	**.271**	**4.38**
R^2_{P+}	.204	10.35	.220	4.40
R^2_{OP+}	.205	10.39	.220	4.40
R^2_{OP5+}	.205	10.39	.220	4.40
R^2_{OP2+}	.205	10.41	.220	4.41
R^2_W	.208	11.49	.267	4.47
R^2_{OP1+}	.208	10.62	.222	4.41
R^2_{W+}	.236	12.01	.273	4.46
R^2_C	.249	14.00	.259	4.60
R^2_{C+}	.278	14.55	.266	4.59
R^2	.495	25.34	–	5.70

Note: Rows with effect sizes for bias with $|\bar{d}| < .20$ are printed in bold.

Cohen, 1988) in the ANOVA, the mean Cohen's *d* across the remaining factors was reported. The same was done for MSE.

3 | RESULTS

Table 1 shows the average Cohen's *d* values for bias and MSE for each estimator, along with the mean values of bias and MSE. The table shows that there are 10 estimators with an average Cohen's *d* for bias lower than .20, namely R^2_E, R^2_S, R^2_{OP}, R^2_{OP5}, R^2_P, R^2_{OP2}, R^2_{OP1}, R^2_{E+}, R^2_{S+}, and R^2_{ML}. These estimators were submitted to the ANOVA with $\hat{\rho}^2_d - \rho^2$ as the dependent variable. Of the 10 methods with $|\overline{d}| > .20$ for bias, seven were positive-part estimators (see Table 1, first column).

As discussed in Section 2, it was originally decided for MSE to submit only the estimators with $|\overline{d}| > .80$ (a large effect according to Cohen, 1988) into a subsequent analysis. However, as can be seen in Table 1, none of the estimators met this criterion. Additionally, the average Cohen's *d* values varied little around $\overline{d} = .20$ (small effect) across the estimators. A similar finding was reported by Shieh (2008, p. 599) for complete-data estimators ρ^2, who concluded with regard to MSE that results for different estimators were inconclusive. Because both earlier research and the current study found inconclusive results regarding MSE it was decided to not further look into the MSE results, and only focus on bias in the subsequent analyses.

Next, the 10 estimators with average Cohen's *d* values below .20 for bias were submitted to the ANOVA. Table 2 shows the effect sizes of the main and interaction effects with partial $\eta^2 > .13$. Since the focus of this study is on how well the different estimators perform in multiply imputed data sets, the two most important factors in this table are estimator and combination strategy. The table shows that the factors that have large interaction effects with combination strategy are sample size, value of population coefficient of determination, percentage of missing data, and number of predictors. Factors that have large interaction effects with estimator are sample size and value of population coefficient of determination. No substantial interaction effects that included both combination strategy and estimator were found. Thus, we present two tables: one showing the results for the combination of factors that had substantial interaction effects with combination strategy, and one showing the results for the combination of factors that had substantial interaction effects with estimator.

Table 3 shows the results for combination strategy. For comparison, the results of the original data are shown as well. To save space, the results of sample sizes $N = 100$ and $N = 250$ are not shown in the table. For each combination of sample size, number of predictors, ρ^2, and percentage of missingness, the combination strategy with the least bias is printed in bold.

It can be seen from the table that the results for the original data (0% missing data) usually have a smaller bias than the results for the multiply imputed data. Furthermore, in general, for low ρ^2 (0, .20, and mostly .50) pooled estimators based on standardisation before pooling have lower bias than estimators based on averaging. For the higher values of ρ^2 (.75 and .9) this is reversed. However, for higher ρ^2 the differences in bias between both combination strategies are smaller than for low ρ^2. In other words, on average, estimators based on standardisation before pooling have smaller bias.

As for the other factors in Table 3, firstly, sample size has a substantial influence on the bias. In going from $N = 50$ to $N = 500$, the bias seems to drop by almost a factor of 10, on average. For example, compare $\rho^2 = 0$, $p = 6$, 25% missing data, method SP, and $N = 50$ ($M = .124$, $SD = .139$) with $\rho^2 = 0$, $p = 6$, 25% missing data, method SP, and $N = 500$ ($M = .011$, $SD = .013$). Furthermore, for the imputed data, bias seems to increase when the number of predictors increases. The same cannot be found in the complete data. This increase becomes larger as ρ^2 drops ($\rho^2 = 0, \rho^2 = .2$), the sample becomes smaller ($N = 50$), and the percentage of missingness increases (25%).

The results for estimator can be found in Table 4. It becomes clear from the table that for $\rho^2 = 0$ there is on average a substantial difference between the bias of the original data and the bias of the imputed data (medium effect according to $|\overline{d}|$), for all estimators. As ρ^2 increases, the difference in bias between original data and imputed data decreases. With the exception of $\rho^2 = .2$ when $N < 500$, for all the other values

TABLE 2 Effect sizes of the ANOVA with bias as the dependent variable of the large effects (partial $\eta^2 > .13$) according to Cohen (1988).

Effect	Partial η^2
Method	.815
Method × n	.775
Method × ρ^2	.631
Method × p	.568
Method × Percent	.542
Method × n × ρ^2	.553
Method × n × p	.501
Method × n × Percent	.491
Method × ρ^2 × p	.290
Method × ρ^2 × Percent	.297
Method × p × Percent	.295
Method × n × ρ^2 × p	.222
Method × n × ρ^2 × Percent	.237
Method × n × p × Percent	.237
Estimator	.314
Estimator × n	.220
Estimator × ρ^2	.472
Estimator × n × ρ^2	.330

of ρ^2 the difference in bias between original data and imputed data is on average smaller than a small effect. The results for $N = 100$ and $N = 250$ are not shown, but for these sample sizes, the difference in bias between the original data and imputed data does not meet the criterion for a small effect size either.

Furthermore, the results for estimator do not show a clear best-performing estimator regarding bias. For $\rho^2 = 0$ and $\rho^2 = .2$, R^2_{ML} usually has the largest bias whereas R^2_E has the smallest bias. For $\rho^2 = .5$, $\rho^2 = .75$, and $\rho^2 = .9$, the differences between methods are less clear. In general, for these values of ρ^2 the bias is close to 0.

Finally, it should be noted that for some combinations of sample size, estimator, and ρ^2 (especially for $\rho^2 = 0$) there is a difference in bias between the original data and the imputed data with a medium effect size. This result shows that occasionally MI introduces some (additional) bias to the estimators. As the results in Table 4 are aggregations across combination strategy as well, it cannot be seen whether this additional bias occurs both for estimators based on averaging and estimators based on standardisation before pooling. However, inspection of effect sizes for combination strategy separately revealed that occasionally, both combination strategies introduced additional bias to the estimators, although standardisation before pooling did so to a smaller degree (results not shown).

4 | DISCUSSION

In this study pooled versions of different estimators of ρ^2 in multiply imputed data sets were proposed, for which previously no solutions were available. In a simulation study the statistical properties of these proposed estimators were studied. Two quality measures were studied, namely bias and MSE. Regarding MSE the results were inconclusive, as was found in earlier research as well (Shieh, 2008). For this reason we mainly focus on bias in this discussion. Although in a design with this many factors (some of which having up to 20 levels) it may be difficult to detect clear trends, some general conclusions may still be drawn about the current study. These conclusions are discussed below.

TABLE 3 Bias (standard deviations in parentheses) $\times 10^3$ for all combinations of sample size, population coefficient of determination, number of predictors, percentage of missingness, and combination method, aggregated across estimators of the coefficient of determination.

			Percentage of missingness						
			0%	6.25%		12.5%		25%	
				Method					
N	ρ^2	p		Mean	SP	Mean	SP	Mean	SP
50	0	2	6 (39)	17 (42)	**12 (42)**	31 (53)	**19 (53)**	58 (66)	**32 (65)**
		4	10 (58)	32 (64)	**20 (64)**	64 (75)	**36 (76)**	149 (101)	**80 (106)**
		6	9 (74)	49 (84)	**29 (85)**	97 (98)	**52 (100)**	242 (128)	**124 (139)**
	.2	2	−3 (107)	4 (111)	**0 (111)**	8 (116)	**−1 (118)**	24 (124)	**4 (128)**
		4	−3 (112)	14 (115)	**4 (117)**	35 (122)	**12 (125)**	100 (133)	**42 (146)**
		6	−2 (120)	26 (124)	**9 (127)**	61 (129)	**23 (135)**	171 (139)	**74 (158)**
	.5	2	1 (100)	1 (105)	**−1 (106)**	**2 (114)**	−3 (115)	**5 (123)**	−6 (127)
		4	−2 (109)	5 (112)	**0 (113)**	12 (116)	**−2 (120)**	43 (128)	**7 (139)**
		6	−1 (109)	11 (110)	**1 (113)**	31 (114)	**7 (119)**	90 (122)	**26 (140)**
	.75	2	−7 (65)	**−9 (68)**	−9 (69)	**−9 (73)**	−11 (74)	**−12 (83)**	−17 (86)
		4	−4 (67)	**−2 (71)**	−5 (72)	−2 (75)	−9 (77)	**5 (86)**	−14 (94)
		6	−1 (64)	3 (67)	**−2 (69)**	6 (71)	**−6 (74)**	32 (83)	**−3 (97)**
	.9	2	−2 (28)	**−2 (30)**	−2 (30)	**−4 (32)**	−5 (32)	**−6 (38)**	−8 (39)
		4	−1 (29)	**−3 (32)**	−4 (33)	**−3 (36)**	−5 (36)	**−7 (48)**	−15 (52)
		6	−1 (28)	**−2 (31)**	−4 (32)	**−1 (34)**	−6 (36)	**5 (47)**	−12 (55)
500	0	2	0 (4)	2 (5)	**1 (5)**	3 (5)	**2 (5)**	3 (5)	**2 (5)**
		4	0 (5)	3 (6)	**2 (6)**	5 (7)	**3 (7)**	12 (10)	**6 (10)**
		6	1 (7)	4 (8)	**3 (8)**	8 (9)	**4 (9)**	20 (13)	**11 (13)**
	.2	2	1 (32)	1 (34)	**1 (34)**	2 (35)	**1 (35)**	2 (34)	**2 (34)**
		4	0 (33)	2 (34)	**1 (34)**	3 (36)	**1 (36)**	8 (41)	**3 (41)**
		6	0 (31)	2 (32)	**1 (32)**	5 (34)	**3 (35)**	13 (40)	**6 (40)**
	.5	2	−1 (34)	**−1 (35)**	−1 (35)	**−1 (37)**	−1 (37)	**−2 (37)**	−2 (37)
		4	0 (33)	**0 (35)**	0 (35)	1 (36)	**0 (36)**	4 (40)	**2 (40)**
		6	−2 (31)	**−1 (32)**	−1 (32)	**0 (34)**	−1 (34)	5 (39)	**1 (39)**
	.75	2	−1 (19)	**−2 (20)**	−2 (20)	**−2 (21)**	−2 (21)	**−2 (22)**	−2 (22)
		4	0 (19)	**0 (21)**	0 (21)	**0 (23)**	−1 (23)	**0 (25)**	−1 (25)
		6	0 (20)	0 (21)	**0 (21)**	**0 (22)**	−1 (26)	−2 (26)	**−1 (26)**
	.9	2	0 (9)	**−1 (9)**	−1 (9)	**−1 (9)**	−1 (9)	**−1 (10)**	−1 (10)
		4	0 (9)	**0 (9)**	0 (9)	**0 (10)**	−1 (10)	**−1 (13)**	−1 (13)
		6	0 (9)	**−1 (9)**	−1 (9)	**−1 (11)**	−1 (11)	**0 (13)**	−1 (13)

Note: The combination method with the least bias for a specific sample size, population coefficient of determination, number of predictors, and percentage of missingness is printed in bold.

4.1 | Additional bias introduced by sample size and number of predictors

Before getting into the conclusions regarding the effects of MI, the combination strategies, and the estimators, it should first be noted that sample size substantially influenced the bias. For $N = 50$, the bias could occasionally be a factor of 10 higher than for $N = 500$. Additionally, the higher the number of predictors, the more bias was found in the estimators, while this was not true for the complete data.

TABLE 4 Bias (standard deviations in parentheses) × 10^3 for all combinations of sample size, estimator, and value of ρ^2, aggregated across the remaining design factors.

			ρ^2				
N	**Estimator**	**Data sets**	**0**	**.2**	**.5**	**.75**	**.9**
50	R^2_E	Original	1 (61)	−8 (112)	−7 (106)	−8 (66)	−3 (29)
		Imputed	57 (103)**	29 (134)*	7 (122)	−8 (79)	−7 (39)
	R^2_S	Original	3 (61)	−6 (112)	−6 (106)	−8 (66)	−3 (29)
		Imputed	59 (103)**	30 (133)*	8 (121)	−7 (79)	−6 (39)
	R^2_{OP}	Original	1 (64)	−2 (115)	4 (106)	0 (65)	0 (28)
		Imputed	59 (106)**	36 (136)*	17 (122)	0 (78)	−3 (38)
	R^2_{OP5}	Original	1 (64)	−2 (115)	4 (106)	0 (65)	0 (28)
		Imputed	59 (106)**	36 (136)*	17 (122)	0 (78)	−3 (38)
	R^2_P	Original	1 (63)	−2 (115)	3 (106)	0 (65)	0 (28)
		Imputed	60 (106)**	35 (136)*	17 (122)	0 (78)	−3 (38)
	R^2_{OP2}	Original	1 (64)	−1 (115)	4 (106)	0 (65)	0 (28)
		Imputed	60 (106)**	36 (136)*	17 (122)	0 (78)	−3 (38)
	R^2_{OP1}	Original	4 (63)	0 (114)	4 (106)	0 (65)	0 (28)
		Imputed	62 (105)**	37 (135)*	17 (121)	0 (78)	−3 (38)
	R^2_{E+}	Original	24 (44)	−7 (111)	−7 (106)	−8 (66)	−3 (29)
		Imputed	70 (93)**	30 (132)*	7 (122)	−8 (79)	−7 (39)
	R^2_{S+}	Original	24 (44)	−6 (111)	−6 (106)	−8 (66)	−3 (29)
		Imputed	71 (93)**	31 (132)*	8 (121)	−7 (79)	−6 (39)
	R^2_{ML}	Original	28 (48)	4 (110)	−1 (104)	−6 (65)	−3 (29)
		Imputed	77 (95)**	40 (131)*	13 (119)	−6 (78)	−6 (39)
500	R^2_E	Original	0 (6)	0 (32)	−1 (33)	−1 (19)	−1 (9)
		Imputed	5 (9)**	3 (36)	0 (36)	−1 (23)	−1 (10)
	R^2_S	Original	0 (6)	0 (32)	−1 (33)	−1 (19)	−1 (9)
		Imputed	5 (9)**	3 (36)	0 (36)	−1 (23)	−1 (10)
	R^2_{OP}	Original	0 (6)	1 (32)	0 (33)	0 (19)	0 (9)
		Imputed	5 (9)**	3 (36)	1 (36)	0 (22)	0 (10)
	R^2_{OP5}	Original	0 (6)	1 (32)	0 (33)	0 (19)	0 (9)
		Imputed	5 (9)**	3 (36)	1 (36)	0 (22)	0 (10)
	R^2_P	Original	0 (6)	1 (32)	0 (33)	0 (19)	0 (9)
		Imputed	5 (9)**	3 (36)	1 (36)	0 (22)	0 (10)
	R^2_{OP2}	Original	0 (6)	1 (32)	0 (33)	0 (19)	0 (9)
		Imputed	5 (9)**	3 (36)	1 (36)	0 (22)	0 (10)
	R^2_{OP1}	Original	0 (6)	1 (32)	0 (33)	0 (19)	0 (9)
		Imputed	5 (9)**	3 (36)	1 (36)	0 (22)	0 (10)
	R^2_{E+}	Original	2 (4)	0 (32)	−1 (33)	−1 (19)	−1 (9)
		Imputed	6 (9)**	3 (36)	0 (36)	−1 (23)	−1 (10)
	R^2_{S+}	Original	2 (4)	0 (32)	−1 (33)	−1 (19)	−1 (9)
		Imputed	6 (9)**	3 (36)	0 (36)	−1 (23)	−1 (10)
	R^2_{ML}	Original	2 (5)	1 (32)	−1 (33)	−1 (19)	−1 (9)
		Imputed	7 (9)**	4 (36)	0 (36)	−1 (23)	−1 (10)

*$.2 \leq |\bar{d}| < .5$. **$.5 \leq |\bar{d}| < .8$.

Because this trend was seen for all estimators of ρ^2 and for both combination strategies, this finding has no consequences for general advice on which combination strategy or estimator to use. It is, however, important to keep in mind that for low sample sizes and many predictors, on average more bias is to be expected when data are incomplete and handled using MI.

4.2 | Additional bias introduced by multiple imputation

Regardless of the combination strategy and the estimator, MI generally introduced bias additional to the intrinsic bias of the estimator (if there was any). Especially for the (approximately) unbiased Olkin–Pratt estimators, this bias seems remarkable at first. To see whether this bias can be explained we must look at a necessary condition for an MI procedure to lead to unbiased results.

One necessary condition for an MI procedure to lead to unbiased results is that the procedure is *proper* (Rubin, 1987). Although we will not get into the technical details of what defines a proper MI method, one important property that is implied by its definition is that the statistic of interest has a normal sampling distribution in the case of complete data (e.g., Schafer, 1997, pp. 108–109). This is not the case for any of the estimators of ρ^2, including the Olkin–Pratt estimators (see, for example, Olkin & Pratt, 1958). This might explain why MI introduces bias into the estimators, including those that are assumed to be (approximately) unbiased in complete data. It is, however, not clear in exactly what way this violation of the assumption causes bias in the estimators. According to Schafer (1997, p. 145), it is extremely difficult to determine whether an imputation method is proper, except in some trivial cases. Considering the fact that in complete data R^2 is intrinsically biased, and that most of the estimators in the current paper are complex functions of R^2, the situations studied in the current paper will almost certainly not qualify as trivial cases.

4.3 | Standardisation before pooling is preferred over averaging

A second conclusion we can draw from this study is that on average, the bias of the standardisation-before-pooling combination strategy was smaller than the bias of the averaging combination strategy. This smaller average bias was largely caused by the fact that for low values of ρ^2 averaging produced a substantially larger positive bias than standardisation before pooling. For high values of ρ^2 the biases of both combination strategies were much closer together, and in fact averaging generally had smaller bias than standardisation before pooling. However, differences in bias between both combination strategies for high values of ρ^2 were so small that for practical purposes one may gain little by switching from standardisation before pooling to averaging when a high ρ^2 is to be expected. Besides, values of ρ^2 of .75 and .9 are extremely rare in social sciences. Thus, although there may be situations in which averaging is slightly preferred over standardisation before pooling, from a practical point of view we recommend using standardisation before pooling as a combination strategy at all times.

4.4 | Bias of the estimators

The Ezekiel estimator turned out to be the estimator with the smallest bias, on average. As expected, R^2 had the most bias. Ten out of the 20 estimators studied had a bias with an average Cohen's *d* greater than .20. Of these estimators, seven were positive-part estimators. This leads to the conclusion that in general, positive-part estimators produce more bias than non-positive-part estimators, which is in line with results based on complete data (Karch, 2020; Shieh, 2008).

However, when in practice negative values of an estimator of ρ^2 are found, it does not make sense to interpret this negative estimate, since a negative ρ^2 is not possible. This raises the question whether

unbiased estimators of ρ^2 should be striven for in the first place, because in order for them to be unbiased, estimates occasionally need to be negative. A lot can be said about the usefulness of unbiased estimators of ρ^2 (see, for example, Karch, 2020). However, it was not the goal of the current study to start a discussion on this. The goal of the current study was to propose and investigate MI variants of various estimators for ρ^2. Given that a researcher finds the reporting of an unbiased or less biased estimator of ρ^2 useful for his/her purposes, we now have more insight into how the specific estimator will be affected by the treatment of missing data using MI.

4.5 | No substantial effect of missingness mechanism

The MAR results were not discussed in Section 3, but they can be found in the Appendix. Although effect sizes found under MAR are on average smaller than for MCAR, the conclusions remain largely the same. The standardisation before pooling estimators produce substantially less bias than estimators based on averaging, and the Ezekiel estimator is the least biased on average.

4.6 | General conclusions and future research

To summarise, although MI may introduce some bias into each estimator, when using standardisation before pooling as a combination strategy rather than averaging, the bias remains relatively small. Furthermore, the Ezekiel estimator based on standardisation before pooling ($\widehat{R}^2_{E,SP}$) generally produced the smallest bias. Thus, if the goal is to obtain an estimator for ρ^2 in multiply imputed data with minimal bias, then $\widehat{R}^2_{E,SP}$ is the preferred estimator. However, if the user has a different goal when reporting an estimator for ρ^2 in multiply imputed data (e.g., providing one that cannot be negative), then it is mostly up to the user to decide which estimator for ρ^2 to use. However, having decided which estimator to use, the recommended way to get to a pooled version of this estimator in multiply imputed data is to calculate $\widehat{R}^2_{SP}$ and insert this estimator into the function of the specific estimator for ρ^2.

Finally, in this paper we only focused on missingness mechanisms MCAR and MAR. The reason for only studying MCAR and MAR was that we wanted to study the statistical properties of the pooled estimators of ρ^2 without the interference of additional bias due to missingness mechanisms that were MNAR. Future research could focus on more MAR mechanisms, and possibly MNAR mechanisms, to see whether the results of this study may be generalized to other missingness mechanisms as well.

AUTHOR CONTRIBUTIONS

Joost R. van Ginkel: Methodology; formal analysis; writing – review and editing; supervision; writing – original draft. **Julian D. Karch:** Investigation; writing – review and editing; software.

CONFLICT OF INTEREST STATEMENT

There are no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Joost R. van Ginkel https://orcid.org/0000-0002-4137-0943

REFERENCES

Alf, E. F., & Graf, R. G. (2002). A new maximum likelihood estimator for the population squared multiple correlation. *Journal of Educational and Behavioral Statistics*, *27*, 223–235. https://doi.org/10.3102/10769986027003223

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement*, *2*, 595–607. https://doi.org/10.1177/014662167800200414

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

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

Fay, R. E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, *81*, 354–365. https://doi.org/10.1080/01621459.1986.1047827

Fisher, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society of London, Series A*, *121*, 654–673. https://doi.org/10.1098/rspa.1928.0224

Galimard, J. E., Chevret, S., Protopopescu, C., & Resche-Rigon, M. (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Statistics in Medicine*, *35*, 2907–2920. https://doi.org/10.1002/sim.6902

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

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*, 475–492.

Karch, J. (2020). Improving on adjusted R-squared. *Collabra Psychology*, *6*(1), 45. https://doi.org/10.1525/collabra.343

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

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*(2), 331–334. https://doi.org/10.2466/03.11.PMS.112.2.331-348

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A*, *163*, 445–459. https://doi.org/10.1111/1467-985X.00177

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, *29*, 201–211. https://doi.org/10.1214/aoms/1177706717

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1997). Methodology review: Estimation of population validity and cross-validity, and the use of equal weights in prediction. *Applied Psychological Measurement*, *21*, 291–305. https://doi.org/10.1177/01466216970214001

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

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

Seaman, S. R., Bartlett, J., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMC Medical Research Methodology*, *12*, 46. https://doi.org/10.1186/1471-2288-12-46

Shieh, G. (2008). Improved shrinkage estimation of squared multiple correlation coefficient and squared cross-validity coefficient. *Organizational Research Methods*, *11*, 387–407. https://doi.org/10.1177/1094428106292901

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

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*, 1049–1064. https://doi.org/10.1080/10629360600810434

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

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. https://doi.org/10.1080/00273171.2018.1540967

Van Ginkel, J. R. (2020). Standardized regression coefficients and newly proposed estimators for R^2 in multiply imputed data. *Psychometrika*, *85*, 185–205. https://doi.org/10.1007/s11336-020-09696-4

Yin, P., & Fan, X. (2001). Estimating R^2 shrinkage in multiple regression: A comparison of different analytical methods. *Journal of Experimental Education*, *69*, 203–224. https://doi.org/10.1080/00220970109600656

How to cite this article: van Ginkel, J. R., & Karch, J. D. (2024). A comparison of different measures of the proportion of explained variance in multiply imputed data sets. *British Journal of Mathematical and Statistical Psychology*, *00*, 1–22. https://doi.org/10.1111/bmsp.12344

APPENDIX

In addition to the results discussed in Section 3, results were also generated under MAR. MAR was simulated in the following way. First, an $N \times (p+1)$ matrix **A** with uniform random numbers ranging from 0 to 1 was generated. Next, the matrix $\mathbf{W} = \left[\mathbf{0}_N, X_1\mathbf{1}'_p - \left(\min(X_1) - .01\right) \times \mathbf{1}_N\mathbf{1}'_p\right]$ was computed. Suppose c is the percentage of missing data. Then to simulate MAR, the $c \times N \times p$ highest entries in matrix **WA** were removed in matrix $[X, Y]$. In this way, X_1 was always observed, and the probability of missing data on the other variables increased as the values on X_1 increased. This way of simulating MAR is in accordance with Van Ginkel (2020).

The same results displayed in Table 1 for MCAR, are displayed in Table A1 but now for MAR. As can be seen, the order of the methods regarding effect size is the same as in Table 1, except for the fact that R^2_W is substantially less biased than in Table 1. However, R^2_W is still more biased on average than the 10 least biased methods in Table 1.

The estimators in Table A1 that were printed in bold were subjected to an ANOVA with the same factors as the ANOVA in the MCAR situation (note that in this ANOVA more estimators were included than for MCAR). Inspecting the effect sizes of this ANOVA (see Table A2) revealed that fewer effects met the criterion of partial $\eta^2 > .13$ than for MCAR, but those that did also met this criterion under MCAR (see Table 2). Consequently, it would not be very informative to show the results from Tables 3 and 4 for MAR, as this would largely show the same pattern.

TABLE A1 Cohen's d of the bias, and bias of each estimator of ρ^2, averaged across all design cells, in ascending order with respect to the value of d for bias, for missingness mechanism MAR.

Estimator	Average Cohen's d of the bias across design cells	Average bias across design cells $\times 10^3$	Average Cohen's d of the MSE across design cells	Average MSE across design cells $\times 10^3$
R^2_E	**.048**	**3.41**	**.165**	**3.70**
R^2_S	**.052**	**3.71**	**.167**	**3.70**
R^2_{OP}	**.094**	**6.05**	**.152**	**3.78**
R^2_{OP5}	**.094**	**6.05**	**.152**	**3.78**
R^2_P	**.094**	**6.02**	**.153**	**3.77**
R^2_{OP2}	**.094**	**6.07**	**.153**	**3.78**
R^2_{OP1}	**.100**	**6.37**	**.158**	**3.78**
R^2_{E+}	**.120**	**4.76**	**.194**	**3.65**
R^2_{S+}	**.124**	**5.01**	**.195**	**3.66**
R^2_{ML}	**.153**	**7.05**	**.235**	**3.70**
R^2_W	**.162**	**8.42**	**.227**	**3.78**
R^2_{P+}	**.166**	**7.40**	**.185**	**3.73**
R^2_{OP+}	**.166**	**7.44**	**.185**	**3.73**
R^2_{OP5+}	**.166**	**7.44**	**.185**	**3.73**
R^2_{OP2+}	**.167**	**7.45**	**.185**	**3.73**
R^2_{OP1+}	**.169**	**7.66**	**.187**	**3.73**
R^2_{W+}	**.198**	**9.03**	**.239**	**3.77**
R^2_C	.206	10.91	.209	3.89
R^2_{C+}	.242	11.55	.222	3.87
R^2	.470	22.44	–	

Note: Rows with effect sizes for bias with $|\bar{d}| < .20$ are printed in bold.

TABLE A2 Effect sizes of the ANOVA with bias as the dependent variable of the large effects (partial $\eta^2 > .13$) according to Cohen (1988) under missingness mechanism MAR.

Effect	Partial η^2
Method	.470
Method × n	.452
Method × ρ^2	.241
Method × p	.324
Method × Percent	.163
Method × n × ρ^2	.217
Method × n × p	.286
Method × n × Percent	.177
Method × ρ^2 × p	.134
Method × p × Percent	.152
Method × n × p × Percent	.149
Estimator	.270
Estimator × n	.187
Estimator × ρ^2	.325
Estimator × n × ρ^2	.210