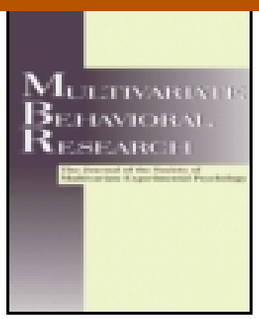

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

Significance Tests and Estimates for R^2 for Multiple Regression in Multiply Imputed Datasets: A Cautionary Note on Earlier Findings, and Alternative Solutions

Joost R. van Ginkel

To cite this article: Joost R. van Ginkel (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, DOI: 10.1080/00273171.2018.1540967

To link to this article: https://doi.org/10.1080/00273171.2018.1540967

MULTIVARIATE BEHAVIORAL RESEARCH
https://doi.org/10.1080/00273171.2018.1540967

Significance Tests and Estimates for R^2 for Multiple Regression in Multiply Imputed Datasets: A Cautionary Note on Earlier Findings, and Alternative Solutions

Joost R. van Ginkel

Department of Methodology and Statistics, Leiden University, Leiden, The Netherlands

ABSTRACT

Whenever multiple regression is applied to a multiply imputed data set, several methods for combining significance tests for R^2 and the change in R^2 across imputed data sets may be used: the combination rules by Rubin, the Fisher z-test for R^2 by Harel, and F-tests for the change in R^2 by Chaurasia and Harel. For pooling R^2 itself, Harel proposed a method based on a Fisher z transformation. In the current article, it is argued that the pooled R^2 based on the Fisher z transformation, the Fisher z-test for R^2, and the F-test for the change in R^2 have some theoretical flaws. An argument is made for using Rubin's method for pooling significance tests for R^2 instead, and alternative procedures for pooling R^2 are proposed: simple averaging and a pooled R^2 constructed from the pooled significance test by Rubin. Simulations show that the Fisher z-test and Chaurasia and Harel's F-tests generally give inflated type-I error rates, whereas the type-I error rates of Rubin's method are correct. Of the methods for pooling the point estimates of R^2 no method clearly performs best, but it is argued that the average of R^2's across imputed data set is preferred.

KEYWORDS

Missing data; multiple imputation; multiple regression; coefficient of determination

Introduction

In multiple regression, the coefficient of determination, R^2, is the squared correlation between the observed values of the outcome variable y, and its predicted values. To test whether the population coefficient of determination, denoted ρ^2, is 0, an F-test is used. Suppose k is the number of predictors in the regression model, and N is the sample size, the F-test is computed as

$$F = \frac{R^2/k}{(1-R^2)/(N-k-1)} \tag{1}$$

which, under the assumption of normality of the errors, has an F-distribution with k numerator degrees of freedom, and $N-k-1$ denominator degrees of freedom.

When researchers want to test a large model with k_2 predictors against a smaller model with k_1 ($k_1 < k_2$) predictors, an F-test may be used for testing the change in R^2 for significance, denoted ΔR^2. Suppose that R_1^2 is the R^2 of the smaller model and R_2^2 is the R^2 of the larger model. The F-test for testing ΔR^2 for significance is given by

$$F = \frac{(R_2^2 - R_1^2)/(k_2 - k_1)}{(1-R_2^2)/N - k_2 - 1}. \tag{2}$$

For an overview of regression and its statistical tests, see Chatterjee and Hadi (1999).

The computation of both $(\Delta)R^2$ and the F-tests may be complicated by missing data. A highly recommended technique to handle missing data is multiple imputation (Rubin, 1987; Van Buuren, 2012). The complete multiple imputation process consists of three steps: (1) the missing data are estimated several times (M) using a stochastic model that accurately describes the data, creating M plausible complete versions of the incomplete data set, (2) each completed data set is analyzed using the same statistical analysis, resulting in M different outcomes of this analysis, and 3) the M analyses are combined into one analysis, using specific formulas that take into account the additional uncertainty due to the missing data in the standard errors and statistical tests. Such formulas for obtaining

CONTACT Joost R. Van Ginkel jginkel@fsw.leidenuniv.nl Department of Methodology and Statistics, Leiden University, PO Box 9500, 2300 RB Leiden, The Netherlands.

This article has been republished with minor changes. These changes do not impact the academic content of the article.

overall statistics from multiply imputed data sets are henceforth denoted *combination rules*.

Rubin (1987) provided a general set of combination rules for a parameter estimate of a statistical model, its standard error, its error degrees of freedom, and its significance test. Additionally, Rubin provided multivariate extensions of these rules for testing multiple parameter estimates for significance simultaneously. Barnard and Rubin (1999) and Reiter (2007) developed improved error degrees of freedom for these combination rules. Simulation studies (Barnard & Rubin, 1999; Grund, Lüdtke, & Robitzsch, 2016; Li, Raghunathan, & Rubin, 1991; Liu & Enders, 2017; Reiter, 2007; Schafer, 1997) have shown that these combination rules generally give type-I error rates close to the theoretical type-I error rates.

In the specific context of regression, Harel (2009) proposed combination rules for R^2 and its significance test. Chaurasia and Harel (2014) offered combination rules for ΔR^2 and its significance test. Simulations (Chaurasia & Harel, 2014; Harel, 2009) showed that their methods gave type-I error rates close to the theoretical type-I error rates. However, both methods have some theoretical flaws. Additionally, the situations under which these methods were studied have, to the author's opinion, limited relevance in practice. Given the flaws and the circumstances under which the methods were studied, the question is whether these results will generalize to slightly different, more relevant situations.

Fortunately, the general combination rules (Barnard & Rubin, 1999; Li, Meng, Raghunathan, & Rubin, 1991; Reiter, 2007; Rubin, 1987) can also serve for pooling the *F*-values for R^2 and ΔR^2. However, besides the references for these rules being rather technical, they either only briefly or implicitly state the suitableness of these methods for testing R^2 and ΔR^2 for significance. Harel (2009) and Chaurasia and Harel (2014) on the other hand, are very explicit in stating that their methods are meant for this purpose. Consequently, the average applied researcher may not be aware of the existence of a better alternative than their methods. Furthermore, no alternatives to the methods by Harel (2009) and Chaurasia and Harel (2014) for pooling the point estimates for R^2 and ΔR^2 exist as of yet. Consequently, applied researchers who want to carry out a regression to a multiply imputed data set may use Harel's and Chaurasia and Harel's methods, and might end up drawing incorrect conclusions. The above issues were the motivation for the current article.

The current article has three goals. The first goal is to show the theoretical flaws of the methods by Harel (2009) and Chaurasia and Harel (2014). The second goal is to explicitly formulate the earlier combination rules (Rubin, 1987) as a suitable alternative for testing R^2 and ΔR^2 for significance, and propose alternative combination rules for the point estimate of R^2. The third goal is to empirically demonstrate the flaws of the methods by Harel (2009) and by Chaurasia and Harel (2014), and compare them with the proposed alternatives, in more relevant situations. To this end, two simulation studies were carried out. Besides apparent similarities between these studies and the studies by Harel (2009) and Chaurasia and Harel (2014), both studies also had some overlap with a more recent study by Liu and Enders (2017). The specific similarities and differences between these studies and the current two studies will be discussed in the methods section.

In the next sections, the general combination rules for multiple imputation (Rubin, 1987), and the combination rules by Harel (2009), and Chaurasia and Harel (2014) along with their flaws are discussed. Next, it is explained how the general combination rules can be used for testing R^2 and ΔR^2 for significance, and alternative pooled measures for R^2 are proposed. After that, two simulation studies comparing the different methods are discussed, and all methods are applied to an empirical data example. Finally, conclusions are drawn about the results, and guidelines for pooling the results of (significance tests of) R^2 and ΔR^2 in multiply imputed data, are given.

Rubin's combination rules

Single-parameter estimates

Suppose $\hat{Q}$ is the sample estimate of parameter Q for complete data, and U is its variance. Each imputed data set m ($m = 1, \ldots, M$) has an estimate of $\hat{Q}$, denoted $\hat{Q}_m$, and a variance U_m. The overall estimate of Q is

$$\overline{Q} = \frac{1}{M}\sum_{m=1}^{M}\hat{Q}_m. \tag{3}$$

The overall variance T of $\overline{Q}$ consists of two parts, namely the *within-imputation variance* $\overline{U}$, and the *between-imputation variance* B, and are computed as follows:

$$\overline{U} = \frac{1}{M}\sum_{m=1}^{M}U_m, \tag{4}$$

and

$$B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{Q}_m - \overline{Q}\right)^2, \tag{5}$$

respectively. The overall variance T then becomes

$$T = \overline{U} + (1 + M^{-1})B. \tag{6}$$

The idea behind the additional term $(1 + M^{-1})B$ is that the additional uncertainty caused by the missing data are incorporated in the variance (and thus in the standard error) of $\overline{Q}$. Consequently, this adjusts the p-values and confidence intervals for the additional uncertainty due to the missing data. To test whether a parameter is equal to a specific population value Q_0, the following statistic is used:

$$t_{Ru} = \frac{\overline{Q} - Q_0}{\sqrt{T}}, \tag{7}$$

which has an approximate t-distribution with ν^{BR} (Barnard & Rubin, 1999) degrees of freedom. ν^{BR} is computed as:

$$\nu^{BR} = \left(\frac{1}{\nu^1} + \frac{1}{\nu_{obs}}\right)^{-1}, \tag{8}$$

$$\nu^1 = (M-1)\left[1 + \frac{\overline{U}}{(1 + M^{-1})B}\right],^2$$

$$\nu_{obs} = \left[1 - \frac{(1 + M^{-1})B}{T}\right]\nu^*_{com},$$

$$\nu^*_{com} = \left(\frac{\nu_{com} + 1}{\nu_{com} + 3}\right)\nu_{com}.$$

where ν_{com} is the number of degrees of freedom in case of complete data.

The above-described combination rules are available in IBM SPSS 25.0 (2017), using an old approximation (Rubin, 1987) of the number of degrees of freedom. The approximation from Barnard and Rubin (1999) can be applied using an SPSS macro by Van Ginkel (2010). Other software packages that include these combination rules are SAS 9.4 (SAS Institute, Inc., 2013) in the procedure MIAnalyze (Yuan, 2011), Stata 14.0 (ICE; StataCorp, 2015), and the `pool()` function of the mice package (Van Buuren & Groothuis-Oudshoorn, 2011) in R (R Core Team, 2017).

Multiparameter estimates

For testing several parameters for significance simultaneously, several solutions are available (Li, Raghunathan, et al., 1991; Li et al., 1991; Meng & Rubin, 1992; Rubin, 1987). Of these solutions, the most promising one (Li, Raghunathan, et al., 1991; Rubin, 1987) according to several simulation studies (Grund et al., 2016; Li, Raghunathan, et al., 1991; Liu & Enders, 2017; Reiter, 2007) is a set of formulas that are multivariate extensions of Equations (3)–(8). This solution will be compared with the combination rules by Harel (2009) and by Chaurasia and Harel (2014).

Suppose $\hat{\mathbf{Q}}$ is a k-dimensional vector of estimates of parameter vector $\mathbf{Q}$ that would have been obtained if no data were missing, and $\mathbf{U}$ is its covariance matrix. For imputed data set m an estimate of $\hat{\mathbf{Q}}$ is denoted $\hat{\mathbf{Q}}_m$, and its covariance matrix $\mathbf{U}_m$. The overall estimate $\overline{\mathbf{Q}}$ is

$$\overline{\mathbf{Q}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\mathbf{Q}}_m. \tag{9}$$

The within-imputation covariance matrix is computed as

$$\overline{\mathbf{U}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{U}_m, \tag{10}$$

and the between-imputation covariance matrix $\mathbf{B}$ is computed as

$$\mathbf{B} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\mathbf{Q}}_m - \overline{\mathbf{Q}}\right)\left(\hat{\mathbf{Q}}_m - \overline{\mathbf{Q}}\right)'. \tag{11}$$

The overall covariance matrix $\mathbf{T}$ is

$$\mathbf{T} = (1 + r)\overline{\mathbf{U}}, \tag{12}$$

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

where r is the *relative increase in variance due to nonresponse*. To test whether the parameter vector is equal to the vector under the null hypothesis, $\mathbf{Q}_0$, we use

$$F_{Ru} = \left(\overline{\mathbf{Q}} - \mathbf{Q}_0\right)'\mathbf{T}^{-1}\left(\overline{\mathbf{Q}} - \mathbf{Q}_0\right)/k, \tag{13}$$

which has an approximate F-distribution with k numerator degrees of freedom and ν^{Rt} (Reiter, 2007) denominator degrees of freedom. The latter is computed as follows:

$$\nu^{Rt} = 4 + \frac{1}{z}, \tag{14}$$

$$z = \frac{1}{\nu^*_{com} - 4(1+a)} + \frac{1}{q-4}\left(\frac{a^2\left[\nu^*_{com} - 2(1+a)\right]}{(1+a)^2\left[\nu^*_{com} - 4(1+a)\right]}\right)$$

$$+ \frac{1}{q-4}\left(\frac{8a^2\left[\nu^*_{com} - 2(1+a)\right]}{(1+a)\left[\nu^*_{com} - 4(1+a)\right]^2} + \frac{4a^2}{(1+a)\left[\nu^*_{com} - 4(1+a)\right]}\right)$$

$$+ \frac{1}{q-4}\left(\frac{4a^2}{\left[\nu^*_{com} - 4(1+a)\right]\left[\nu^*_{com} - 2(1+a)\right]} + \frac{16a^2\left[\nu^*_{com} - 2(1+a)\right]}{(1+a)\left[\nu^*_{com} - 4(1+a)\right]}\right)$$

$$+ \frac{1}{q-4}\left(\frac{8a^2}{\left[\nu^*_{com} - 4(1+a)\right]^2}\right),$$

$$a = \frac{rq}{q-2},$$

$$q = k(M-1).$$

These combination rules for multiparameter estimates can be applied in SAS 9.4 (SAS Institute, Inc.,

2013; Yuan, 2011), in Stata 14.0 (2015), in the `MIwaldtest()` function of the `miceadds` package in R (Robitzsch, Grund, & Henke, 2017), and SPSS using Van Ginkel's (2010) macro. Also, see Van Ginkel and Kroonenberg (2014) who used this macro for pooling the results of ANOVA obtained from multiply imputed data sets.

Combination rules for regression analysis by Harel, and Chaurasia and Harel

Combination rules for R^2 and its significance test

Harel (2009) argues that averaging R^2 (Equation (3), applied to R^2) is not justified because the rules for single-parameter estimates were defined under the assumption that the sampling distribution of the parameter estimate is normal. R^2 is not normally distributed as it is bounded to values between 0 and 1. Schafer (1997, p. 109) argues that for pooling significance tests of correlations and pooling correlations themselves, it is advisable to use a Fisher z transformation and to apply the combination rules for single parameter estimates to the Fisher z transformed correlation.

Harel (2009) applied Schafer's procedure to the square root of R^2 in multiple regression, justifying this by arguing that $\sqrt{R^2}$ is a correlation as well. Suppose R^2_m is the coefficient of determination for imputed data set m. First, R^2_m is transformed to a Fisher z_F-score:

$$z_{F,m} = \frac{1}{2}\ln\frac{1+\sqrt{R^2_m}}{1-\sqrt{R^2_m}}. \tag{15}$$

Next, let $\hat{Q}_m = z_{F,m}$ and substitute this result into Equations (3) and (5). Under the null hypothesis of no association, the variance of $z_{F,m}$ is $U_m = 1/(N-3)$, and may be substituted into Equation (4). Next, using Equations (6) and (7), R^2 is tested for significance. The resulting statistic is denoted $t_{Ru,Ha}$. Finally, to get a pooled estimate of R^2, the pooled Fisher z score $\overline{Q}$, obtained from Equation (3), is transformed back:

$$\Re^2 = \left[\frac{\exp(2\overline{Q})-1}{\exp(2\overline{Q})+1})\right]^2. \tag{16}$$

The above pooling procedure has been implemented in the mice package in R (see the `pool.r.squared()` function by Van Buuren & Groothuis-Oudshoorn, 2011).

Potential problems of Harel's combination rules

Incorrect justification. Harel's argument of nonnormality of R^2 would be valid if R^2 were tested using a t-test as if it were a single parameter estimate. However, this is not how R^2 is tested for significance. In complete data, R^2 is tested using the F-test from Equation (1). In this F-test the normality assumption does not concern the sampling distribution of R^2 but the distribution of the residuals in the regression model (Fox, 2016, pp. 107, 112; Tabachnick & Fidell, 2013, pp. 124–126). As long as the distribution of the residuals is normal, no assumption of normality is violated. Likewise, when R^2 is tested for significance in a multiply imputed data set, we need an equivalent of the F-test in Equation (1) for multiply imputed data sets. In using this equivalent the distribution of the residuals needs to be normal, not the sampling distribution of R^2. In short, Harel's argument for transforming nonnormally distributed parameters does not apply here.

Incorrect assumption of normality of the Fisher z transformation. While a Fisher z transformation of a correlation between two variables is approximately normally distributed, this is not the case for $z_{F,m}$ (Equation (15)). The Fisher z transformation stretches the lower bound of a correlation of -1 to minus infinity, and the upper bound of $+1$ to plus infinity. $\sqrt{R^2}$ differs from an ordinary correlation in that it can only range from 0 to 1. Consequently, $z_{F,m}$ ranges only from 0 to plus infinity, and so it is not normally distributed.

Since only the positive value of $\sqrt{R^2}$ is used in the formulation, dividing p by 2 will solve this problem for complete data. However, this would not work for multiply imputed data. To illustrate this, consider the bivariate case where the Fisher z transformed correlations of the M imputed data sets are substituted for the $\hat{Q}_m$'s in Equation (3). If the same is done with the M Fisher z transformed $\sqrt{R^2_m}$'s, this will not give the same absolute value of $\overline{Q}$ when in some of the imputed data sets the correlation is negative. The resulting $\Re^2$ (Equation (16)) will be overestimated because in averaging it ignores the signs of the M correlations. Likewise, in a multiple regression model possible sign differences among the M regression coefficients of a specific predictor are ignored when Harel's method is applied to the $\sqrt{R^2_m}$'s. Again, the resulting $\Re^2$ will be overestimated, and $t_{Ru,Ha}$ will be too large. This problem may especially occur when the relation between a predictor and y is weak.

Ignoring the model degrees of freedom in the significance test. $\sqrt{R^2}$ differs from an ordinary correlation in that it is not a correlation between two different

variables, but a correlation between the observed and expected values of the same variable y. The expected values have been obtained from a regression model with k predictors. Normally k is incorporated in the F-test as the model degrees of freedom. Different numbers of model degrees of freedom result in different critical F-values. If a Fisher z-test is used, the model degrees of freedom are not incorporated, which will consequently lead to incorrect p-values.

Combination Rules for ΔR^2 and its Significance Test

For pooling the significance test of ΔR^2, Chaurasia and Harel (2014) proposed the following procedure. Suppose $\Re_1^2$ is the pooled estimate of R^2 of the smaller model and $\Re_2^2$ is the pooled estimate of R^2 of the larger model (both obtained using Equation (16)). Furthermore, substitute $z_{F2,m}$ for $\hat{Q}_m$ in Equations (3) and (5), use $U_m = 1/(N-k_2-2)$ in Equation (4), compute T using Equation (6), and compute r in Equation (12) using $\overline{\mathbf{U}}$ and $\mathbf{B}$ (which are for a single-parameter equivalent to $\overline{U}$ and B, respectively). Finally, using these quantities, compute either ν_2^{BR} (Equation (8)) or ν_2^{Rt} (Equation (14)) using $\nu_{com}^* = N-k_2-1$. Two pooled F-values across M imputed data sets are computed as follows:

$$F_{BR} = \frac{(\Re_2^2 - \Re_1^2)/(k_2 - k_1)}{(1-\Re_2^2)/\nu_2^{BR}}, \tag{17}$$

and

$$F_{Rt} = \frac{(\Re_2^2 - \Re_1^2)/(k_2 - k_1)}{(1-\Re_2^2)/\nu_2^{Rt}}, \tag{18}$$

respectively, both with $k_2 - k_1$ model degrees of freedom.

Potential problems of Chaurasia and Harel's combination rules

Underestimation of ΔR^2. Like in Harel's (2009) method, the estimates $\Re_1^2$ and $\Re_2^2$ may suffer from the fact that possible sign differences among the M regression coefficients of a specific predictor are ignored. For $\Re_1^2$ this problem will be more severe than for $\Re_2^2$ because $\Re_1^2$ is closer to 0. As a result, $\Re_2^2 - \Re_1^2$ might be an underestimation of ΔR^2. This problem may especially occur for weak relations between the predictors of the smaller model on the one hand and y on the other hand.

Inclusion of between-imputation variance in error degrees of freedom. A potentially more serious problem is the way the additional variation due to the missing data B (Equation 5) is incorporated in both F_{BR} and F_{Rt}. In both tests B is included in the error degrees of freedom. Both ν_2^{BR} and ν_2^{Rt} are only approximations of the actual number of degrees of freedom. It is unknown how the use of these approximations would affect the outcome of F_{BR} and F_{Rt}. Some work (Van Ginkel & Kroonenberg, 2014) suggests that for large percentages of missingness and small M, ν_2^{Rt} may become too low for what is considered a reasonable estimate of the error degrees of freedom. When used for constructing the reference distribution, this may not be problematic because for fairly large N, critical F-values do not vary much across different numbers of error degrees of freedom. However, the use of these approximations may be more influential when used for calculating a pooled F-test itself.

No justification for calculation within-imputation variance. Chaurasia and Harel (2014, p. 435) use $U_m = 1/(N-k_2-2)$ as the within-imputation variance for computing ν_2^{BR} and ν_2^{Rt} in Equations (17) and (18), respectively. However, they neither give a reference, nor a justification for this generalization of $1/(N-3)$ to the multivariate case. This raises the question whether the use of this variance estimate is justified, and performs well at all times.

Although Chaurasia and Harel (2014, p. 435) did not explicitly mention the above-mentioned problems, they do recognize that their methods are "ad hoc." They justify their procedures by their ease in terms of computation and implementation. However, due to the issues raised above it may be wondered whether their methods just happened to perform well in the specific situations studied, possibly because the several potential sources of bias may have canceled each other out.

Alternatives to the methods by Harel, and Chaurasia and Harel

Rubin's combination rules used as significance tests for R^2 and ΔR^2

Pooling significance tests of R^2. Let $\mathbf{b}_m$ be a vector of all regression coefficients (excluding the intercept), $\mathbf{X}_m$ be a matrix of predictors of the regression model, and $s_{\varepsilon,m}^2$ be the error variance, in imputed data set m. The covariance matrix of y_m is $\mathbf{V}_m = s_{\varepsilon,m}^2 \mathbf{I}_N$. Next, let $\hat{\mathbf{Q}}_m = \mathbf{b}_m$, and $\mathbf{U}_m = (\mathbf{X}_m' \mathbf{V}_m^{-1} \mathbf{X}_m)^{-1}$. Using these values for $\hat{\mathbf{Q}}_m$ and $\mathbf{U}_m$, and using $\nu_{com} = N-k-1$ (Equation (8)) for computing ν_{com}^* in Equation (14), F_{Ru} for testing all regression coefficients simultaneously is computed using Equations (9)–(14). Since the null hypothesis that all population regression coefficients are 0 is equivalent to testing the null hypothesis that $\rho^2 = 0$,

Rubin's (1987) F_{Ru} actually tests the null hypothesis of $\rho^2 = 0$.

Pooling significance tests of ΔR^2. Let $\mathbf{b}_{2,m}$ be a vector of regression coefficients of the larger model, and let $s^2_{\varepsilon 2,m}$ be the error variance of the larger model, in imputed data set m. Define $\mathbf{b}_{k2-k1,m}$ as a subset of $\mathbf{b}_{2,m}$ with dimension (k_2 - k_1), containing only the regression coefficients of the newly added predictors, and let $\hat{\mathbf{Q}}_m = \mathbf{b}_{k2-k1,m}$. Third, let $\mathbf{X}_{k2-k1,m}$ be an $N \times (k_2 - k_1)$ matrix containing only the newly added predictors, and let $\mathbf{U}_m = (\mathbf{X}_{k2-k1,m}{}'\mathbf{V}^{-1}_{2,m}\mathbf{X}_{k2-k1,m})^{-1}$ be the covariance matrix of the regression coefficients of the newly added predictors ($\mathbf{V}_{2,m} = s^2_{\varepsilon 2,m}\mathbf{I}_N$), in imputed data set m. Using these newly defined $\hat{\mathbf{Q}}_m$ and $\mathbf{U}_m$, and using $\nu_{com} = N - k_2 - 1$, we can compute F_{Ru} for ΔR^2 using Equations (9)–(14). Since testing the null hypothesis that $\Delta\rho^2 = 0$ is equivalent to testing the null hypothesis that all population coefficients of the newly added predictors are 0, F_{Ru} in this context tests the null hypothesis of $\Delta\rho^2 = 0$.

Alternative methods for pooling R^2

Two alternative combination methods to $\Re^2$ will be proposed below. Firstly, since the justification for using a Fisher z transformation in calculating $\Re^2$ is incorrect, it could be argued that this transformation might be put aside altogether, and average all R^2_m's directly. This pooled version of R^2 is denoted $\overline{R^2}$. Although this procedure is ad hoc, it is not based on an *incorrect* justification, and it is simpler to calculate than $\Re^2$.

Because both $\overline{R^2}$ and $\Re^2$ have a lower bound of 0, a problem of both measures is that the sampling errors of the R^2_m's and $z_{F,m}$'s in the direction of the middle will on average be larger than the sampling errors in the direction of the lower bound when ρ^2 is close to 0. As a result, both measures may move to the center. Additionally, as $\overline{R^2}$ has an upper bound of 1, this problem may occur for $\overline{R^2}$ when ρ^2 is close to 1 as well.

Alternatively, a pooled R^2 can be constructed that does not use any averaging at all. By back-transforming F to R^2 using the relation between the F-test and R^2 (Equation (1)), R^2 can be written as $R^2 = F/[F + (N - k - 1)k^{-1}]$. This formula may easily be generalized to multiply imputed data by means of:

$$R^2_F = \frac{F_{Ru}}{F_{Ru} + \nu^{Rt}/k}. \tag{19}$$

A potential disadvantage of R^2_F is that it uses ν^{Rt} in its calculation. As already mentioned, some work (Van Ginkel & Kroonenberg, 2014) suggests that ν^{Rt} as an a approximation may be too low for large percentages of missingness and low M.

To summarize, $\overline{R^2}$ is the simplest way to pool R^2 and has neither a correct nor incorrect justification. However, for values of ρ^2 close to 0 or 1, its estimate may move toward the center of the scale because of floor and ceiling effects. R^2_F does not have this problem, but uses approximation ν^{Rt} which might not always be accurate. The question is which of these problems are more influential, and how both estimates compare to the already existing $\Re^2$.

Table 1 gives an overview of all the pooled significance tests and point estimates for $(\Delta)R^2$ in multiply imputed data sets.

Method

Two simulation studies assessed the performance of the methods by Harel and Chaurasia and Harel. In study 1, the performance of Harel's (2009) $\Re^2$ and its significance test $t_{Ru,Ha}$ were compared with the performance of the newly proposed $\overline{R^2}$ and R^2_F, and significance test F_{Ru} (Rubin, 1987). In study 2, the performance of $\Delta\Re^2$ and its significance tests, F_{Rt} and F_{BR} (Chaurasia & Harel, 2014), were compared with $\Delta\overline{R^2}$ and ΔR^2_F, and the significance test F_{Ru}, respectively.

Properties of the simulation study were largely based on Harel's (2009) study. The general form of the regression model that will be the basis for the simulations is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \tag{20}$$

(a regression model with three predictors). The simulations were programmed in R.

It should be noted that in a study by Liu and Enders (2017) about combination techniques for regression analysis of multiply imputed data, the number of predictors in the regression model was varied. The purpose of the study by Liu and Enders was to study the robustness of several combination rules under many different circumstances, among which different numbers of predictors. The purpose of the current studies was to specifically show the flaws of the methods by Harel (2009) and Chaurasia and Harel (2014). The number of predictors was not expected to have any influence on the visibility of these flaws.

Constant factors. In both studies, within each design cell, twenty-five hundred ($D = 2500$) replications were drawn of $N = 100$. Unlike Liu and Enders (2017) who studied different sample sizes, sample size was kept constant here, because if the flaws of Harel's method already become visible under a fairly large N, it is unnecessary to study it under a smaller, or other

Table 1. Overview of significance tests and pooled estimates for $(\Delta)R^2$ in multiply imputed data sets.

Parameter	Significance test (equation no.)	Author(s) and Author(s) who contributed	Pooled estimate (equation no.)	Description pooled estimate for $(\Delta)R^2$	Author(s) and author(s) who contributed
ρ^2	F_{Ru} (13)	Rubin (1987) and Reiter (2007)	R^2_F (19)	R^2 based on F_{Ru}, backwards computation	Van Ginkel (current article), Reiter (2007) and Rubin (1987)
			$\bar{R}^2$ (3*)	Average R^2 across imputed data sets.	Van Ginkel (current article), Rubin (1987)
	$z_{F,m}$ (15)	Harel (2009), Schafer (1997) and Rubin (1987)	$\mathfrak{R}^2$ (16)	Average R^2 based on Fisher z transformation	Harel (2009), Schafer (1997), and Rubin (1987)
$\Delta\rho^2$	F_{Ru} (13)	Rubin (1987) and Reiter (2007)	$\Delta R^2_F = R^2_{F,2}$ (19) $-R^2_{F,1}$ (19)	ΔR^2 based on F_{Ru}, backwards computation	Van Ginkel (current article), Reiter (2007) and Rubin (1987)
			$\Delta\bar{R}^2 = \bar{R}^2{}_2$ (3*) $-\bar{R}^2{}_1$ (3*)	Average ΔR^2 across imputed data sets.	Van Ginkel (current article), Rubin (1987)
	F_{BR} (17)	Chaurasia and Harel (2014), Barnard and Rubin (1999), and Schafer (1997)	$\Delta\mathfrak{R}^2 = \mathfrak{R}^2_2$ (16) $-\mathfrak{R}^2_1$ (16)	Average ΔR^2 based on Fisher's z transformation	Chaurasia and Harel (2014), Schafer (1997)
	F_{Rt} (18)	Chaurasia and Harel (2014), Reiter (2007) and Schafer (1997)	$\Delta\mathfrak{R}^2 = \mathfrak{R}^2_2$ (16) $-\mathfrak{R}^2_1$ (16)	Average ΔR^2 based on Fisher's z transformation	Chaurasia and Harel (2014), Schafer (1997)

*Where $\hat{Q}_m = R^2_m$

N. See, Chaurasia and Harel (2014, p. 436), who used the reversed reasoning for using only one small N to show the robustness of their method.

In both studies, missing data were simulated under missingness mechanism *missing completely at random* (MCAR; Little & Rubin, 2002, p. 10). Under MCAR missing data are randomly scattered across the data and are not related to observed background variables, as in *missing at random* (MAR; Little & Rubin, 2002, p. 10), or on unobserved data, as in *not missing at random* (NMAR; Little & Rubin, 2002).

Besides MCAR being sufficient for the scope of the current paper (i.e., showing the flaws of the methods by Harel and by Chaurasia and Harel), MCAR also allows including complete case analysis (CCA; i.e., deleting all cases with at least one missing value from the analysis) as a lower benchmark (to be discussed shortly). CCA may give an impression of the lowest possible power from the incomplete data, but this method is not guaranteed to give unbiased results under MAR and NMAR.

In both studies, complete data were simulated using a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_y) = (2{,}5{,}10{,}20)$, and two different covariance structures, to be discussed in the independent variables section. The method for multiple imputation was chosen to be fully conditional specification using regression (Van Buuren, 2012, pp. 108–116; Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006).

Independent variables. In study 1, two covariance structures were studied: one where $\rho^2 = 0$, and one where $\rho^2 = 0.18$. More specifically, the covariance structures under $\rho^2 = 0$ and $\rho^2 = 0.18$, were

$$\rho^2 = 0 \qquad\qquad \rho^2 = 0.18,$$

$$\boldsymbol{\Sigma} = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & y \end{matrix} \\ \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 5 & 1 & 0 \\ 0 & 1 & 5 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix} \end{array} \quad \boldsymbol{\Sigma} = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & y \end{matrix} \\ \begin{bmatrix} 5 & 0 & 0 & 3 \\ 0 & 5 & 1 & 0 \\ 0 & 1 & 5 & 0 \\ 3 & 0 & 0 & 10 \end{bmatrix} \end{array}.$$

In study 2, it was assumed that variable x_1 in the regression model of Equation (20) was the only predictor in the smaller model and that variables x_2 and x_3 were added to the smaller model to form the larger model. Chaurasia and Harel (2014) added only one additional variable in their study. The reason for using two additional variables here was that when one variable is added, the significance test can already be pooled using Rubin's rules for single-parameter estimates (Equations (3)–(8)). This was considered a trivial case that neither required the development of new combination methods (F_{Rt} and F_{BR}) nor an

explicit formulation of Rubin's rules for this context. This led to two covariance structures in study 2, one with $\Delta\rho^2 = 0$ and one with $\Delta\rho^2 = 0.14$:

$$\Delta\rho^2 = 0 \qquad\qquad \Delta\rho^2 = 0.14,$$

$$\boldsymbol{\Sigma} = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & y \end{matrix} \\ \begin{bmatrix} 5 & 0 & 0 & 3 \\ 0 & 5 & 1 & 0 \\ 0 & 1 & 5 & 0 \\ 3 & 0 & 0 & 10 \end{bmatrix} \end{array} \quad \boldsymbol{\Sigma} = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & y \end{matrix} \\ \begin{bmatrix} 5 & 1 & 1 & 3 \\ 1 & 5 & 1 & 2 \\ 1 & 1 & 5 & 3 \\ 3 & 2 & 3 & 10 \end{bmatrix} \end{array}.$$

Contrary to Liu and Enders (2017) who studied three different effect sizes along with the null model ($\rho^2 = 0$), in the current studies only one alternative effect size was studied. When the robustness of several combination rules for regression in multiply imputed data is studied (like Liu & Enders, 2017) it is important to study the behavior of combination techniques under various circumstances. In the current article, however, it was only important to show that the methods by Harel (2009) and Chaurasia and Harel (2014) would give incorrect type-I error rates under the null model, and that their observed power values would deviate more from those of the original data without missing data, than the combination rules by Rubin (1987).

In both studies of the current article, four percentages of missingness were simulated: 6.25%, 12.50%, 25%, and 50%. Finally, the number of imputations was varied to be $M = 6$, $M = 25$, $M = 100$, and $M = 250$ in both studies. Note that for $M = 5$ (which is a more commonly used small number of imputations. See Schafer, 1997), ν^{Rt} (Equation (14)) will reduce to 4 when $k = 1$, regardless of the sample size. Because this will heavily influence the value of F_{Rt} (Equation (18)) in study 2, the smallest number of imputations was chosen to be $M = 6$, rather than $M = 5$.

Dependent variables. In study 1, three different pooled versions of R^2 were studied: $\Re^2$, $\overline{R^2}$, and R^2_F. In study 2, $\Delta\Re^2$, $\Delta\overline{R^2}$, and ΔR^2_F were studied. Additionally, in both studies $(\Delta)R^2$ based on CCA was studied as a lower benchmark. Note that when $\rho^2 = 0$, the R^2 measures are expected to be positively biased because when $\rho^2 = 0$, the sample estimate R^2 can only be 0 or larger. However, the question is how much larger the bias in the three measures will be compared to the bias in R^2 of the same data without missing values, how much smaller they will be than for CCA, and how much they will differ with each other. The inclusion of different combination rules for $(\Delta)R^2$ in the current two studies extends the work by Liu and Enders (2017), who only studied properties of combination rules for statistical tests of R^2.

Furthermore, Harel (2009) used a simulation model with $\rho^2 = 0.75$, and studied the percentage of times the hypothesis of $\rho^2 = 0.75$ was rejected for $t_{Ru,Ha}$. In study 1, the percentage of times the null hypothesis of $\rho^2 = 0$ was rejected across D replications was studied for both $t_{Ru,Ha}$ and F_{Ru}, regardless of the actual value of ρ^2. There were two reasons for testing against $\rho^2 = 0$ for all simulation models. Firstly, the author wanted to keep $t_{Ru,Ha}$ comparable with the F-test, which can only test against $\rho^2 = 0$.[1] Secondly, it is common practice to test against $\rho^2 = 0$, which makes the current situation more relevant for practice than Harel's (2009) study. Under $\rho^2 = 0$, the percentage of times the null hypothesis is rejected is the observed type-I error rate; under $\rho^2 = 0.18$ this is the observed power. Ideally, a pooled significance test for R^2 should have type-I error rates close to 5% when $\rho^2 = 0$, substantially higher observed power than CCA when $\rho^2 = 0.18$, and only slightly lower observed power than the original data under $\rho^2 = 0.18$. Finally, in study 2, the percentage of times the null hypothesis of $\Delta\rho^2 = 0$ was rejected across replications, was studied for methods F_{Ru}, F_{BR}, and F_{Rt}.

Results

Results simulation study 1

The results of study 1 are shown in Table 2. For the different pooled measures of R^2, both the means and the standard deviations (between brackets) across the D replications, are reported. Note that for 50% missingness, results for CCA are not displayed because often too few cases were left for the intended regression.

In general, it can be seen that differences between $\overline{R^2}$ and $\Re^2$ (first two columns) are small, but that R^2_F, (third column), is on average higher than the other two measures for $M = 6$ imputations, but lower for higher numbers of imputations. For F_{Ru} (fourth column) the type-I error rate is close to 5% for all percentages of missing data and all M. Harel's $t_{Ru,Ha}$ however (fifth column), produces substantially larger type-I error rates, and increase even more as the percentage of missingness increases, or M decreases. Even in the original data and in CCA, the Fisher z-test gives type-I error rates far off the theoretical 5%.

For $\rho^2 = 0.18$, $\overline{R^2}$ and $\Re^2$ (first two columns, lower half) increase as the percentage of missingness increases, and are higher than R^2 from the original data, but lower than R^2 from CCA. Again, differences

[1] A noncentral F-distribution may test against other values than 0. However, a noncentral F-test is not implemented in R's lm(), SPSS's "MIXED" or Regression procedure, or SAS's "reg".

Table 2. Means of $\Re^2$, $\bar{R}^2$, and R_F^2 (SDs between brackets), and rejection rate using F_{Ru} and $t_{Ru,Ha}$, of study 1. Results for R^2 of the original data and of CCA, and results of the Fisher z-tests of the original data and of CCA are displayed for comparison. When $\Delta\rho^2 = .180$, rejection rates represent power.

						Rejection rate	
ρ^2	% Missing data	M	$\Re^2$	$\bar{R}^2$	R_F^2	F_{Ru}	$t_{Ru,Ha}$
.00	Original		.030 (.024)	.030 (.024)	.030 (.024)	.053	.290
	6.25%	6	.039 (.029)	.039 (.029)	.040 (.033)	.058	.306
		25	.038 (.029)	.039 (.029)	.032 (.027)	.054	.314
		100	.038 (.029)	.039 (.029)	.031 (.026)	.055	.322
		250	.038 (.029)	.039 (.029)	.031 (.026)	.055	.321
		CCA	.040 (.033)	.040 (.033)	.040 (.033)	.055	.408
	12.5%	6	.048 (.033)	.049 (.033)	.057 (.045)	.056	.338
		25	.047 (.031)	.049 (.031)	.034 (.027)	.047	.354
		100	.047 (.031)	.049 (.031)	.032 (.025)	.045	.354
		250	.047 (.031)	.049 (.031)	.031 (.025)	.045	.361
		CCA	.051 (.039)	.051 (.039)	.051 (.039)	.038	.540
	25%	6	.071 (.045)	.075 (.045)	.097 (.074)	.051	.402
		25	.070 (.042)	.074 (.042)	.042 (.034)	.050	.446
		100	.070 (.041)	.074 (.041)	.034 (.027)	.048	.454
		250	.070 (.042)	.075 (.041)	.032 (.026)	.048	.459
		CCA	.094 (.074)	.094 (.074)	.094 (.074)	.052	.759
	50%	6	.091 (.061)	.095 (.061)	.119 (.093)	.040	.429
		25	.090 (.059)	.095 (.059)	.048 (.041)	.042	.492
		100	.090 (.058)	.095 (.059)	.035 (.029)	.045	.521
		250	.089 (.058)	.095 (.058)	.033 (.027)	.043	.524
.180	Original		.202 (.070)	.202 (.070)	.202 (.070)	.976	.998
	6.25%	6	.206 (.074)	.206 (.074)	.224 (.082)	.938	.994
		25	.205 (.073)	.205 (.073)	.188 (.067)	.946	.996
		100	.206 (.075)	.206 (.075)	.186 (.070)	.946	.996
		250	.206 (.074)	.206 (.073)	.185 (.069)	.946	.996
		CCA	.207 (.079)	.207 (.079)	.207 (.079)	.919	.996
	12.5%	6	.213 (.081)	.214 (.081)	.268 (.103)	.874	.984
		25	.211 (.079)	.212 (.079)	.180 (.072)	.896	.993
		100	.212 (.081)	.213 (.080)	.169 (.070)	.896	.992
		250	.213 (.080)	.214 (.079)	.168 (.068)	.896	.993
		CCA	.216 (.091)	.216 (.091)	.216 (.091)	.816	.990
	25%	6	.227 (.092)	.229 (.091)	.326 (.126)	.694	.954
		25	.226 (.090)	.228 (.088)	.173 (.077)	.770	.983
		100	.229 (.091)	.231 (.089)	.145 (.068)	.782	.987
		250	.228 (.091)	.230 (.089)	.139 (.066)	.786	.988
		CCA	.247 (.119)	.247 (.119)	.247 (.119)	.536	.984
	50%	6	.243 (.103)	.245 (.101)	.342 (.136)	.571	.912
		25	.240 (.101)	.242 (.098)	.171 (.083)	.658	.976
		100	.243 (.101)	.245 (.098)	.135 (.068)	.670	.984
		250	.242 (.100)	.245 (.098)	.128 (.065)	.676	.985

between $\overline{R^2}$ and $\Re^2$ are small, but R_F^2 (third column) is overestimated compared to the original data for $M = 6$, and underestimated for higher M. The power of F_{Ru} (fourth column, lower half) decreases as both the percentage of missingness and M increase, and stays above the power of CCA. The power of the Fisher z-test for both the original data and the multiply imputed data on the other hand (fifth column) is close to 100% across all design cells.

Results simulation study 2

Table 3 shows the results of simulation study 2. When $\Delta\rho^2 = 0$, both $\Delta\overline{R^2}$ and $\Delta\Re^2$ (first two columns, upper half) increase and deviate more from ΔR^2 of the original data as the percentage of missingness increases. Here, $\Delta\Re^2$ increases somewhat more than $\Delta\overline{R^2}$, but again differences are small. The number of imputations does not influence $\Delta\overline{R^2}$ and $\Delta\Re^2$ much. On the other hand, ΔR_F^2 (third columns) is substantially lower than $\Delta\overline{R^2}$, $\Delta\Re^2$, and ΔR^2 of CCA, and is even negative occasionally, especially for low M and high percentages of missingness. For F_{Ru} (fourth column), the type-I error rate is close to 5% across all M and all percentages of missingness. However, the type-I error rates of F_{Rt} and F_{BR} (last two columns) heavily depend on both M and the percentage of missingness. For F_{Rt} it varies from .000 ($M = 6$, 50% missingness) to .378 ($M = 250$, 50% missingness), while for F_{BR} it varies from .015 ($M = 6$, 50% missingness) to .094 ($M = 250$, 25% missingness).

For $\Delta\rho^2 = 0.14$, both $\Delta\overline{R^2}$ and $\Delta\Re^2$ (first two columns, lower half) are stable across different M, but increase as the percentage of missingness increases. Of these two measures, $\Delta\Re^2$ increases most. Furthermore, when the percentage of missingness increases, $\Delta\overline{R^2}$ and $\Delta\Re^2$ remain closer to ΔR^2 of the original data than of CCA. In general, differences between $\Delta\overline{R^2}$ and $\Delta\Re^2$ are small. Again, ΔR_F^2 (third column) is

Table 3. Means of $\Delta\Re^2$, $\Delta\bar{R}^2$, and ΔR_F^2 (SDs between brackets), and rejection rate using F_{Ru}, F_{Rt}, and F_{BR}, of study 2. Results for of the original data and of CCA are displayed for comparison. When $\Delta\rho^2 = .141$, rejection rates represent power.

						Rejection rate		
$\Delta\rho^2$	% Missing data	M	$\Delta\Re^2$	$\Delta\bar{R}^2$	ΔR_F^2	F_{Ru}	F_{Rt}	F_{BR}
.00	Original		.016 (.016)	.016 (.016)	.016 (.016)	.048	.048	.048
	6.25%	6	.022 (.020)	.022 (.020)	−.177 (.133)	.051	.005	.034
		25	.022 (.020)	.022 (.020)	.007 (.020)	.049	.065	.052
		100	.022 (.020)	.022 (.020)	.013 (.019)	.050	.085	.058
		250	.022 (.020)	.022 (.020)	.014 (.019)	.049	.087	.060
		CCA	.022 (.022)	.022 (.022)	.022 (.022)	.052	.052	.052
	12.5%	6	.028 (.024)	.028 (.024)	−.255 (.121)	.052	.004	.024
		25	.028 (.023)	.028 (.023)	−.009 (.026)	.048	.070	.050
		100	.028 (.023)	.028 (.023)	.008 (.021)	.048	.116	.062
		250	.028 (.022)	.028 (.022)	.010 (.020)	.049	.122	.065
		CCA	.028 (.027)	.028 (.027)	.028 (.027)	.044	.044	.044
	25%	6	.045 (.034)	.044 (.034)	−.247 (.101)	.054	.002	.022
		25	.045 (.032)	.044 (.031)	−.033 (.039)	.056	.065	.041
		100	.045 (.032)	.044 (.031)	−.001 (.026)	.054	.209	.078
		250	.045 (.032)	.044 (.031)	.005 (.025)	.058	.246	.094
		CCA	.053 (.052)	.053 (.052)	.053 (.052)	.052	.052	.052
	50%	6	.058 (.046)	.057 (.045)	−.225 (.111)	.048	.000	.015
		25	.059 (.043)	.057 (.042)	−.040 (.048)	.045	.061	.038
		100	.059 (.043)	.057 (.042)	.043 (.028)	.048	.286	.076
		250	.059 (.043)	.057 (.044)	.003 (.027)	.044	.383	.086
.141	Original		.152 (.058)	.152 (.058)	.152 (.058)	.976	.976	.976
	6.25%	6	.155 (.064)	.154 (.063)	−.020 (.148)	.952	.452	.897
		25	.156 (.063)	.155 (.063)	.134 (.062)	.959	.968	.960
		100	.155 (.063)	.155 (.063)	.137 (.060)	.960	.971	.965
		250	.156 (.063)	.155 (.063)	.137 (.060)	.961	.973	.964
		CCA	.154 (.068)	.154 (.068)	.154 (.068)	.929	.929	.929
	12.5%	6	.160 (.068)	.159 (.068)	−.094 (.146)	.900	.169	.706
		25	.161 (.067)	.159 (.066)	.115 (.066)	.929	.949	.927
		100	.161 (.067)	.160 (.066)	.122 (.061)	.931	.970	.944
		250	.161 (.066)	.159 (.066)	.123 (.060)	.934	.971	.953
		CCA	.158 (.076)	.158 (.076)	.158 (.076)	.827	.827	.827
	25%	6	.170 (.078)	.168 (.077)	−.092 (.143)	.744	.038	.385
		25	.171 (.076)	.168 (.075)	.079 (.079)	.794	.840	.778
		100	.171 (.075)	.168 (.074)	.093 (.063)	.806	.949	.880
		250	.171 (.075)	.168 (.074)	.094 (.061)	.805	.964	.896
		CCA	.175 (.104)	.175 (.104)	.175 (.104)	.570	.570	.570
	50%	6	.186 (.089)	.182 (.087)	−.067 (.154)	.646	.035	.312
		25	.186 (.086)	.182 (.084)	.072 (.087)	.729	.773	.696
		100	.186 (.086)	.182 (.084)	.086 (.067)	.732	.953	.844
		250	.186 (.085)	.182 (.083)	.087 (.066)	.740	.972	.865

substantially lower than $\Delta\overline{R^2}$, $\Delta\Re^2$, and ΔR^2 of both the original data and CCA, and is heavily influenced by both M and the percentage of missingness.

Finally, under $\Delta\rho^2 = 0.141$, the observed power increases for all tests as either the percentage of missingness decreases or M increases (last three columns). For F_{Rt} and F_{BR}, the power increases more with increasing M than for F_{Ru}, with the power of these tests being lower than F_{Ru} for $M = 6$, but higher than F_{Ru} for $M > 6$.

Empirical data example

In this section, the three measures of $(\Delta)R^2$ and the significance tests of $(\Delta)R^2$ are applied to an empirical data example from a study about obesity in young children (Camfferman, Van der Veek, & Mesman, 2017). The original data set has $N = 101$ children. Some of the variables of interest are Body Mass Index (BMI) of both the father and the mother, restrictions (the parents' control of their child's eating behavior by restriction of the type or amount of food), pressure (attempts by the parents to increase children's food consumption), and approaching eating habits by the child (the extent to which a child is tended to approach food rather than to avoid it). Restrictions and pressure by the parents are measured using two subscales of the Children's Feeding Questionnaire (CFQ; Birch, Fisher, Grimm-Thomas, Markey, Sawyer, & Johnson, 2001); approaching eating habits are measured by a subscale of the Children's Eating Behavior Questionnaire (CEBQ; Wardle, Guthrie, Sanderson, & Rapoport, 2001). The data used in this example are a random subsample ($n = 55$) of the complete data set, meant for illustrative purposes only. In the reduced data set, BMI of the father has 11 missing values, BMI of the mother is completely observed, restriction has 6 missing values, pressure has 4 missing values, and approaching eating habits has 1

Table 4. Pooled regression analysis of the multiply imputed data from Camfferman et al. (2017) about obesity.

	Model 1			Model 2		
Effect	*B*	*SE*	*p*	*B*	*SE*	*p*
Intercept	−0.07	0.15	0.64	−0.03	0.143	0.82
BMI Mother	−0.09	0.18	0.61	0.04	0.170	0.84
BMI Father	0.05	0.18	0.76	0.06	0.158	0.72
CFQ pressure to eat				−0.33	0.166	0.06
CFQ restriction				0.38	0.153	0.02*
$\Re^2$		0.014			0.199	
$\bar{R}^2$		0.017			0.202	
R_F^2		0.005			0.157	
F_{Ru}, *p*		$F(2, 48) = 0.13$, $p = 0.88$			$F(4, 47) = 2.17$, $p = 0.09$	
$t_{Ru,Ha}$, *p*		$Z = 1.02$, $p = 0.31$			$Z = 3.58$, $p < 0.001$*	
$F_{Ru}(\Delta R^2)$, *p*					$F(2, 46) = 4.09$, $p = 0.02$*	
F_{Rt}, *p*					$F(2, 44) = 5.11$, $p = 0.01$*	
F_{BR}, *p*					$F(2, 33) = 3.86$, $p = 0.03$*	

missing value. Missing values were imputed $M = 100$ times. In the intended analysis, the score of approaching eating habits is predicted by BMI of both the father and mother (block 1), and by pressure and restrictions (block 2). The complete data set and its analysis are described in Camfferman et al. (2017). The results of the combined analysis using the three measures for $(\Delta)R^2$ and the significance tests are shown in Table 4. The pooled coefficients and their *t*-tests were computed using Equations (3)–(8).

The table shows that differences between $\Delta\overline{R^2}$ and $\Re^2$ are small, and that R_F^2 is substantially lower than $\Delta\overline{R^2}$ and $\Re^2$. The *p*-values of $t_{Ru,\ Ha}$ are not any way near those of F_{Ru}. $F_{Ru}(\Delta R^2)$ has a *p*-value higher than that of F_{Rt}, but lower than that of F_{BR}.

Discussion

Significance Tests for R^2 and ΔR^2

In this study different procedures for pooling statistical tests for $(\Delta)R^2$ in multiply imputed data were compared. In advance, it was argued that methods $t_{Ru,\ Ha}$, F_{Rt}, and F_{BR} would give biased type-I error rates, despite earlier simulations (Chaurasia & Harel, 2014; Harel, 2009), which showed that these methods performed well. New simulations supported the theoretical objections against these statistics. It should be noted that the current two studies were not exact replications of Harel (2009) and Chaurasia and Harel (2014).

However, one of the reviewers suggested to exactly replicate both studies and to provide the results as supplemental material. Both studies were partly replicated (see, supplemental material). The replicated results of Harel's (2009) study were largely in accordance with his findings. Apparently, for the specific situations studied by Harel (2009) his method works better than for the situations studied in the current paper. To some extent this makes sense. When ρ^2 is as high as 0.75 (like in Harel's study) and $\Re^2$ is tested against this value, the problem of the Fisher z transformation of $\sqrt{R^2}$ having a lower bound of 0 is not much of an issue because it will hardly ever come close to 0. It remains unclear how the exclusion of the model degrees of freedom of the regression model in the Fisher z test may have had such a small influence on the results of Harel's (2009) study and the replication study. However, given that in practice such high values of ρ^2 rarely occur, and that normally the null hypothesis of $\rho^2 = 0$ is tested, both this question and these findings are largely irrelevant.

The results by Chaurasia and Harel (2014) could not (exactly) be replicated: The type-I error rates were on average 2.2% higher for F_{BR} and 3.7% higher for F_{Rt} than found by Chaurasia and Harel. Additionally, the observed power rates for F_{Rt} were on average 5.8% higher while observed power rates for F_{BR} were similar to the ones found by Chaurasia and Harel. Because of this failure to replicate, the author asked Chaurasia for the programming code of his methods. Additionally, whenever specific details of Chaurasia and Harel's study were open to interpretation, the author asked for clarification. Comparison of both codes and clarification of these details revealed no differences in both procedures that could explain the differences in findings. Thus, the exact cause of failure to replicate remains unclear.

However, even if programming errors had been found in this replication study explaining the differences, then still the theoretical objections to both methods put forward in this article, would have remained. Because of a lack of theoretical justification, any promising result found for both methods might as well have been only context-specific. Thus, to rule out this possibility, more research under more various circumstances would have been necessary. However, given the availability of an alternative of which the

statistical properties are largely known (F_{Ru}; Rubin, 1987), further investigation of F_{BR} and F_{Rt} seems rather unnecessary anyway.

Estimates for R^2

Along with the proposed pooling techniques for the significance tests for R^2 and ΔR^2 Harel (2009) proposed a pooling technique for R^2, denoted $\Re^2$. This measure was compared with two newly proposed alternatives: $\overline{R^2}$ and R_F^2. It turned out that $\Re^2$ and $\overline{R^2}$ produced very similar results. Unfortunately, the newly proposed R_F^2 did not perform as well as was hoped for. This pooled version underestimated ρ^2, and was lower on average than lower benchmark CCA as well. In study 2, for some situations ΔR_F^2 was even negative. This probably lies in the fact that ν^{Rt} in Equation (14) includes the relative increase in variance, r (Equation (12)). This value of r is not constant across two (or more) competing regression models as newly added variables come with new missing values, consequently changing r. Especially for weak relations between the newly added variables and the outcome variable, and a high percentage of missingness, this could result in negative ΔR_F^2's. For practical purposes a negative ΔR_F^2 is not a problem because it may be interpreted as no relationship between the newly added predictors and the outcome variable. Nevertheless, the severe underestimation of R^2 using the R_F^2 statistic in general, is problematic.

Implications and conclusions

The calculation of $\Re^2$ (Harel, 2009) is more complicated than that of $\overline{R^2}$ and relies on an incorrect justification. Although $\Re^2$ is on average close to R^2 of the original data, its results hardly differ from the simpler $\overline{R^2}$, which neither has a correct nor an incorrect justification. On the other hand, R_F^2 is on average substantially lower than R^2 of the original data, and even lower than that of CCA. The above considerations lead to the conclusion that $\overline{R^2}$ is the preferred pooled version of R^2 in multiple imputation.

As for the combination rules for the significance tests for $(\Delta)R^2$, although earlier studies show that $t_{Ru,Ha}$ (Harel, 2009), F_{Rt}, and F_{BR} (Chaurasia & Harel, 2014) give satisfactory type-I error rates, the current studies indicate otherwise. Although Chaurasia and Harel admit that their method is ad hoc, they justify its use by its ease of computation (2014, p. 433) and its higher observed power (p. 439), compared to F_{Ru}. However, the former argument only holds when F_{Ru} has to be calculated manually. Nowadays, F_{Ru} is available in both SPSS 25.0 using the macro by Van Ginkel (2010), in SAS 9.4, and in R. In Appendix A, the procedure for calculating F_{Ru} for R^2 and ΔR^2 in SPSS is explained using the reduced data set from Camfferman et al. (2017). Appendices B and C show the procedure for SAS an R, respectively.

As for the argument of better power, in study 2 of the current article, it was found that when $M > 6$, F_{Rt} and F_{BR} indeed correctly reject the null hypothesis more frequently than F_{Ru}. One interpretation is that these tests have more power than F_{Ru}. However, given that under $\Delta\rho^2 = 0$ these tests have deflated type-I error rates when $M = 6$ and inflated type-I error rates when $M > 6$, a more logical interpretation is that for low M these F-tests are biased downwards and that for high M biased upwards, that is, for the situations studied here (assuming that no programming errors were made in the current study). The bottom line is that the argument of increased power only holds when the type-I error rates of these tests are not heavily affected by factors that should not affect them, such as the number of imputations and the percentage of missingness.

To conclude, $\overline{R^2}$ is the preferred point estimate for R^2, and F_{Ru} is the preferred method for testing $(\Delta)R^2$ for significance. Using $\Re^2$ as a point estimate for R^2 will give estimates that hardly differ from the recommended $\overline{R^2}$, but considering its incorrect justification and the fact that is more difficult to compute, there does not seem to be any reason to prefer $\Re^2$ over $\overline{R^2}$. Finally, based on the current results, R_F^2 and the statistics by Harel (2009) and Chaurasia and Harel (2014) must be avoided.

Article information

Conflict of interest disclosures: The author signed a form for disclosure of potential conflicts of interest. The author did not report any financial or other conflicts of interest in relation to the work described.

Ethical principles: The author affirms having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The author would like to thank Dr Ashok Chaurasia for sharing his programming code, regardless of

possible discrepancies in findings between his study and the current one.

The author would also like to thank Dr. Alexander Robitzsch for helping him update the R code in Appendix C.

The ideas and opinions expressed herein are those of the author alone, and endorsement by the author's institution is not intended and should not be inferred.

References

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*, 949–955. doi:10.1093/biomet/86.4.948

Birch, L. L., Fisher, J. O., Grimm-Thomas, K., Markey, C. N., Sawyer, R., & Johnson, S. L. (2001). Confirmatory factor analysis of the Child Feeding Questionnaire: A measure of parental attitudes, beliefs and practices about child feeding and obesity proneness. *Appetite*, *36*(3), 201–210. doi:10.1006/appe.2001.0398

Camfferman, R., Van der Veek, S. M. C., & Mesman, J. (2017). *Lack of received sensitivity during mealtime is related to overweight in early childhood*. Manuscript submitted for publication.

Chatterjee, S., & Hadi, A. S. (1999). *Regression analysis by example* (3rd ed.). Hoboken, NJ: Wiley.

Chaurasia, A., & Harel, O. (2014). Partial *F*-tests with multiply imputed data in the linear regression framework via coefficient of determination. *Statistics in Medicine*, *34*(3), 432–443. doi:10.1002/sim.6334

Fox, J. (2016). *Applied linear regression and generalized linear models* (3rd ed.). Thousand Oaks, CA: Sage.

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

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

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

Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment based statistics and an *F* reference distribution. *Journal of the American Statistical Association*, *86*(416), 1065–1073. doi:10.2307/2290525

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

Liu, Y., & Enders, C. K. (2017). Evaluation of multi-parameter test statistics for multiple imputation. *Multivariate Behavioral Research*, *52*(3), 371–390. doi:10.1080/00273171.2017.1298432

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

R Core Team (2017). *R: A language and environment for statistical computing.*, Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

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

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

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

SAS Institute, Inc. (2013). *SAS® 9.4 [Computer software]*. Cary, NC: Author.

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

SPSS, Inc. (2017). *SPSS 25.0 for Windows [Computer software]*. Chicago, IL: Author.

StataCorp (2015). *Stata Statistical Software: Release 14 [Computer software]*. College Station, TX: Author.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson

Van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: 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*(12), 1049–1064. doi: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. (2010). *MI-MUL2.SPS [Computer code]*. Retrieved from https://www.universiteitleiden.nl/en/staffmembers/joost-van-ginkel#tab-1

Van Ginkel, J. R. (2016). *MI-MUL2.pdf [Software manual]*. Retrieved from https://www.universiteitleiden.nl/en/staffmembers/joost-van-ginkel#tab-1

Van Ginkel, J. R., & Kroonenberg, P. M. (2014). Analysis of variance of multiply imputed data. *Multivariate Behavioral Research*, *49*(1), 78–91. doi: 10.1080/00273171.2013.855890

Wardle, J., Guthrie, C. A., Sanderson, S., & Rapoport, L. (2001). Development of the Children's Eating Behaviour Questionnaire. *Journal of Child Psychology and Psychiatry*, *42*(7), 963–970. doi:10.1111/1469-7610.00792

Yuan, Y. C. (2011). Multiple Imputation using SAS Software. *Journal of Statistical Software*, *45*, 1–25.

Appendix A

In this appendix, the complete procedure for obtaining F_{Ru} for the complete model and for ΔR^2 using the SPSS macro by Van Ginkel (2010), is described. The procedure is demonstrated using the example data set (Camfferman et al., 2017). The interested reader can download this subset from the author's personal webpage https://www.universiteitleiden.nl/en/staffmembers/joost-van-ginkel#tab-1 and reproduce the results from Table 4. From now on it is assumed that the data set is stored in C:\MyData\ CamffermanEtAl.sav.

First we open the data set in SPSS:

GET FILE= 'C:\MyData\CamffermanEtAl.sav'.

In order for the multiply imputed data set to be analyzed as a multiply imputed data set by SPSS, a split file must be carried out, using the variable Imputation_ as a split variable (in the menu: Data, Split File):

```
SORT CASES BY Imputation_.
SPLIT FILE LAYERED BY Imputation_.
```

The SPSS macro (Van Ginkel, 2010) reads SPSS data files in which the results of the statistical analyses are stored. To create such a data file, we use the OMS command (in the menu: Utilities, OMS Control Panel):

```
OMS
  /SELECT TABLES
  /IF COMMANDS = ['Mixed'] SUBTYPES =
  ['Covariance Matrix'
'Parameter Estimates']
  /DESTINATION FORMAT = SAV
NUMBERED = TableNumber_
  OUTFILE='C:\MyData\Parameters1.sav' VIEWER = YES.
```

What is particularly needed are the regression coefficients and the covariance matrix of the regression coefficients.

This can be found in the line: /IF COMMANDS = ['Mixed'] SUBTYPES = ['Covariance Matrix' 'Parameter Estimates'].

More extensive examples of how to use the OMS command are given in the manual of the SPSS macro (Van Ginkel, 2016).

Once the OMS option has been carried out, the regression analysis may be carried out on the multiply imputed data sets. However, this regression must be carried out in Mixed models (In the menu: Analyze, Mixed Models, Linear) as Van Ginkel's (2010) macro can only process the regression coefficients and covariance matrices when they are stored in the format that is generated by Mixed Models.

```
MIXED ZCEBQ_APPROACH WITH ZM_BMI_combi
ZF_BMI_combi
  /CRITERIA = CIN(95) MXITER(100) MXSTEP(10)
SCORING(1) SINGULAR(0.000000000001)
HCONVERGE(0, ABSOLUTE) LCONVERGE(0,
ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
  /FIXED = ZM_BMI_combi ZF_BMI_combi | SSTYPE(3)
  /METHOD = REML
  /PRINT = COVB SOLUTION.
```

More examples of how to carry out analyses using Mixed Models may be found in the manual as well. Next, the OMS command is ended by the following statement:

```
OMSEND.
```

After ending the OMS command, the macro has to be called, and arguments must be specified. We assume the macro MI-MUL2.sps is saved to the directory C:\MyData\. The syntax code for running the macro to get pooled regression coefficients of Model 1 (Table 4) and their significance tests, plus the pooled *F*-test of the complete model, is

```
INCLUDE 'C:\MyData\MI-mul2.sps'.
RULESMIMUL FILE = 'C:\MyData\Parameters1.sav'
/ESTIMATE = Estimate
/COV = Intercept to ZF_BMI_combi
/LEVELSIND =1,1
/DF = df
/M = 100.
```

Here, the INCLUDE line calls the macro. The RULESMIMUL FILE line specifies the data set in which the results of the analyses are stored. The /ESTIMATE line specifies the variable in the file Parameters1.sav that contain the regression coefficients. The /COV line specifies the variables that contain the covariance matrices of the regression coefficients. In the /LEVELSIND command, the number of predictor variables taking part in the analysis, and their number of levels are specified. In the specific example, there are two continuous predictors (ZF_BMI_combi and ZF_BMI_combi), all with one level. Continuous predictors are specified as variables with one level (Van Ginkel, 2016, p. 15). The /DF line specifies the variable that contains the number of degrees of freedom. This is needed for the calculations of ν^{Rt} (Equation (14)). The last line (/M) specifies the number of imputed data sets.

To carry out the analysis of Model 2, the same procedure must be followed as for Model 1. First, the analysis must be "carried out" on each imputed data set separately, and the results must be stored in a data file, using the OMS option:

```
GET FILE= 'C:\MyData\CamffermanEtAl.sav' .
SORT CASES BY Imputation_.
SPLIT FILE LAYERED BY Imputation_.
OMS
  /SELECT TABLES
/IF COMMANDS = ['Mixed'] SUBTYPES = ['Covariance
Matrix'
'Parameter Estimates']
  /DESTINATION FORMAT = SAV NUMBERED =
TableNumber_
  OUTFILE=' C:\MyData\Parameters2.sav' VIEWER = YES.
MIXED ZCEBQ_APPROACH WITH ZM_BMI_combi
ZF_BMI_combi ZCFQpressuretoeat ZCFQ_Arestriction
  /CRITERIA = CIN(95) MXITER(100) MXSTEP(10)
SCORING(1) SINGULAR(0.000000000001) HCONVERGE(0,
  ABSOLUTE) LCONVERGE(0, ABSOLUTE)
PCONVERGE(0.000001, ABSOLUTE)
  /FIXED = ZM_BMI_combi ZF_BMI_combi ZCFQ
  pressuretoeat
ZCFQ_Arestriction | SSTYPE(3)
  /METHOD = REML
  /PRINT = COVB SOLUTION.
OMSEND.
```

To compute the F_{Ru} for ΔR^2 of Model 2 against Model 1 (Table 4) the following syntax code has to be used:

```
INCLUDE 'C:\MyData\MI-mul2.sps'.
RULESMIMUL FILE = 'C:\MyData\Parameters2.sav'
/ESTIMATE = Estimate
/COV = Intercept to ZCFQ_Arestriction
/LEVELSIND =1,1,2
/DF = df
/M = 100.
```

The *F*-test at the top of the output is the F_{Ru} for the complete Model 2 (Table 4). In the /LEVELSIND command, the latter two variables have been joined into one "variable" with two levels. The last pooled *F*-test in the resulting output is the F_{Ru} for ΔR^2 (Table 4, the fifth row of bottom panel).

Appendix B

Below the complete procedure for obtaining F_{Ru} for both R^2 and ΔR^2 in SAS 9.4 is described, using the example data set (Camfferman et al., 2017). In order for SAS to carry out the pooled tests, the SPSS data set has to be converted to SAS format in SPSS first.

Multiply imputed data sets in SAS do not include the incomplete data set without imputed values on top of the data file, while in SPSS this incomplete file is included, indicated by an imputation number of 0 (Variable Imputation_ in the example data set).

This incomplete data set must be deleted first before the data can be saved in SAS format:

```
GET FILE='C:\MyData\CamffermanEtAl.sav '.
FILTER OFF.
USE ALL.
SELECT IF (Imputation_ ~= 0).
EXECUTE.
```

Once this has been done, the data can be saved in SAS format (Menu: Save As…, and set "Save as type:" to SAS v9 + Windows (*.sas7bdat)):

```
SAVE TRANSLATE OUTFILE='C:\MyData\Camfferman
EtAl.sas7bdat'
  /TYPE = SAS
  /VERSION =9
  /PLATFORM = UNIX
  /ENCODING='UTF8'
  /MAP
  /REPLACE.
```

Next, we open SAS and read the data file. In SAS the variable containing the imputation number is called _Imputation_. Thus, the variable Imputation_ in the data set must be renamed first, and the data must be saved to a new file named CamffermanEtAl2.sas7bdat:

```
data
"C:\MyData\CamffermanEtAl2.sas7bdat";set"
  C:\MyData\CamffermanEtAl.sas7bdat";
  Rename Imputation_=_Imputation_;
run;
```

Next, the regression analysis of the smaller model is carried out for each imputed data set separately:

```
proc
reg data="C:\MyData\CamffermanEtAl2.sas
7bdat"outest="
C:\MyData\parametersSPSS.sas7bdat" covout;
model ZCEBQ_APPROACH=ZM_BMI_combi
ZF_BMI_combi;
by _Imputation_;
run;
```

The results of the smaller model can be pooled by means of:

```
proc mianalyze data=
"C:\MyData\parametersSPSS.sas7bdat" edf=52;
modeleffects Intercept ZM_BMI_combi ZF_BMI_
combi;
test ZM_BMI_combi=ZF_BMI_combi=0/mult;
run;
```

Next, analysis of the larger model is carried out for each imputed data set separately:

```
proc reg data="C:\MyData\CamffermanEtAl2.sas7
bdat"
outest="C:\MyData\parametersSPSS.sas7b
dat" covout;
model ZCEBQ_APPROACH=ZM_BMI_combi
ZF_BMI_combi ZCFQpressuretoeat ZCFQ_
Arestriction;
by _Imputation_;
run;
```

The F_{Ru} for R^2 and for ΔR^2 can be calculated by means of:

```
proc mianalyze data="C:\MyData\
parametersSPSS. sas7bdat" edf=50;
modeleffects Intercept ZM_BMI_combi
ZF_BMI_combi ZCFQpressuretoeat
ZCFQ_Arestriction;
test ZM_BMI_combi=ZF_BMI_combi=ZCFQ
pressuretoeat=
ZCFQ_Arestriction=0/mult;
test ZCFQpressuretoeat=ZCFQ_Arestriction
=0/mult;
run;
```

Finally, it should be noted that for this example the already imputed data file was used rather than multiply imputing the incomplete data set in SAS, to ensure that the results would be identical to the results in Table 4.

Appendix C

In R, F_{Ru} can be calculated using the miceadds package (Robitzsch, Grund, & Henke, 2017). After installation, this package is called:

```
library(miceadds)
```

Furthermore, this package is only applicable to a data set that has been multiply imputed within R using the mice package (Van Buuren & Groothuis-Oudshoorn, 2011). Therefore, we cannot directly use the multiply imputed data set as in the SPSS and SAS examples. Instead, the incomplete data set from Camfferman et al. is read in R as a plain text file, named CamffermanEtAl.dat (missings are indicated by -999 in the file), available from the author's personal page:

```
CamffermanData <-
read.csv("http://leidenuniv.nl/fsw/
Psychologie/CamffermanEtAl.dat",
header=TRUE, sep = " ")
```

Next, the data are multiply imputed using the mice package:

```
library(mice)
CamffermanDataImputed <- mice(CamffermanData,
m=100, maxit =100, seed =361)
```

In order to obtain the F_{Ru} for testing all parameters of the smaller model simultaneously, we create a function in which one can specify the multiply imputed data, regression model and the set of parameters that one wants to test for significance simultaneously. More specifically, this function requires four arguments: (1) `data set`: a multiply imputed data set obtained from the mice procedure (here: `CamffermanDataImputed`), (2) `response`: the name of the response variable (in the data set from Camfferman et al. this variable is called `ZCEBQ_APPROACH`), (3) `predictors`: a string vector containing the names of the predictors in the model (the names of the predictors depend on the specific model that is evaluated), and (4) `testpredictors`: a string vector containing the names of a subset of predictors inthe model, which are tested for significance. The code for the specific function is:*

```
library(mitools)
Fru <- function(dataset, response,
predictors, testpredictors){
#Creating a string of the specific linear

model
eq <- paste(response, " ~ ",
paste(predictors, collapse = " + ")
)
#Computing the pooled model using the pool
function in mice
model <- with(dataset, lm(as.formula(eq)))
pooled model <- pool(model)

#Creating arrays containing the regression
coefficients and covariance matrices of each
# imputed dataset, denoted qhat and u,
#respectively.
qhat <- MIextract (model$analyses,
fun=coef)
  u <- MIextract (model$analyses,
  fun=vcov)

#Creating a vector containing the parameter
names of the model.
pars <- names(qhat[[1]])

#creating a design matrix indicating which
of the
#parameters inqhat are tested simultan-
eously. The
#create.designMatrices.waldtest function
facilitates
#the creation of the design matrix.
Since the
#miceadds manual (Robitzsch, Grund, &
Henke, 2017),
#pp. 103-107) gives some clear examples
of this
#function, the next lines are not fur-
ther explained.
design <- create.designMatrices.waldtest
(pars = pars,
k = length(testpredictors))
Cdes <- design$Cdes
rdes <- design$rdes
ii <- 0
for (predictor in testpredictors) {
ii <- ii +1
Cdes[ii, predictor] <- 1
}
#The MIwaldtest function in the
miceadds package
#calculates a pooled F value testing the
parameters in vector testparameters¤
#for significance
Wald <- MIwaldtest(qhat, u, Cdes, rdes)
summary(Wald)
}
```

Once the above code has been run, F_{Ru} testing all parameters of Model 1 simultaneously can be obtained by means of:

```
response <- "ZCEBQ_APPROACH"
predictors <- c("ZM_BMI_combi",
"ZF_BMI_combi")
result1 <- Fru(
CamffermanDataImputed,
response,
predictors,
predictors
)
```

Note that the results for this F_{Ru} are not identical to the results of the SPSS and SAS examples since the original reduced multiply imputed data set by Camfferman et al. (2017) could not be read by the `miceadds` package, and the incomplete data set had to be re-imputed using the `mice` package instead.

Next, R^2 of Model 2 is tested. This is done in a similar way as for the smaller model:

```
predictors <- c("ZM_BMI_combi",
"ZF_BMI_combi",
"ZCFQpressuretoeat","ZCFQ_Arestriction")
result2 <- Fru(
CamffermanDataImputed,
response,
predictors,
predictors
)
```

Finally, the F_{Ru} for testing ΔR^2 for significance is obtained using:

```
testpredictors <- c("ZCFQpressuretoeat",
"ZCFQ_Arestriction")
result3 <- Fru(
CamffermanDataImputed,
response,
predictors,
testpredictors
)
```

*Disclaimer: shortly after acceptance of this article the `mice` package was updated such that the code of the originally accepted draft did not work anymore. Fortunately the code in this appendix could be changed before appearing online. However, because of the rapid developments in `mice` and in R in general, the author cannot guarantee that the code provided will still work in future versions of R.