¦ 2016 ■ Vol. 12 ■ no. 1

DOI: 10.20982/tqmp.12.1.p057

Impact on Cronbach's α of simple treatment methods for missing data

Sébastien Béland [a,✉], François Pichette [b] & Shahab Jolani [c]

[a] Département d'administration et fondements de l'éducation, faculté des sciences de l'éducation, Université de Montréal, Canada
[b] TÉLUQ-Université du Québec, Canada
[c] School of Health Professions Education, FHML, Maastricht University, Maastricht, The Netherlands

Abstract ■ The scientific treatment of missing data has been the subject of research for nearly a century. Strangely, interest in missing data is quite new in the fields of educational science and psychology (Peugh & Enders, 2004; Schafer & Graham, 2002). It is now important to better understand how various common methods for dealing with missing data can affect widely-used psychometric coefficients. The purpose of this study is to compare the impact of ten common fill-in methods on Cronbach's α (Cronbach, 1951). We use simulation studies to investigate the behavior of α in various situations. Our results show that multiple imputation is the most effective method. Furthermore, simple imputation methods like Winer imputation, item mean, and total mean are interesting alternatives for specific situations. These methods can be easily used by non-statisticians such as teachers and school psychologists.

Keywords ■ Cronbach's α coefficient, missing data, missing completely at random, missing at random, simulation study, R software

✉ sebastien.beland@umontreal.ca

Introduction

Have you smoked marijuana during the last month? Do you know if your partner is HIV positive? Do you believe God created the world? Do you think homosexuality is wrong? Do you think President Obama is a socialist? These kinds of questions could be of interest in psychology, political science, criminology, and sexology. However, they can make some examinees so uncomfortable that they would be unwilling to answer them.

In the context of educational testing, the problem of missing data is also an important issue. Explanations for nonresponse can include fatigue (e.g., absence of answers for the last items of a long questionnaire), distraction (e.g., a student forgot to answer the back side of a copy, which was left blank), and item difficulty (the student skipped or ignored some items).

The scientific treatment of missing data has been the subject of research for nearly a century. Strangely, interest in missing data is quite new in the fields of educational science and psychology (Peugh & Enders, 2004; Schafer & Graham, 2002). Many reasons can explain this apparent disinterest. One reason can be related to lack of statistical training (Giguère, Hélie, & Cousineau, 2004; Lazaraton, Riggenbach, & Ediger, 1987; Schmidtke, Spino, & Lavolette, 2012; Yang, 2010). Sharpe (2013) invoked resistance to statistical innovation to describe that phenomenon. Another argument refers to the fact that providing information about participant exclusion based on their missing data can raise too many questions and decrease the chance of being accepted in a peer-reviewed journal (Pichette, Béland, Jolani, & Leśniewska, 2015). There also remains the issue of the effect these nonresponses can have on the psychometric properties of tests.

Effect of missing data on the psychometric properties of tests

According to Huisman (2000), "in the presence of item nonresponse, (...) the measurement task is much harder and the quality of measurement can be seriously affected (p. 332)". For example, Finch (2008) and Huisman and Molenaar (2001) have shown that missing data can create problems when estimating parameters from item response models. Similar conclusions emerge from Rose, Von Davier, and Xu (2010), who used the PISA 2006 data set to show that missing data do have an impact on the estimation of multidimensional item response models.

Other studies show the effectiveness of various methods. For example, Bernaards and Sijtsma (1999) demonstrated that the EM algorithm was the most effective treatment of missing data if one wishes to use factor analysis. Huisman (2000) used simple methods to illustrate that the amount and type of missing data, as well as the characteristics of the analysed matrix (sample size and number of items) have an effect on Cronbach's α. He also demonstrated that procedures which involve relationships between items tend to perform best. In another study, (Sijtsma & Van der Ark, 2003) showed that response-function methods are superior to Person mean, Two-way imputation, and Mean response-function in recovering several statistical properties of the original complete data from incomplete data sets.

Finally, Pichette et al. (2015) studied Cronbach's α and discovered that replacing a nonresponse by the mean for that item is a better alternative than simple methods like leaving the square empty, or filling in with zero, as for an incorrect answer. However, that study had certain limitations, such as not including advanced methods like multiple imputation.

The treatment of missing data

The previous section argued that missing data can affect the psychometric properties of tests. The fact that missing data are a test validity issue underscores the need to investigate this problem. At this moment, two major elements must be discussed: the type of missing data, and possible treatments methods.

Missing data mechanisms

Rubin (1976) classified missing data into three broad categories that reflect their most probable cause: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The MCAR mechanism is when missingness happens totally by chance, for example when a participant forgets to answer the back side of a questionnaire. MAR occurs when the cause of missingness does not depend on the missing data themselves, but can be explained by the observed data. Consider a group of foreign students who might not answer specific historical questions, for example, because they are unfamiliar with the history of their country of residence. In such a case, the missing items can be explained by factors related to the students' background. The last type of missing data mechanism is MNAR, when the cause of missingness depends on the missing data themselves. For example, some students might not answer questions regarding sexual orientation or criminal activities, due to their sensitive nature.

It is worth noting that deletion methods are only valid (i.e., they yield unbiased results) if the missing data mechanism is MCAR. However, this assumption is very hard to justify in practice. The specific missing data mechanism in presence is usually difficult to identify, first because all the participants would have to be interviewed on their reasons for not providing answers, and they would have to know and remember those reasons. Second, defending any single mechanism for a data matrix supposes that all participants failed to provide answers for the same reason, while in reality one may have found items too difficult while another may have simply run out of time, sometimes for the same items across participants. Finally, even for a specific item for a single individual, a combination of mechanisms might concur to explain the missing answer. For example, the participant may have skipped a difficult item (MNAR) in the hope of answering it later, but never found the time to get back to it (MAR). For those reasons, researchers should consider other options in addition to MCAR, such as MAR and MNAR. However, because the MNAR mechanism involves highly complex issues, we only consider in this paper the possibility of MCAR and MAR.

Methods

Many methods have been recommended to deal with missing data. In this paper, we will focus on three categories: deletion methods, simple imputation methods, and advanced methods . [1]

There are two types of deletion methods: Complete case (or listwise deletion) and Available case (or pairwise deletion). Complete-case deletion (C.C.) consists of removing the data for any participant that has at least one missing value. This is sometimes what researchers refer to when they mention the exclusion of their participants. In other words, if a participant declines to answer only one question or item, the whole questionnaire will be discarded for that person. Obviously, this method results in considerable loss of information and proves inefficient. The other method-Available case (A.C.)- consists of mitigating the loss of information by eliminating missing data on a case-by-case analysis. More specifically, with this method the researcher only discards the missing answers in a questionnaire, while keeping all the other answers obtained on that same questionnaire.

Regarding simple imputation techniques, those may produce biased results even for an ideal situation of MCAR. For example, one could think of replacing missing data by "0" without any explicit proof the student knows the answer. Another example is in a true/false questionnaire: the choice "true" is generally coded as "0".

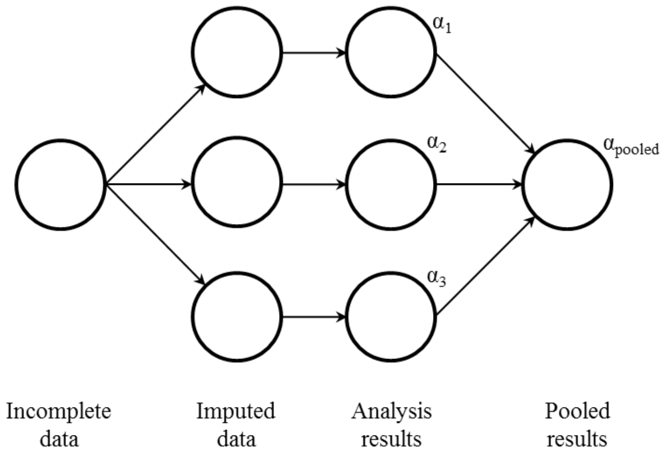
There are other simple imputation methods, some of which are listed here. One approach consists of substituting the item's mean for all participants with missing values on that item; In this case, the item mean (It. mean) of the observed cases is imputed for every missing value of an item:

$$It.\,mean = \sum_{i \in obs(j)} X_{ij}/\#obs(j) \qquad (1)$$

where $obs(j)$ denotes the items for which an answer is

[1] The interested reader can read Enders (2010) or Allison (2001) for more information.

Figure 1 ■ Steps of the multiple imputation methods (van Buuren, 2012)



available. A second example is Participant mean (Part. mean) substitution. In this case, the mean score obtained by each participant on the rest of the items is used to impute the missing values of that participant:

$$Part.\,mean = \sum_{j\in obs(i)} X_{ij}/\#obs(i)\,. \tag{2}$$

Here, $obs(i)$ denotes respondents who answered a specific question.

Winer (1971) proposed an alternative - here called Winer Imputation (W.I.)- that combines It. mean and Part. Mean and imputes the missing values by:

$$W.I. = \frac{It.\,mean + Part.\,mean}{2}. \tag{3}$$

Bernaards and Sijtsma (1999) have proposed Two-way imputation (T.-W.I.) where

$$T.-W.I. = Part.mean + It.\,mean - Tot.mean \tag{4}$$

where Tot. mean is the total mean of the test:

$$Tot.\,mean = \sum_{j\in obs}\sum_{i\in obs} X_{ij}/\#obs. \tag{5}$$

It is worth mentioning that Van Ginkel, Van der Ark, Sijtsma, and Vermunt (2007) focused on a Bayesian version of the Two-way imputation, but this method will not be covered in this paper due to our focus on simple techniques.

Modern techniques include multiple imputation. According to van Buuren (2012, p.17), multiple imputation consists of three main steps: imputation, analysis and pooling. As seen in Figure 1, the first step consists of imputing the missing data from an incomplete data set to produce several completed (imputed) data sets. Note that all the imputed data sets are different in order to represent the uncertainty about which value to impute.

In general, the imputation step leads to three to five complete data sets, although more imputations can be generated (e.g., 20 or 50). The next step is to perform the desired statistical analysis on each imputed data set. For example, we can compute the Cronbach's α for each imputed data set. Finally, the results are pooled to obtain a single statistics for inference, which is Cronbach's α in this example. Following Rubin (1987), the pooled estimate of the Cronbach's α (over imputations) is simply the arithmetic average of M replications

$$\bar{\alpha} = \frac{1}{M}\sum_{i}\alpha_k, \tag{6}$$

and its standard error (S.E.) is defined as

$$S.E.\,(\bar{\alpha}) = \sqrt{\frac{1}{M}\sum_k s_k^2 + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_k(\alpha_k-\bar{\alpha})^2} \tag{7}$$

where M is the number of replications, α_k is the Cronbach's α value in replication k, sk is the standard error estimate of α_k, , and $\bar{\alpha}$ is the mean of all Cronbach's α estimates.

Purpose of this study

The purpose of this study is to compare the impact on Cronbach's α of ten common fill-in methods of handling missing data for normal-size matrices in educational testing, using simulated data. The next sections will explain every step of our methodology.

Method

We examined the impact of missing data on Cronbach's α coefficients of the methods we described above. Note that we will present the results in this specific order:

(i) multiple imputation (M.I.);
(ii) Winer imputation (W.I.);
(iii) Two-way imputation (T.-W.I.);
(iv) replacement by the participant's mean (Part. mean);
(v) replacement by the item's mean (It. mean);
(vi) replacement by the total mean (Tot. mean);
(vii) replacement by "0";
(viii) replacement by "1";
(ix) only complete cases (C.C.); and
(x) all available cases (A.C.).

In addition we will add the empirical mean of α ($\alpha_{\text{Empirical}}$) obtained over all replications (1,000 in this study) before introducing missing data. Finally note that the M.I. method is added to this analysis following a recommendation by Sijtsma and Van der Ark (2003), who claimed that any investigation into missing data should include multiple imputation.

Cronbach's α coefficient

In language research and educational science, the Cronbach's α coefficient (Cronbach, 1951) has long been one of the most commonly used measure for assessing the internal consistency of tests and questionnaires. According to Sijtsma (2009), "probably no other statistic has been reported more often as a quality indicator of test scores than Cronbach (1951, 's) α coefficient (p. 107)". Peterson (1994) adds that "Not only is coefficient alpha the most widely used estimator of reliability, but also it has been the subject of considerable methodological and analytical attention" (p.382). Today, the use of α in research cannot be ignored: it is the best known coefficient to assess internal consistency, and that coefficient is widely available in popular software like SPSS, SAS, and R. Mathematically, this coefficient can be represented as

$$\alpha = \frac{j}{j-1}\left(1-\frac{\sum s_i^2}{s_T^2}\right) \qquad (8)$$

where j is the number of items on the test, s_i^2 is the variance of the ith items and s_T^2 is the total variance on all items. Consequently, in a case where $\alpha = 1$, all items are perfectly related to one another. On the contrary, if $\alpha = 0$, there is no link between items in the test. It is very important to understand that α quantifies the level of interrelatedness in a series of items, and that a high coefficient does not necessarily imply unidimensionality. Furthermore, the interpretation of Cronbach's α is relatively easy: Bland and Altman (1997) and Nunnally and Bernstein (1994) mention that an acceptable value for α is above 0.70. In a meta-analysis about this coefficient, Peterson (1994) analyzed 4,286 alpha coefficients, from 1,030 samples, and found a mean α coefficient of 0.77.

The simulation study details

Our procedure was based on the collection and analysis of dichotomous data, e.g., in the form of true/false questions or good/bad appraisals. We generated data sets for 20, 50, 250 and 500 participants and for 20 and 60 items. For this study, we chose two percentages of missing answers: 5% and 20%. Missing values were then created under MCAR and MAR mechanisms. For MCAR, missing values were randomly created in the data set. In the case of the MAR mechanism, we adopted a methodology (explained in full details in van Buuren, Brand, Groothuis-Oudshoorn, and Rubin, 2006) for creating intermittent missing values under MAR. This procedure ensures that, for each participant, the probability for an item to be missing only depends on the observed items of that participant. Finally, for each combination (item × participant × percentage), 1,000 matrices were generated. We will systematically report the Cronbach's α mean and the Cronbach's α standard deviation.

A didactic example of our R code

The R software was used for every analysis in this paper. Here is a short didactic example of how our analyses were performed. First, the user needs to load the R code provided in Listing 1 to 6 at the end of this article. Second, the estimated Cronbach's α can be obtained using the following function:

```
FUN(k1 = k1, k2 = k2, mech = mech, rate
    = rate)
```

where k1 is the number of participants, k2 is the number of items, mech is the missing data mechanism (with only two possibilities, "mar" and "mcar"), and rate is the rate of missing data. As an example, the following code gives Cronbach's alpha coefficient for a data matrix with 100 response patterns, 20 items, MAR mechanism, and 10% missing data:

```
FUN(k1 = 100, k2 = 20, mech = "mar",
    rate = 0.1).
```

Results

In this section, results will be presented for the MCAR and MAR mechanisms respectively.

Results for the MCAR condition

Table 1 shows the results for a hypothetical 20-item questionnaire with 5% missing data. For sample sizes of 20 and 50, our results show that the empirical means computed with It. mean and Tot. mean are very close to the $\alpha_{\text{Empirical}}$. M.I. presents a mean slightly below the $\alpha_{\text{Empirical}}$ value. All the other methods present a mean higher than the $\alpha_{\text{Empirical}}$. When the sample size increases to 250 and

Figure 2 ■ Standard deviation for 20 items, a nonresponse rate of 0.05, and MCAR

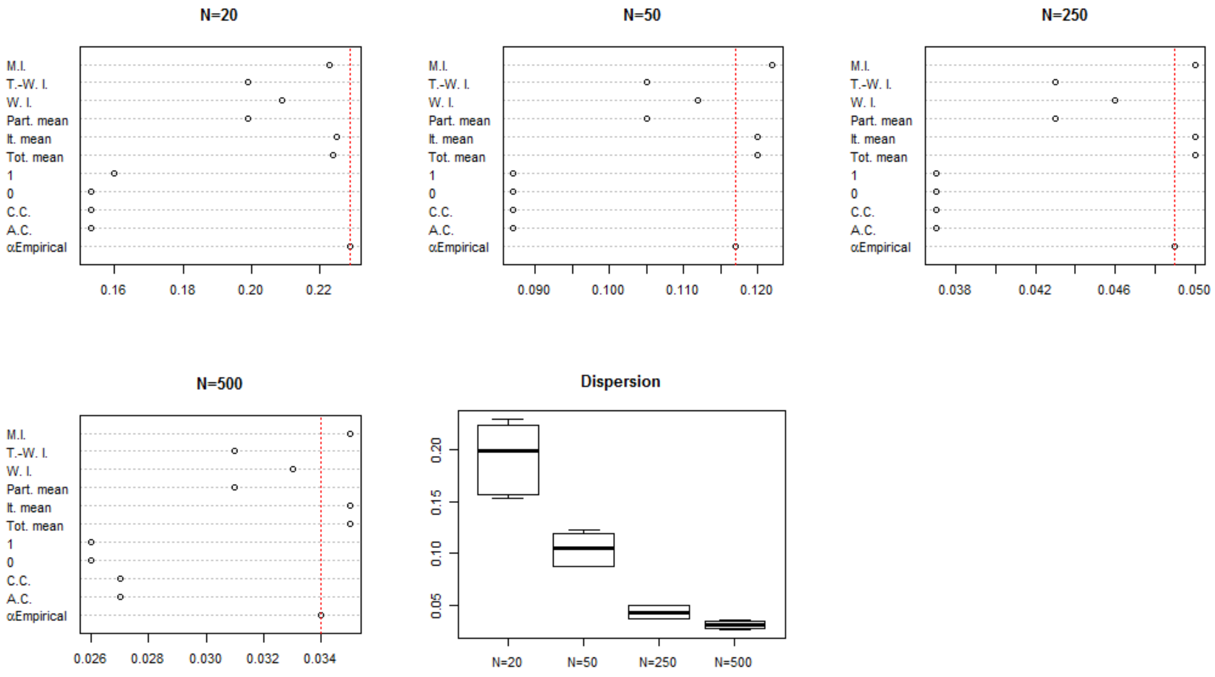


Table 1 ■ Cronbach's α mean for 20 items, a nonresponse rate of 0.05, and MCAR

	$N = 20$	$N = 50$	$N = 250$	$N = 500$
M.I.	0.413	0.458	0.490	0.495
T.-W.I.	0.485	0.522	0.542	0.545
W.I.	0.452	0.489	0.511	0.514
Part. mean	0.485	0.522	0.542	0.545
It. mean	0.422	0.460	0.484	0.488
Tot. mean	0.421	0.460	0.484	0.488
0	0.541	0.563	0.578	0.580
1	0.544	0.566	0.577	0.581
C.C.	0.544	0.565	0.576	0.580
A.C.	0.544	0.565	0.576	0.579
$\alpha_{\text{Empirical}}$	0.427	0.469	0.492	0.495

Table 2 ■ Cronbach's α mean for 60 items, a nonresponse rate of 0.05, and MCAR

	$N = 20$	$N = 50$	$N = 250$	$N = 500$
M.I.	0.744	0.756	0.774	0.777
T.-W. I.	0.768	0.786	0.793	0.794
W.I.	0.755	0.774	0.781	0.783
Part. mean	0.768	0.786	0.793	0.794
It. mean	0.743	0.764	0.772	0.773
Tot. mean	0.743	0.764	0.772	0.773
0	0.814	0.822	0.825	0.826
1	0.812	0.822	0.826	0.826
C.C.	0.812	0.822	0.825	0.826
A.C.	0.812	0.822	0.825	0.826
$\alpha_{\text{Empirical}}$	0.751	0.769	0.777	0.778

500 response patterns, M.I. presented the empirical mean that was closest to the $\alpha_{\text{Empirical}}$.

Figure 2 shows the results for the standard deviation, which was obtained from the estimated Cronbach's α over all replications. For every sample size, we see that Tot. mean, It. mean, M.I. and W.I. present the closest standard deviation to the $\alpha_{\text{Empirical}}$ value (i.e., the empirical standard deviation before inducing missing data). Furthermore, the dispersion of the standard deviation becomes less important as the sample size increases progressively to 500 response patterns.

Table 2 shows the results for 60-item data matrices with 5% missing data. For small sample sizes (N = 20 and N = 50), we notice that W.I., It. mean, Tot. mean, and M.I. presents the closest estimate to $\alpha_{\text{Empirical}}$. When the sample size increases to 250 and 500, It. mean, Tot. mean, M.I. and W.I. are the methods who yield empirical means closest to the $\alpha_{\text{Empirical}}$. Again, we see that A.C., C.C., "0", "1", Part. mean, and T.-W.I have the strongest impact on α.

Figure 3 shows the results for the standard deviation.

Figure 3 ■ Standard deviation for 60 items, a nonresponse rate of 0.05, and MCAR

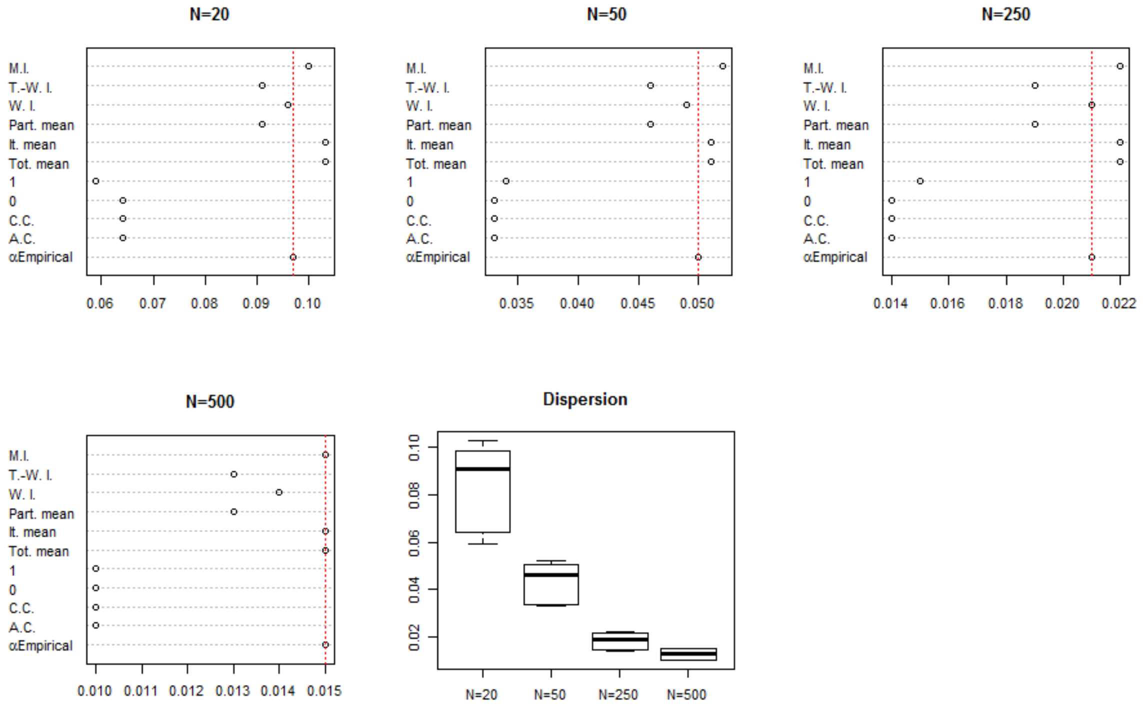


Table 3 ■ Cronbach's α mean for 20 items, a nonresponse rate of 0.20, and MCAR

	$N = 20$	$N = 50$	$N = 250$	$N = 500$
M.I.	0.329	0.421	0.483	0.491
T.-W. I.	0.624	0.653	0.667	0.669
W.I.	0.515	0.554	0.571	0.574
Part. mean	0.624	0.653	0.667	0.669
It. mean	0.368	0.430	0.455	0.459
Tot. mean	0.367	0.429	0.455	0.459
0	0.637	0.663	0.666	0.667
1	0.644	0.652	0.665	0.667
C.C.	0.627	0.636	0.649	0.651
A.C.	0.627	0.636	0.649	0.651
$\alpha_{\text{Empirical}}$	0.425	0.470	0.492	0.495

Table 4 ■ Cronbach's α mean for 60 items, a nonresponse rate of 0.20, and MCAR

	$N = 20$	$N = 50$	$N = 250$	$N = 500$
M.I.	0.724	0.705	0.759	0.769
T.-W. I.	0.819	0.831	0.837	0.837
W.I.	0.773	0.789	0.797	0.796
Part. mean	0.819	0.831	0.837	0.837
It. mean	0.714	0.738	0.751	0.751
Tot. mean	0.713	0.738	0.751	0.751
0	0.863	0.868	0.871	0.87
1	0.865	0.868	0.87	0.871
C.C.	0.859	0.862	0.864	0.864
A.C.	0.859	0.862	0.864	0.864
$\alpha_{\text{Empirical}}$	0.752	0.769	0.777	0.777

For $N = 20$, $N = 50$, and $N = 250$, W. I. yields the closest value to the $\alpha_{\text{Empirical}}$. For $N = 500$, we see that Tot.mean, It. mean, M.I. and W.I. present the closets standard deviation with the αEmpirical. As was the case with figure 2, we see that the dispersion of the standard deviation becomes less important as the sample size increases.

Table 3 displays results for 20 items with 20% missing data. For a sample size of 20 and 50 participants, the methods with the value closest to the $\alpha_{\text{Empirical}}$ are Tot. mean, It. mean, and M.I. When the sample size increases to 250 and 500 participants, M.I. is systematically the most precise method.

Figure 4 presents the results for the standard deviation. Tot.mean, It. mean, M.I. and W.I. present, in every situation, the standard deviation closest to the $\alpha_{\text{Empirical}}$. Finally, the dispersion of the standard deviation shrinks considerably when the sample size increases.

The results for 60 items with 20% of missing data are

Figure 4 ■ Standard deviation for 20 items, a nonresponse rate of 0.20, and MCAR

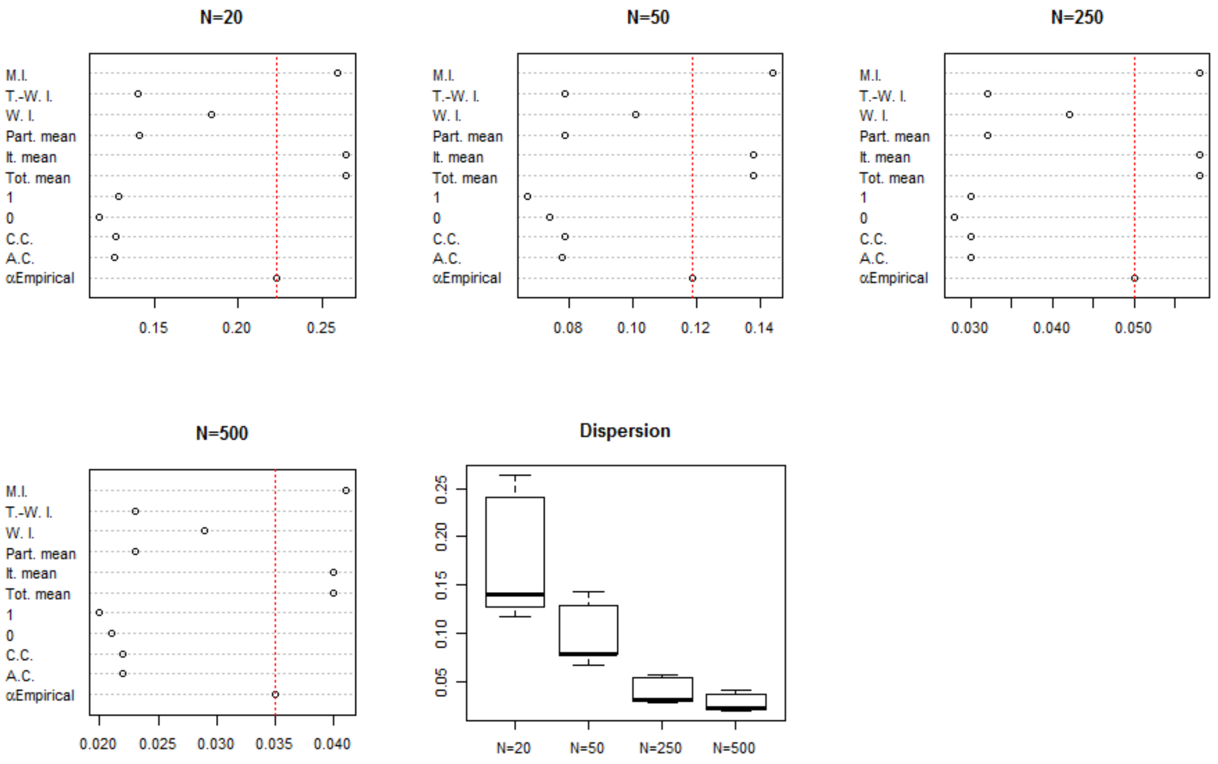


Table 5 ■ Cronbach's α mean for 20 items, a nonresponse rate of 0.05, and MAR

	$N = 20$	$N = 50$	$N = 250$	$N = 500$
M.I.	0.401	0.460	0.489	0.494
T.-W. I.	0.495	0.531	0.550	0.553
W.I.	0.453	0.493	0.512	0.516
Part. mean	0.495	0.531	0.550	0.553
It. mean	0.409	0.458	0.480	0.483
Tot. mean	0.409	0.458	0.480	0.483
0	0.519	0.550	0.567	0.569
1	0.539	0.567	0.582	0.586
C.C.	0.537	0.566	0.581	0.585
A.C.	0.537	0.566	0.581	0.585
$\alpha_{\text{Empirical}}$	0.428	0.471	0.491	0.495

Table 6 ■ Cronbach's α mean for 60 items, a nonresponse rate of 0.05, and MAR

	$N = 20$	$N = 50$	$N = 250$	$N = 500$
M.I.	0.746	0.751	0.772	0.775
T.-W. I.	0.773	0.787	0.794	0.795
W.I.	0.757	0.772	0.780	0.781
Part. mean	0.773	0.787	0.794	0.795
It. mean	0.741	0.759	0.768	0.769
Tot. mean	0.741	0.759	0.768	0.769
0	0.802	0.815	0.821	0.821
1	0.807	0.819	0.825	0.826
C.C.	0.806	0.818	0.825	0.825
A.C.	0.806	0.818	0.825	0.825
$\alpha_{\text{Empirical}}$	0.752	0.768	0.776	0.778

presents in the next table. We only report the results for the two best methods for each situation. For $N = 20$ the best methods are respectively W.I. and M.I. For $N = 50$, W.I. is again the best method, followed by It. mean and Tot. mean. When the sample size rises to 250 participants, W.I. and M.I. are the methods that yield means closest values to the $\alpha_{\text{Empirical}}$. Finally, W.I. and M.I. are the most precise methods for $N = 500$.

Figure 5 shows that for every sample size, W.I. present the standard deviation closest to the $\alpha_{\text{Empirical}}$. Without any surprise, the dispersion of the standard deviation becomes less important when the sample size increases to $N = 500$.

Figure 5 ■ Standard deviation for 60 items, a nonresponse rate of 0.20, and MCAR

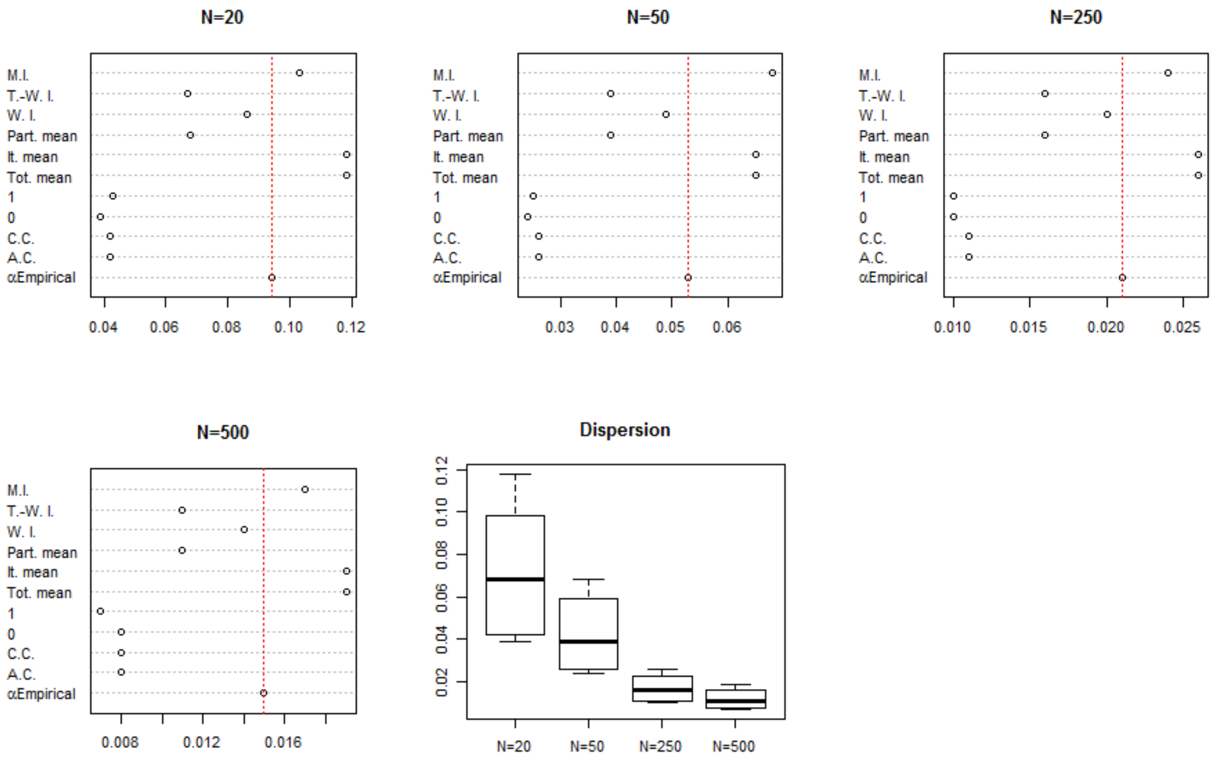


Table 7 ■ Cronbach's α mean for 20 items, a nonresponse rate of 0.20, and MAR

	$N = 20$	$N = 50$	$N = 250$	$N = 500$
M.I.	0.330	0.413	0.479	0.490
T.-W. I.	0.642	0.666	0.679	0.680
W.I.	0.530	0.559	0.575	0.577
Part. mean	0.642	0.666	0.679	0.680
It. mean	0.365	0.416	0.441	0.446
Tot. mean	0.365	0.415	0.441	0.446
0	0.598	0.614	0.628	0.631
1	0.668	0.685	0.690	0.690
C.C.	0.647	0.667	0.674	0.675
A.C.	0.647	0.668	0.674	0.675
$\alpha_{\text{Empirical}}$	0.428	0.473	0.492	0.495

Table 8 ■ Cronbach's α mean for 60 items, a nonresponse rate of 0.20, and MAR

	$N = 20$	$N = 50$	$N = 250$	$N = 500$
M.I.	0.722	0.698	0.755	0.768
T.-W. I.	0.824	0.834	0.840	0.840
W.I.	0.773	0.788	0.794	0.795
Part. mean	0.824	0.834	0.840	0.840
It. mean	0.701	0.729	0.739	0.741
Tot. mean	0.701	0.729	0.739	0.741
0	0.844	0.854	0.857	0.857
1	0.871	0.875	0.877	0.878
C.C.	0.864	0.869	0.871	0.872
A.C.	0.864	0.869	0.871	0.872
$\alpha_{\text{Empirical}}$	0.752	0.770	0.776	0.778

Results for the MAR condition

This section presents the results for the MAR condition. Table 5 reports the results for a 20-item test with 5% of missing data. For sample sizes comprising 20 and 50 participants, our results show that W.I., It. mean and Tot. mean provide the closest α means to $\alpha_{\text{Empirical}}$. When the sample size increases, it is M.I. that yields means that are closest to the $\alpha_{\text{Empirical}}$.

Figure 6 illustrates the results for the standard deviation. For every sample size, we see that Tot.mean, It. mean, M.I. and W.I. present the standard deviations closest to the $\alpha_{\text{Empirical}}$. Furthermore, the dispersion of the standard deviation becomes smaller when the sample size becomes larger.

Figure 6 ■ Standard deviation for 20 items, a nonresponse rate of 0.05, and MAR

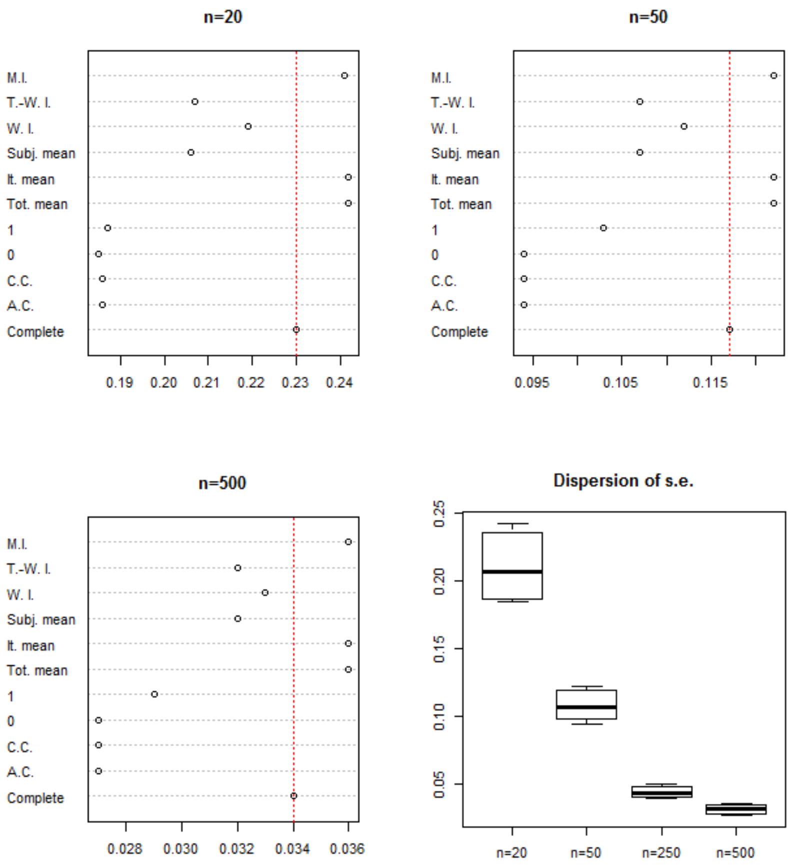


Table 9 ■ Mean estimated as a function of sample size ($N = 20, 50, 250$ or 500) across number of items (20 or 60), across nonresponse rate (.05 or .20) and across type of missingness (MCAR or MAR)

	$N = 20$	$N = 50$	$N = 250$	$N = 500$
M.I.	0.551	0.583	0.625	0.632
T.-W. I.	0.679	0.701	0.713	0.714
W.I.	0.626	0.652	0.665	0.667
Part. mean	0.679	0.701	0.713	0.714
It. mean	0.558	0.594	0.611	0.614
Tot. mean	0.558	0.594	0.611	0.614
0	0.702	0.719	0.727	0.728
1	0.719	0.732	0.739	0.741
C.C.	0.712	0.726	0.733	0.735
A.C.	0.712	0.726	0.733	0.735
$\alpha_{\text{Empirical}}$	0.589	0.62	0.634	0.636

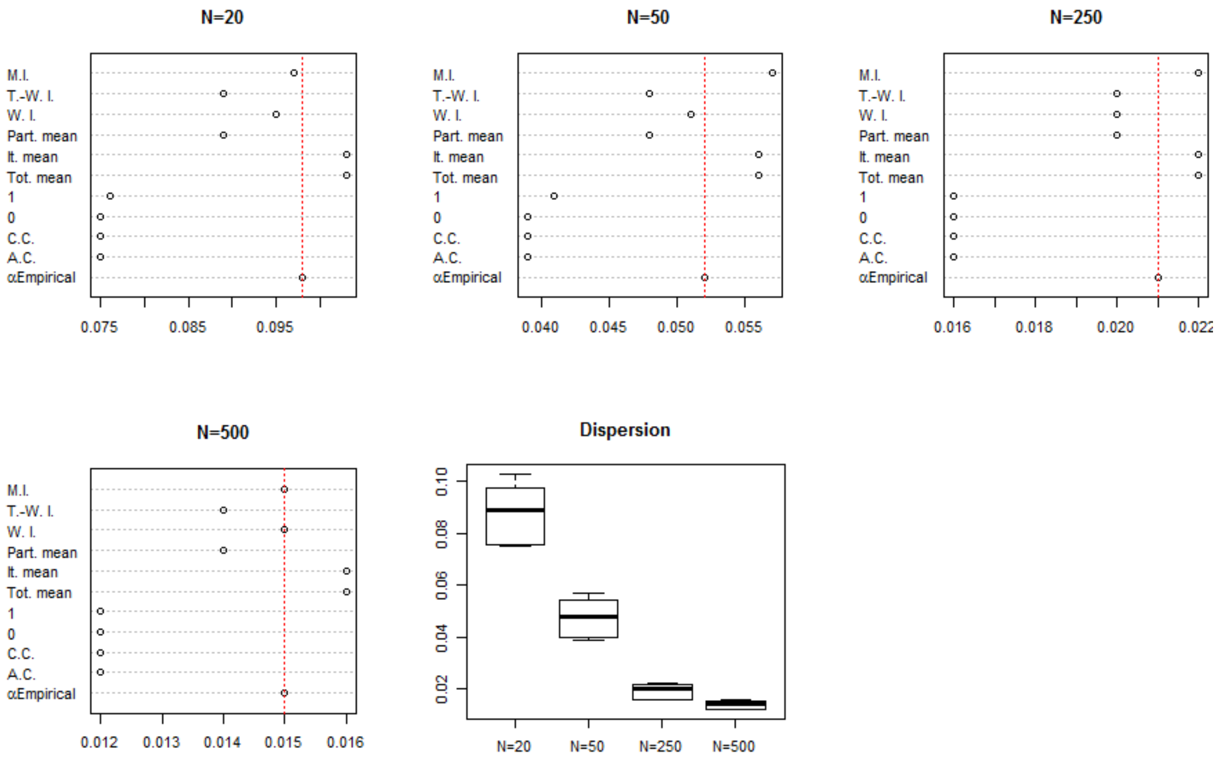
Table 6 shows the results for 60 items with 5% missing data. For 20 participants, the α mean obtained from M.I., W.I., Tot. mean, and It. Mean yield results close to $\alpha_{\text{Empirical}}$. Every other method presents a value above the $\alpha_{\text{Empirical}}$. Finally, when the sample size increases, M.I. is the methods which yield the closest values to $\alpha_{\text{Empirical}}$.

As we can read in Figure 7, M.I. presents the smallest standard deviation for $N = 20$. For $N = 50$, the best methods becomes W.I. In the case of a sample size equal to 250, Tot. mean, It. mean, Part. mean, M.I. and W.I. present the standard deviations closest to the $\alpha_{\text{Empirical}}$. Finally, for $N = 500$, M. I. and W.I. are very close to the $\alpha_{\text{Empirical}}$.

Table 7 shows the results for 20 items with 20% missing data. For a sample size of 20 and 50 response patterns, the methods with the values closest to the $\alpha_{\text{Empirical}}$ are It. mean, Tot. mean, and M.I.. When the sample size increases to 250 and 500 response patterns, M.I. is systematically the most precise method.

The next figure shows the standard deviations. For every sample size, we see that Tot.mean, It. mean, M.I. and W.I. present the standard deviations closest to the $\alpha_{\text{Empirical}}$. In addition, we see that the dispersion of the

Figure 7 ■ Standard deviation for 60 items, a nonresponse rate of 0.05, and MAR



standard deviation becomes smaller with larger sample sizes.

The results for 60 items with 20% of missing data are presented in Table 8. We report once again the results for the two best methods for each situation. Except for $N = 50$, M.I. is the most precise method to estimate the $\alpha_{\text{Empirical}}$.

The analysis of standard deviations shows us that W.I. presents the closest proximity to the $\alpha_{\text{Empirical}}$. Again, we see that the dispersion of the standard deviation becomes smaller with larger sample sizes.

Summary of the results

We further provide a synthesis table that summarizes all simulation results presented in Tables 1-8. Table 9 shows the mean estimated α for every sample size across a number of items, nonresponse rates, and types of missingness.

Next, Table 10 shows the mean bias (i.e., the estimate minus $\alpha_{\text{Empirical}}$) per sample size. We see that most techniques improve slowly as N increases. The exception is M.I., which rapidly outperforms the other techniques when the sample size is large enough (the improvement is negligible when $N \leq 50$).

Finally, our results show that under all simulation conditions M.I., W.I., It. mean, and Tot. mean are the most precise methods. In the last column of Table 10, we can distinguish that M.I, It.mean, Tot.Mean, and W.I are in the cluster of the best method. All other methods are in another cluster of methods presenting the worst estimates for $\alpha_{\text{Empirical}}$.

Discussion

The most common ways used by researchers to deal with missing data (replacing by zero, deleting the data, or excluding the participant) happen to be among the methods that have the strongest impact on Cronbach's α coefficients. As highlighted by Little (1988), it is important to understand the effect of naive imputations, because their effect can be worse than not doing anything about missing data.

Like Huisman (2000), our results show that the amount and type of missing data, and the characteristics of the matrix (sample size and number of items) have an effect on Cronbach's α. Sijtsma and Van der Ark (2003) also show that T.-W.I. overestimates Cronbach's α. These authors demonstrate that the bias in Cronbach α is slightly higher for Part. mean than for T.-W.I.

Many authors have shown and pleaded for the efficacy of M.I. (Schafer & Graham, 2002; van Buuren, 2012). Our

Figure 8 ■ Standard deviation for 20 items, a nonresponse rate of 0.20, and MAR

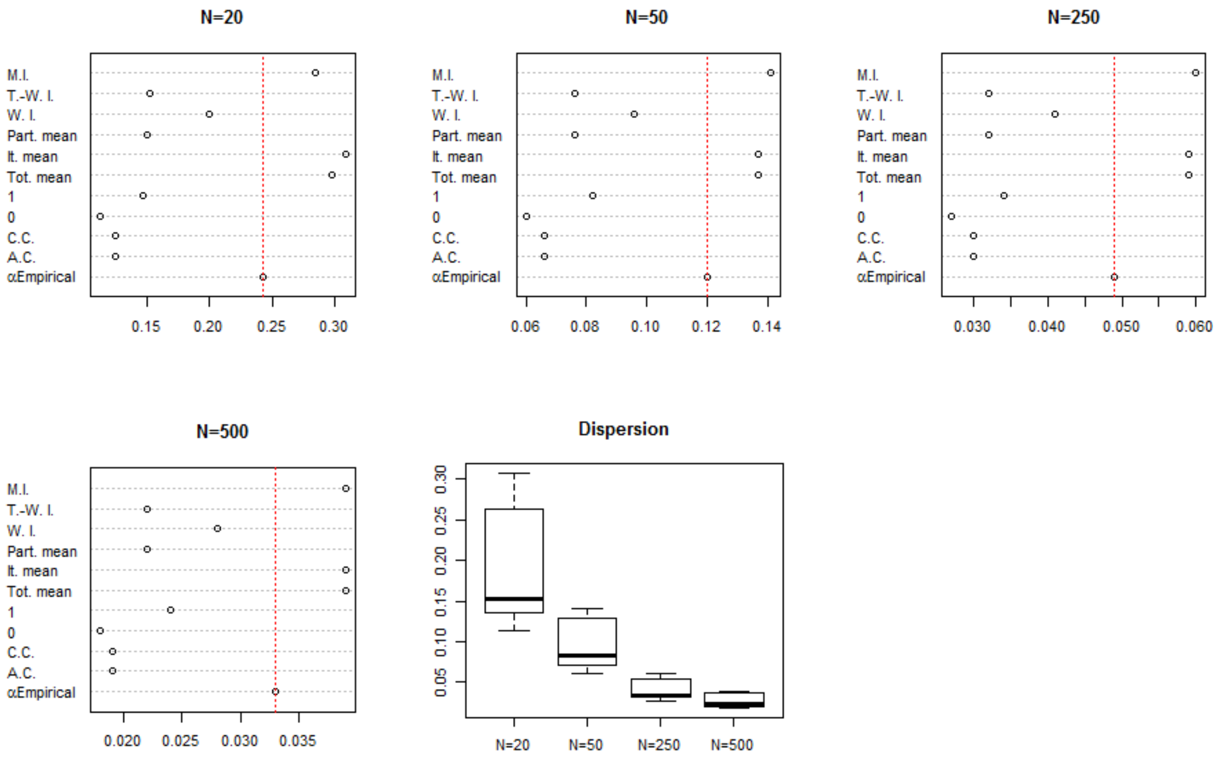


Table 10 ■ Mean bias as a function of sample size ($N = 20, 50, 250$ or 500) across the number of items (20 or 60), across nonresponse rate (.05 or .20) and across types of missingness (MCAR or MAR). The last column is the mean across sample size as well.

	20	50	250	500	Mean
M.I.	-0.038	-0.037	-0.009	-0.004	-0.022
T.-W. I.	0.089	0.081	0.079	0.078	0.082
W.I.	0.037	0.032	0.031	0.031	0.033
Part. mean	0.089	0.081	0.079	0.078	0.082
It. mean	-0.032	-0.026	-0.023	-0.023	-0.026
Tot. mean	-0.032	-0.026	-0.023	-0.023	-0.026
0	0.113	0.099	0.092	0.091	0.099
1	0.129	0.112	0.105	0.104	0.112
C.C.	0.123	0.106	0.099	0.098	0.106
A.C.	0.123	0.106	0.099	0.098	0.106

results also showed that M.I. generally yields Cronbach coefficients that are close to the $\alpha_{\text{Empirical}}$, especially when the sample is greater than 250 participants.

Finally, simple imputation methods based on arithmetic means, such as W.I., Tot. mean and It. mean are very interesting options because they present results relatively close to the $\alpha_{\text{Empirical}}$. Like many statisticians, Enders (2010) is not enthusiastic about using mean imputation:

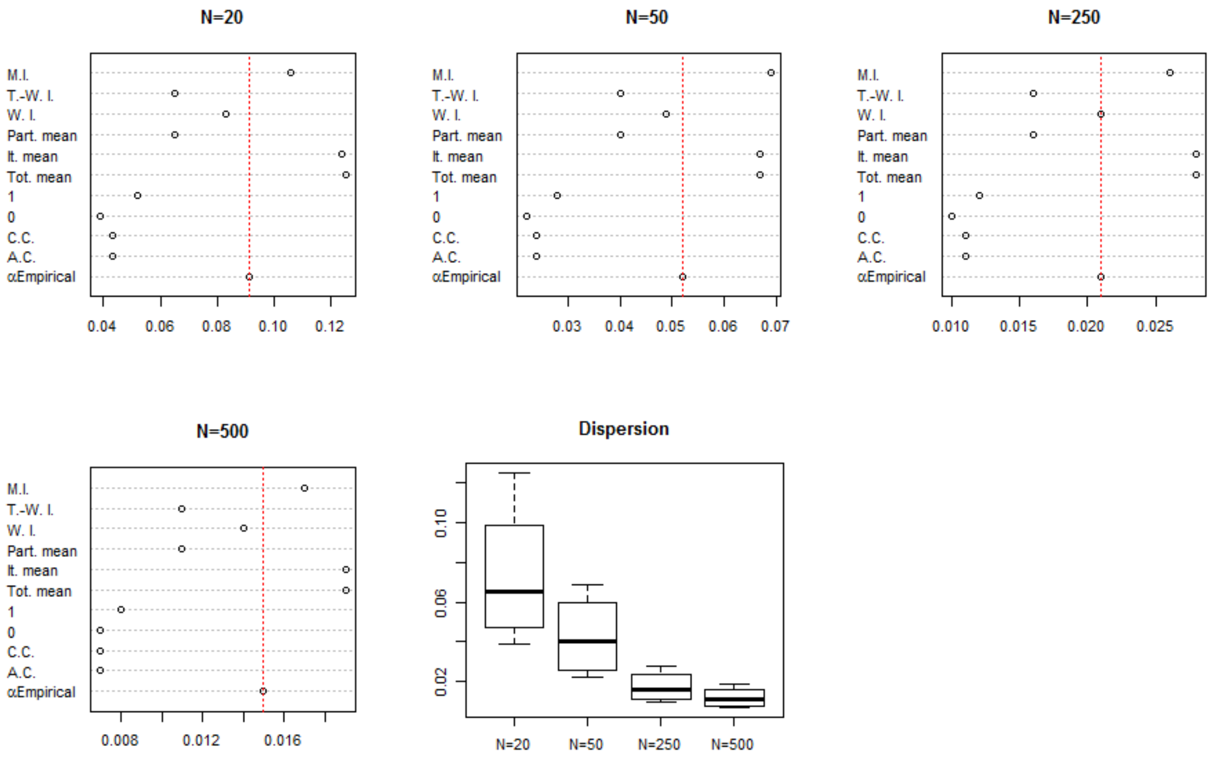
> simulation studies suggest that mean imputation [herein called Tot.mean, It. mean, and Part. mean] is possibly the worst missing data handling method available. Consequently, in no situation is mean imputation defensible, and you should absolutely avoid this approach (p. 43).

Our results seem promising. Furthermore, these simple imputation strategies can be easily used by non-statisticians like teachers and school psychologists.

Conclusion

This simulation study analyzed the effect on Cronbach's α of ten methods for handling missing data. Our simulation allowed us to discover two different categories for those strategies. Based on how little they impact Cronbach's alpha coefficient and how little they distort the instrument's internal consistency, the four best methods for dealing with missing data are It. mean, Tot. mean, W.I. and

Figure 9 ■ Standard deviation for 60 items, a nonresponse rate of 0.20, and MAR



M.I. These methods also present the smallest standard deviation across every simulation scenario.

Based on the study of their mean α, Tot. mean and It. mean provide more interesting results with small sample sizes. However, their standard deviations are not systematically the lowest. When the sample size increases, M.I. is the most precise method, followed by W.I. These methods also present low standard deviations.

Many people without statistical training do not know what to do with missing data. The first reflex is to ignore them or to insert a zero as if grading an exam. This strategy is not a good option because it provides a distorted view of the effectiveness of the test. Our results suggest that if we have few participants and items, Tot. mean, It. Mean, W.I., and M.I. should be adopted, and if we have a lot of items (60 and above) and participants (e.g., more than 200), M.I. is the best method.

This study presents some limitations, and further research is warranted to better understand the effect of missing data handling on Cronbach's α coefficients. First, we only investigated dichotomous data. It would be of interest to perform the same kind of analysis on polytomous data sets and investigate the impact of deletion and various replacement methods with such data. Furthermore, although Cronbach's α is the dominant reliability measure found in language research publications, other measures also exist. Cronbach's α has been criticized (see Sijtsma, 2009) because other approaches seem to be better at measuring single concepts (e.g., Revelle and Zinbarg, 2009). From this perspective, similar research could be conducted using the ω_t coefficient (McDonald, 1978, 1999), which could allow for similar comparisons of impact between deletion and imputation methods. Lastly, the results of this simulation study need to be compared with results stemming from real data matrices.

References

Allison, P. D. (2001). *Missing data*. Thousand Oaks, Californie: Sage.

Bernaards, C. A. & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research*, *34*, 277–313.

Bland, J. & Altman, D. (1997). Statistics notes: Cronbach's alpha. *British Medical Journal*, 314–275.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.

Giguère, G., Hélie, S., & Cousineau, D. (2004). Manifeste pour le retour des sciences en psychologie. *Revue Québécoise de psychologie*, *25*, 117–130.

Huisman, M. (2000). Imputation of missing item responses: some simple techniques. *Quality & Quantity*, *34*, 331–351.

Lazaraton, A., Riggenbach, H., & Ediger, A. (1987). Forming a discipline: Applied linguists' literacy in research methodology and statistics. *TESOL Quarterly*, *21*, 263–277.

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*, 287–296.

McDonald, R. P. (1978). Generalizability in factorable domains: "domain validity and generalizability": 1. *Educational and Psychological Measurement*, *38*, 75–79.

McDonald, R. P. (1999). *Test theory: a unified treatment*. Hillsdale, NJ: Erlbaum.

Nunnally, J. & Bernstein, L. (1994). *Psychometric theory*. New York: McGraw-Hill.

Peterson, R. A. (1994). A meta-analysis of cronbach's coefficient alpha. *Journal of Consumer Research*, *21*, 381–391.

Peugh, J. L. & Enders, C. K. (2004). Missing data in educational research: a review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*, 525–556.

Pichette, F., Béland, S., Jolani, S., & Lesǹiewska, J. (2015). The handling of missing binary data in language research. *Studies in Second Language Learning and Teaching*, *5*, 153–169.

Revelle, W. & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: comments on sijtsma. *Psychometrika*, *74*, 145–154.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

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

Schafer, J. L. & Graham, J. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Schmidtke, J., Spino, L. A., & Lavolette, B. (2012, October). How statistically literate are we? examining sla professors' and graduate students' statistical knowledge and training. In *31st second language research forum*. Pittsburgh, USA,

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods*, *18*, 572–582.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120.

Sijtsma, K. & Van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, *38*, 505–528.

Van Ginkel, J. R., Van der Ark, L. A., Sijtsma, K., & Vermunt, J. K. (2007). Two-way imputation: a Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Computational Statistics and Data Analysis*, *51*, 4013–4027.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, K., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*, 1049–1064.

Winer, B. J. (1971). *Statistical principles in experimental design (2nd ed.)* New York: McGraw-Hill.

Yang, K. (2010). *Making sense of statistical methods in social research*. London: Sage.

Citation

Béland, S., Pichette, F., & Jolani, S. (2016) Impact on Cronbach's α of simple treatment methods for missing data. *The Quantitative Methods for Psychology*, *12*(*1*), 57-73.

Copyright © 2016 Béland, Pichette, & Jolani. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 23/07/2015 ~ Accepted: 12/10/2015

Listings 1 to 6 follows on next page

Listing 1 ■ R code used for the simulations

```
# Main program

FUN <- function(re = 1, ...){
  resu <- matrix(NA, nrow = re, ncol = 11) # the number of methods investigated
  for (t in 1:re){
    com.data <- DATA.GEN(...)
    # complete data (no missing data)
    tr.alpha <- CRON(com.data)
    # generate missing data
    mis.data <- MISSING.GEN(com.data,...)
    # cronbach's alpha
    est.alpha <- COMPU(mis.data)
    resu[t,] <- c(tr.alpha, est.alpha)
  }
  result <- apply(resu, 2, mean)
  names(result) <- c("True", "AC", "CC", "Zero.I", "One.I",
                       "ToMean.I", "Item.I", "Part.I", "Winer.I", "Two-way.I", "MI"
   )
  #list(result = result, resu = resu)
  return(result)
}
```

Listing 2 ■ R code used in the simulations (continued)

```
# Generate data
##################################
# k1: the number of participants #
# k2: the number of items        #
##################################
DATA.GEN <- function(k1 = 20, k2 = 20, ...){
  # load a package to generate data
  require(psych)
  data <- as.matrix(sim.dichot(nsub = k1, nvar = k2, gloading = 0.3))
  return(data)
}
```

Listing 3 ■ R code used in the simulations (continued)

```
# Generate missing data
#############################00######################################
# y (n * m): data set                                               #
# rate (1 * 1): fraction of missing data                            #
# pat (k * m): missing data patterns                                #
# a (k * m): coefficients                                           #
# f (1 * k): relative frequencies of patterns with respect to sum 1 #
# quant (k * M): quantiles                                          #
# g (k * M): arbitrarily chosen weights                             #
#####################################################################
MISSING.GEN <- function (y, rate = 0.1, alpha = .5, mech = "mcar", ...){
  n <- nrow(y)
  m <- ncol(y)
  # alpha is percentage missing values in a row (alpha=.25 implies 25% missing)
  # the number of incomplete rows is n*rate/alpha
  box <- c(rep(0, m*alpha), rep(1, (m - m*alpha)))
  incomp <- n*rate/alpha
  pat <- matrix(NA, ncol = m, nrow = incomp)
  for (i in 1:incomp) pat[i, ] <- sample(box, m)
  if(mech == "mcar"){
   # candidate rows for being incomplete
   cand <- sample(1:n, incomp)
   y[cand,][pattern == 0] <- NA
  }
  if(mech == "mar"){
    quant <- matrix(c(.33,.66), nrow = incomp, ncol = 2, byrow = TRUE)
    g <- matrix(c(.25, 1), nrow = incomp, ncol = 2, byrow = TRUE)
    u <- runif(n)
    cand <- rep(1,n)
    fc <- 0
    for (i in 1:incomp){
      fc <- fc + 1/incomp
      ct <- as.numeric(u > fc)
      cand <- cand + ct
      ct <- 0
    }
    s <- y%*%t(pat)
    p <- rep(0, n)
    for (i in 1:incomp){
      c <- quantile(s[,i][cand == i], probs =quant[i,], type=1)
      si <- s[cand == i,i]
      cl <- rep(1, length(si))
        for (j in 1:length(c)){
          fl <- rep(c[j], length(si))
          bl <- as.numeric(si > fl)
          cl <- cl + bl
          fl <- bl <- 0
        }
```

Listing 4 ■ R code used in the simulations (continued)

```
      pi <- rep(0, length(si))
      gi <- c(1, g[i,])
      qi <- c(0, quant[i,], 1)
      sum <- 0
      for (k1 in 1:length(gi)){ sum <- sum + (qi[k1+1] - qi[k1])*gi[k1] }
      for (k2 in 1:(length(c) + 1)){ pi[cl == k2] <- 2*rate*gi[k2]/sum }
      p[cand == i] <- pi
    }
    u2 <- runif(n)
    incompl <- as.numeric(u2 <= p)
    cand1 <- cand*incompl
    r <- matrix(0, nrow = n, ncol = m)
    r[cand1 == 0,] <- 1
    for (i in 1:incomp){
      for (j in 1:n){
        if (cand1[j] == i) r[j,] <- pat[i,]
      }
    }
    for (i in 1:n){
      for (j in 1:m){
        if (r[i,j] == 0) y[i,j] <- NA
      }
    }
  }
  return(y)
}
```

Listing 5 ■ R code used in the simulations (continued)

```
# Compute cronbach’s alpha
CRON <- function(y, mis = FALSE, met = FALSE){
  y <- as.matrix(y)
  m <- ncol(y)
  alpha <- if (!met){
    vitem <- sum(diag(var(y, na.rm = mis)))
    vtot <- var(rowSums(y, na.rm = mis))
    (m/(m - 1)) * (1 - (vitem/vtot))
  }
  else {
    vitem <- sum(diag(cov(y, use = "pairwise.complete.obs")))
    vtot <- var(rowSums(y, na.rm = mis))
    (m/(m - 1)) * (1 - (vitem/vtot))
  }
  return(alpha)
}
```

Listing 6 ■ R code used in the simulations (continued)

```r
# Computation of cronbach's alpha for various imputation and deletion methods
COMPU <- function(y){
  out <- vector()
  # DELETION METHODS
  # available case (A.C.)
  out[1] <- CRON(y, mis = TRUE)
  # complete case (C.C.)
  out[2] <- CRON(y, mis = TRUE, met = TRUE)
  # SINGLE IMPUTATION METHOD
  temp1 <- temp <- y
  # zero replacement
  temp[is.na(y)] <- 0
  out[3] <- CRON(temp)
  # one replacement
  temp[is.na(y)] <- 1
  out[4] <- CRON(temp)
  # overall mean imputation (Tot. mean)
  temp[is.na(y)] <- mean(y, na.rm = T)
  out[5] <- CRON(temp)
  # item's mean imputation (It. mean)
  i.mean <- colMeans(y, na.rm = T)
  for (j in 1:ncol(y)){ temp[,j][is.na(y[,j])] <- i.mean[j] }
  out[6] <- CRON(temp)
  # participant's mean imputation (Part. Mean)
  p.mean <- rowMeans(y, na.rm = T)
  for (i in 1:nrow(y)){ temp[i,][is.na(y[i,])] <- p.mean[i] }
  out[7] <- CRON(temp)
  # other single imputation methods
  for (i in 1:nrow(y)){
    for (j in 1:ncol(y)){
      temp[i,j][is.na(y[i,j])] <- (p.mean[i] + i.mean[j])/2
      temp1[i,j][is.na(y[i,j])] <- p.mean[i] + i.mean[j] - mean(y, na.rm = T)
    }
  }
  # average of item and participant's mean (W.I.)
  out[8] <- CRON(temp)
  # Two-way imputation method (T.-W.I.); PM + IM - OM
  out[9] <- CRON(temp1)
  # MULTIPLE IMPUTATION (M.I.)
  mires <- vector()
  imp <- mice(y, print = F)
  # see van Buuren & Groothuis-Oudshoorn (2011)
  for (i in 1:imp$m) mires[i] <- CRON(complete(imp, i), mis = TRUE)
  out[10] <- mean(mires)
return(out)
}
```