Multivariate Behavioral Research, 49:443–459, 2014

ISSN: 0027-3171 print / 1532-7906 online
DOI: 10.1080/00273171.2014.931799

Routledge
Taylor & Francis Group

A Cautious Note on Auxiliary Variables That Can Increase Bias in Missing Data Problems

Felix Thoemmes
Cornell University

Norman Rose
University of Tübingen

The treatment of missing data in the social sciences has changed tremendously during the last decade. Modern missing data techniques such as multiple imputation and full-information maximum likelihood are used much more frequently. These methods assume that data are missing at random. One very common approach to increase the likelihood that missing at random is achieved consists of including many covariates as so-called auxiliary variables. These variables are either included based on data considerations or in an inclusive fashion; that is, taking all available auxiliary variables. In this article, we point out that there are some instances in which auxiliary variables exhibit the surprising property of increasing bias in missing data problems. In a series of focused simulation studies, we highlight some situations in which this type of biasing behavior can occur. We briefly discuss possible ways how one can avoid selecting bias-inducing covariates as auxiliary variables.

The presence of missing data is a prevalent problem in social science research (Peugh & Enders, 2004) and has triggered much research during the last 30 years. This research has culminated in sophisticated methods to deal with missing values, specifically the use of full-information maximum likelihood (FIML) and multiple imputation (MI). Both of these so-called modern missing data techniques are expected to yield consistent estimates of parameters in the presence of missing data given that certain assumptions about missingness hold, in particular that data are missing at random (MAR). It should be noted that especially MI, although conceptually straightforward (Rubin, 1996), can be conducted with various different techniques; see, for example, Schafer (1999); King, Honaker, Joseph, and Scheve (2001); van Buuren and Groothuis-Oudshoorn (2011); or Raghunathan, Lepkowski, Hoewyk, and Solenberger (2001). To make MAR more plausible, sophisticated methods have been developed to include so-called auxiliary variables into both FIML and MI approaches. The goal of this article is to critically examine the use of auxiliary variables in missing data problems by providing examples in which somewhat surprisingly bias increases when auxiliary variables are used.

We first briefly review classic missingness mechanisms and discuss which conditional independencies these mechanisms imply. Then, we review recommendations to include auxiliary variables and highlight certain situations in which auxiliary variables can have detrimental effects and potentially increase bias. Using a series of examples we describe conditions under which this bias can occur and follow up with several simulation studies that explore the biasing behavior of auxiliary variables. We finish with a brief discussion on current practices with regard to auxiliary variables.

Correspondence concerning this article should be addressed to Felix Thoemmes, MVR G62A, Cornell University, Ithaca, NY 14853. E-mail: felix.thoemmes@cornell.edu

MISSING DATA MECHANISM

Based on the seminal work by Rubin (1976), three missing data mechanisms can be distinguished: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Here we present a brief overview of these mechanisms and define them using both equality of conditional probabilities and conditional independence statements.

MCAR

Using standard notation, Y is an $N \times K$ matrix. The rows of Y are individual cases $n = 1, \ldots, N$. The columns of Y are the variables $i = 1, \ldots, K$. It is important to note that Y is not necessarily a single variable but can be a data matrix of many variables. Y is partitioned into an observed part, Y_{obs}, and a missing part, Y_{mis}. To indicate missing data, a matrix R is introduced that codes whether a particular datum is missing or observed. Usually this matrix consists of values coded 0 and 1 to denote either observed or missing values. Just as Y, R is not a single variable but a whole matrix, potentially representing the missingness of many variables.

Missing completely at random (MCAR) is defined as the equivalence of the unconditional probability distribution of missingness $P(R)$ and the conditional probability distribution of missingness given Y_{obs} and Y_{mis}, or simply Y.

$$P(R \mid Y) = P(R \mid Y_{obs}, Y_{mis}) = P(R). \tag{1}$$

One may also express this equivalence using conditional independence statements, as in

$$R \perp\!\!\!\perp (Y_{obs}, Y_{mis}). \tag{2}$$

The independence statement highlights that under MCAR missingness is completely unrelated to any observed or unobserved variables. In an applied context this may be conceptualized that missing values arose from a purely random process and that no variables in the data are related to missingness. In cases in which MCAR holds, even some simple treatments, like listwise deletion of missing data, are expected to yield unbiased results (Enders, 2010). MCAR cannot be "proven" using statistical techniques and thus needs to be argued for based on theoretical assumptions. Tests of homogeneity of means and variances (Little, 1988) only provide necessary but not sufficient evidence for MCAR (Gelman & Hill, 2007; Raykov, 2011).

MAR

Missing at random (MAR) is defined as the equality of the conditional probability distribution of missingness, given the observed part of Y, and the conditional probability distribution of missingness, given both the observed and unobserved part of Y.

$$P(R \mid Y) = P(R \mid Y_{obs}, Y_{mis}) = P(R \mid Y_{obs}). \tag{3}$$

Again, this equality can be expressed using conditional independence notation:

$$R \perp\!\!\!\perp Y_{mis} \mid Y_{obs}. \tag{4}$$

The independence statement expresses that under an MAR mechanism the unobserved portion of Y and the missingness R are independent of each other as long as we condition on the observed portion of Y. Differently said, in an applied research context, we assume that missingness is affected by both observed and unobserved variables but that there is no relationship between missingness and unobserved variables, given the observed variables. Just as MCAR, MAR cannot be tested statistically, but researchers must provide theoretical arguments that MAR holds. Modern missing data mechanisms, such as FIML and MI, yield consistent results under MAR (Enders, 2010).

MNAR

Missing not at random (MNAR) is defined as the conditional stochastic dependence of missingness on unobserved variables considered in the study, given the observed variables. Here, the conditional probability distribution of R, given the observed portion of Y, is *not* equal to the conditional probability distribution of missingness given both observed and unobserved portions of Y. Differently said, MNAR means that the equalities and independencies proposed in Equations 1 to 4 do not hold and instead a dependency between R and Y exists.

$$P(R \mid Y_{obs}, Y_{mis}) \neq P(R \mid Y_{obs}). \tag{5}$$

MNAR is present when there are unobserved variables that affect R that are also related to Y and thus induce dependencies between R and Y. In an applied context, one can imagine at least two distinct situations in which MNAR arises, one in which an unobserved variable affects both missingness and variables with missing data[1] or if the missing portion of a variable affects missingness on this variable directly. As all other mechanisms, MNAR cannot be statistically tested.

In Figure 1, we present a graphical display of the mechanisms. Figure 1 depicts a situation in which the analytic model contains two variables X and Y, with the latter having missing data, denoted by R_Y. In the graphical notation, R is not a matrix that contains missingness for all variables but represents missingness for a particular variable, denoted by the subscript. Disturbance terms ε represent all remaining unobserved causes of variables. In Figure 1(a) an MCAR situation is depicted in which the missingness on Y is completely unrelated to all other parts of the model. In this graph $R_Y \perp\!\!\!\perp (X, Y)$ holds, and therefore the mechanism is MCAR. In the section of Figure 1 labeled (b), missingness and the variable with missing data are not independent of each other but are related due to the two variables X and A (variable A might be a potential auxiliary variable that is not part of the analytic model). Conditioning on X and A can in this example induce independence between Y and R_Y, $Y \perp\!\!\!\perp R_Y|(X, A)$ and therefore MAR holds. Finally, section (c) of Figure 1 depicts an MNAR situation because variable L_1 is unobserved (shown as a dashed circle and not to be confused with a modeled, latent variable) and cannot be used in MI or FIML.

[1]Alternatively, an *observed* variable that has the same properties, but is *ignored* by the researcher, would cause an MNAR situation.

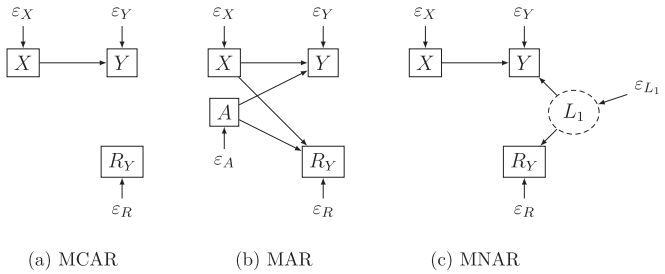

FIGURE 1 Graphical display of missing data mechanisms.

In this situation no conditional independence between Y and R_Y can be induced and MNAR holds.

CURRENT APPROACHES TO AUXILIARY VARIABLES

Although all assumptions of the missingness mechanisms are important insofar as they prescribe which methods will yield biased or unbiased estimates, MAR is an assumption that is frequently invoked by researchers who use "state of the art" methods (Schafer & Graham, 2002), for example, FIML and MI. A pertinent question is therefore how a researcher can achieve MAR or at least make MAR plausible in his or her study. One strategy that has been suggested to achieve MAR is to include so-called auxiliary variables. Auxiliary variables are observed variables that are distinguished from the variables of substantive interest in a particular model. They are added to a model only to improve estimates that pertain to analysis variables with missing data. One hopes that by including these auxiliary variables, MAR will hold, and thus consistent estimates are possible when using FIML or MI. Another reason auxiliary variables are added to a model is to reduce error variance and thus increase statistical power and precision of estimates.

In a FIML model auxiliary variables are typically added as so-called saturated correlates (Graham, 2003). Alternatively, auxiliary variables can be included in an imputation model to predict missing values on analysis variables. Inference based on the multiply imputed data is valid when the missing data mechanism is MAR and the imputation model was correctly specified.

Equations 3 and 4 also hint at the use of auxiliary variables. Both equations express that in order to obtain conditional independence between missingness and variables with missing data, information about observed variables that induce dependence between R and Y_{mis} needs to be collected. In fact, the independence statement directs us to include those variables in the imputation or FIML model that can make Y and R_Y independent of each other. Some of these variables might already be part of the analytic model; others might not be part of the analytic model but might be needed to satisfy the MAR assumption, that is, auxiliary variables. Here we describe current approaches that aim to achieve MAR.

Inclusive Approach

The so-called inclusive approach (Collins, Schafer, & Kam, 2001) to achieve MAR directs researchers to include many auxiliary variables in their model. The reasoning behind the inclusive strategy is that if many variables are included it becomes less likely that variables that are both causes of the missingness and the analytic variables with missing data are omitted. Such omission would be harmful as it would destroy the conditional independence posited in MAR and induce bias. Collins et al. (2001) showed that bias in means, variances, and regression estimates can become substantial if this kind of variable is omitted. A second rationale for adopting an inclusive strategy is that the inclusion of variables that may not be causes of the missingness or causes of the analytic variables with missing data was shown to be "far from being harmful . . . at worst neutral, and at best extremely beneficial" (Collins et al., 2001, p. 349). In particular Collins et al. examined the influence of including variables that are completely uncorrelated to missingness or analytic variables with missing data (so-called trash variables) or only related to analytic variables with missing data but not with the missingness itself. Completely uncorrelated variables did not have any impact on bias, and variables that were only correlated with Y were shown to be able to attenuate bias in MNAR situations and reduce standard errors.

Data-Driven Approach

Even if one fully acknowledges the benefits of an inclusive strategy, such a strategy can reach its limits, especially when applied to large-scale data sets, which may contain hundreds of variables. If analytic models include many variables and many auxiliary variables are added, both MI and FIML will likely encounter convergence problems. To mitigate this problem it has been suggested to examine data for the inclusion of variables as auxiliaries. Schafer (1997) suggested that variables make good candidates for auxiliary variables if they are related to the missingness or the analytic variable that exhibits missingness. The rationale behind this advice is straightforward: a variable that is completely uncorrelated with (or even independent of) the probability of missing cannot induce any dependencies between R_Y and Y. Likewise, a variable that is completely uncorrelated with the analysis variable with missing values can also not induce any dependencies between R_Y and Y. As a demonstration of this principle, consider Figure 2, in which it is of interest to obtain estimates of Y (e.g., the mean), which has missing data, indicated by R_Y. Three auxiliary variables A_1, A_2, and A_3 are available and A_1 induces dependencies between Y and R_Y. The two other variables A_2 and A_3 do not induce dependencies between Y and R_Y and are therefore

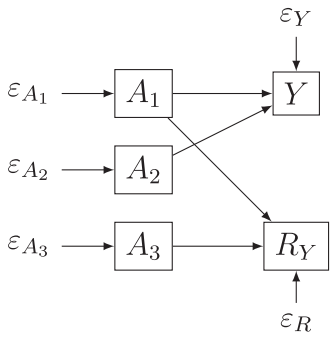

FIGURE 2 A model with several auxiliary variables. Not all of the auxiliary variables are needed for an unbiased estimate.

not needed to render Y and R_Y conditionally independent. Note that these inducing dependencies can be easily determined from the graph in Figure 2 using tracing rules (Wright, 1922) and the so-called d-separation criterion (Hayduk et al., 2003; Pearl, 2010). For more details on reading independence information from graphs with missing data see also Daniel, Kenward, Cousens, and De Stavola (2011); Thoemmes and Rose (2013); or Thoemmes and Mohan (in press).

The data-driven approach advises us to screen our set of potential auxiliary variables as to whether they are related (usually examined using correlations) with any of the analysis variables or any of the missing value indicator variables. Variables that are related to either or both should be included as auxiliary variables, whereas variables that fall below a certain correlation threshold to either should not be used. Particular guidelines on the inclusion and exclusion of auxiliary variables were formulated by van Buuren et al. (1999), who recommend including a variable if the correlation of it with either missingness or the variable with missing data exceeds $\pm$ .1 (or any other chosen threshold; e.g., Collins et al., 2001, suggest correlations with the analysis variables greater than $\pm$ .4). The implicit assumption is that variables that are correlated even lower than the chosen threshold will have little power to induce any dependencies and that variables that are correlated higher than the chosen threshold are assumed to induce biases if they are omitted from the analysis. To our knowledge, there is no empirical evidence whether any given threshold would perform better than another in any applied setting.

Generally, the advice to include auxiliary variables in missing data problems is sensible and has, in both simulation studies (Collins et al., 2001) and theoretical work (Schafer, 1997), been shown to be useful. We now present several situations in which the use of auxiliary variables surprisingly increases bias—contrary to beliefs about auxiliary variables expressed in the inclusive or data-driven approach. One possible reason that this kind of bias was not detected in previous simulation studies was that simplifying assumptions were made (e.g., all variables positively correlated, no unobserved variables) that hid bias-increasing properties of auxiliary variables.

BIAS-ENHANCING AUXILIARY VARIABLES IN THE PRESENCE OF MCAR OR MAR

As we have seen in Equation 4, MAR holds when a conditional independence between R and Y_{mis} can be achieved. Likewise, if such an independence is destroyed, the condition does not hold anymore, MNAR is present, and bias is expected. Consider now a case in which several variables obtained in Y_{obs} are observed alongside a variable A_1. Assume further that the following properties hold:

$$R \perp\!\!\!\perp Y_{mis} \mid Y_{obs} \qquad R \not\perp\!\!\!\perp Y_{mis} \mid (Y_{obs}, A_1) \tag{6}$$

In words, given a set of observed variables contained in Y_{obs}, conditional independence and thus MAR holds, but as soon as one also conditions on A_1, this conditional independence is destroyed, and therefore MNAR holds. What may sound surprising is simply a case in which variables (here R and Y_{mis}) are independent of each other but become dependent within strata of an additional variable (here A_1). How can we imagine such a variable to behave in missing data problems? In Equation 6, we can see that exclusion of A_1 would constitute an MAR situation (conditional independence holds) and unbiased estimates under FIML or MI should be obtained. On the other hand, inclusion of A_1 as an auxiliary variable would create an MNAR situation in which we would typically expect biases. In other words, if MAR held before A_1 was used as an auxiliary variable, it will be destroyed and MNAR will be induced. Every auxiliary variable that fits the definitions of Equation 6 will behave in this undesirable way.

How can we imagine a variable with such peculiar behavior in an applied context? Consider, as examples, the data-generating models presented in Figure 3. Figure 3(a) shows a situation in which the relationships from variable A_1 to Y and R_Y are both simply correlations caused by some unobserved variables, here L_1 and L_2. It is important to note that these variables are truly unobserved and cannot be used as auxiliary variables. This small example depicts a situation in which the conditions in Equation 6 hold. We can confirm this by using tracing rules to determine the covariance between Y and R_Y. The only covariance-inducing sequence of paths between Y and R_Y is via X (note that the sequence of paths via L_1, A_1, and L_2 does not contribute to the covariance between R_Y and Y according to Wright's tracing rules). Conditional on X the covariance between Y and R_Y is therefore zero, thus conditional independence and MAR hold. However, when one conditions on A_1 (meaning that it is used as an auxiliary variable when estimating parameters of the partially observed Y), a covariance between Y and R_Y is induced and MNAR holds. In the causal inference literature a variable that exhibits the characteristics of A_1 is called a collider

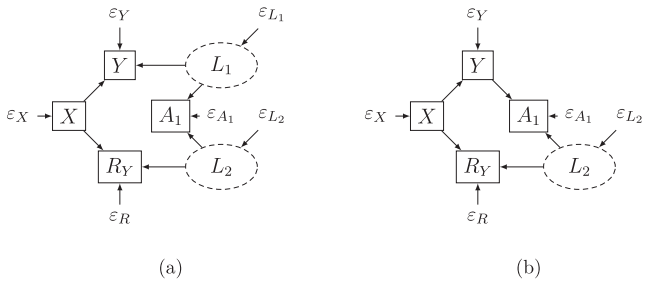

FIGURE 3 Data-generating models that are missing at random (MAR) whenever A_1 is ignored, but that are missing not at random (MNAR) whenever A_1 is included as an auxiliary variable.

variable and it is known that conditioning on A_1 induces dependencies between variables that are connected to this collider variable (Pearl, 2010). The phenomenon that conditioning on a variable can induce dependencies between other variables that were potentially independent before conditioning is also known as Berkson's paradox (Berkson, 1946) in the statistical literature.

Figure 3(b) shows a situation in which variable A_1 is directly caused by a variable with missing data, for example, Y, but at the same time A_1 is only correlated with missingness of Y due to other unobserved variables, in this case L_2. Again, MAR holds when only X is used in an FIML or MI model, but including A_1 as an auxiliary variable yields MNAR and therefore biases estimates of the partially observed Y.

To make this phenomenon a bit more concrete we can imagine that Y in Figure 3(a) is a measure of academic achievement and R_Y denotes missing values on this measure. It is of interest to estimate the mean and variance of academic achievement, Y. X may be a set of variables, such as motivation, socioeconomic status, and so on, that affect both academic achievement and missingness. A_1 on the other hand is a variable like gender. For this example, we may assume that gender does not directly affect academic achievement but may be correlated to it via unobserved variables, in this example L_1. Likewise there is no direct effect from gender to missingness, but again it is correlated with it due to unobserved variables, here L_2. In this particular example, gender would be a variable that would induce bias in the estimates of Y if used as an auxiliary variable in a FIML or MI approach. Note that if the unobserved variables L_1 and L_2 were correlated with each other, the inclusion of A_1 would still induce a dependency between Y and R_Y and thus has the potential to induce bias. However, the direction and magnitude of bias would now also depend on the correlation between L_1 and L_2. We explore biases in models as the ones presented in Figure 3 in the simulation studies that follow but also provide readers with a more formal derivation of bias in Appendix A.

Readers who are interested in additional information on collider variable bias may consult the text by Pearl (2000); the text by Morgan and Winship (2007), which includes many other examples of collider bias; or a recent special issue on this topic in the *European Journal of Personality*, in particular the articles by Asendorpf et al. (2012) and Lee (2012). For a dissenting view on the issue of collider variables see for example, Rubin (2009), who argues that bias due to collider variables may be infrequent in applied settings.

Simulation Study 1

To further demonstrate the point that inclusion of certain auxiliary variables can increase bias, we conducted a focused simulation study. The simulation study roughly followed Collins et al. (2001) in terms of data-generation and evaluation criteria. Broadly explained, data were first generated under a specific model, then missing data were imposed based on a described mechanism, then parameters were estimated using FIML models with varying numbers of auxiliary variables. The auxiliary variables were entered using the Mplus AUXILIARY command, which automatically fits a saturated correlates model in which auxiliary variables are correlated with all other variables or their residuals in case of endogenous variables. We also reran parts of our analysis using multiple imputation with and without auxiliary variables. As expected, results did not differ in any significant form; therefore we only present FIML results. Finally, results of replications were pooled within conditions and performance criteria were assessed. In this first simulation study we only focus on estimates of the population mean. Later we augment these simulations with a case study that evaluates bias in regression coefficients.

The data-generating model for Simulation 1 is shown in Figure 4. In the model, a single independent variable Y is generated with missing data, indicated by R_Y. Auxiliary variable A_1 is correlated with the probability of missing and the outcome Y via two unobserved, uncorrelated variables L_1 and L_2. In the model Y and R_Y are independent of each other but become conditionally dependent given A_1. As such, we would expect unbiased results when A_1 is ignored as an auxiliary variable, and biased results are expected when A_1 is used as an auxiliary variable.

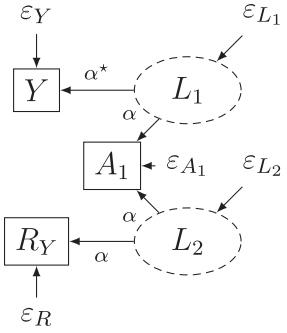

FIGURE 4 Data-generating model for Simulation 1.

All continuous variables were multivariate normally distributed and completely standardized by fixing the total variance of each variable to 1 and setting means to 0. We did not vary sample size but held it constant at 500. This single sample size was also chosen by other authors in similar simulations (Collins et al., 2001; Saris, Satorra, & Van der Veld, 2009) as a somewhat large but still reasonable sample size to consider. Furthermore, changes in sample size usually yield predictable results when other factors are held constant, namely, that standard errors decrease with increased sample size. We also did not vary the amount of missing data but fixed it at a relatively high value of 30%, which was in between the two values chosen by Collins et al. (2001). Varying the amount of missing data is often not very interesting as results of such variation have previously been shown to yield expected results (bias gets worse as missing data increases). All path coefficients in the data-generating model, labeled α, were chosen so that the uniquely explained variance in the outcome variable that these paths were connected to was set to a particular value. Path coefficients were set at 0, .224, .387, .500, .592, and .671. This corresponds to uniquely explained variance of 0%, 5%, 15%, 25%, 35%, and 45%, respectively. See Appendix B for details on how missingness was generated and how explained variance in R_Y was defined. We varied the sign of the coefficient labeled α^* (positive or negative). This sign change of a single path does not alter the magnitude of the bias that is induced but alters the direction. Note that it is not of importance which of the four paths α is varied in sign because the direction of bias is determined by the product of all four constituent paths. In this simulation design we varied all paths labeled α simultaneously, meaning that all path coefficients took on the same value within single conditions. Our primary interest was to observe overall bias and not bias due to differential changes in constituent paths. This simulation design thus yielded 5 conditions with a positive sign, 5 conditions with a negative sign, and 1 condition in which all paths were set to 0, for a total of 11 data-generating conditions. Each condition was analyzed with a FIML model that either included or excluded the auxiliary variable A_1. We replicated each condition 1,000 times. All simulations were conducted using R (R Development Core Team, 2011) and the following packages: MASS (Venables & Ripley, 2002) to generate multivariate normal random data, mice (van Buuren & Groothuis-Oudshoorn, 2011) to impute missing values, MplusAutomation (Hallquist, 2012) to automate code generation and extraction of results from Mplus, and plyr (Wickham, 2011) for general data management. For the generation of graphs we used ggplot2 (Wickham, 2009) and tikzdevice (Sharpsteen & Bracken, 2012).

Performance measures. In order to analyze the results of our simulation study, we assessed a range of standard criteria commonly employed in simulation studies.

1. We assessed standardized bias in the estimates of variables with missing data, defined identical to Collins et al. (2001) as raw bias (average parameter estimate across replications minus true parameter value) divided by the standard error, defined as the standard deviation across all replication estimates. Collins et al. give a rule of thumb that absolute values of .4 or higher are worrisome on the standardized bias metric.
2. We recorded the precision of the estimates defined as the average standard error across all replications. In general it is desirable to have estimates with smaller standard errors and hence narrower confidence intervals and more precise estimates.
3. We computed the root mean square error (RMSE) defined as the square root of the average squared difference between a parameter estimate and the true value of the parameter.
4. Finally, we observed coverage rates, defined as the percentage of replications whose 95% confidence interval included the true parameter estimate. Ideally, one observes 95% coverage rates, as this would indicate that the confidence intervals of the estimator are in the long run accurately capturing the true parameter and have the nominal Type I error rate. Again, relying on rules of thumb by Collins et al. (2001), we regard coverage rates below 90% as worrisome.

Results of Simulation Study 1. The complete results are shown in Table C1 in Appendix C. In order to communicate the most important findings, we display the amount of standardized bias in the estimated means of Y in Figure 5 and coverage values in Figure 6. Both figures show that the model without the auxiliary variable A_1 is unbiased and has perfect coverage across all conditions. The inclusion of A_1 as an auxiliary variable in the FIML estimation induced bias in the mean. Bias emerges in all conditions that used A_1 as an auxiliary variable except the one in which all paths labeled α are set to 0 (essentially an MCAR situation).

The finding that inclusion of an auxiliary variable can increase bias is somewhat contrary to advice that stems from an inclusive or data-driven approach, which would encourage inclusion of a variable like A_1 that is highly correlated to both Y and R_Y. The general pattern in Figures 5 and 6 is that increases in the amount of explained variance of A_1 in both Y and R_Y yield monotonic increases in bias. Little to no bias is observed in conditions of weak path coefficients and stronger biases are observed in more extreme conditions. The standardized bias (and other performance measures) reach a critical threshold, based on the rule of thumb by Collins et al. (2001), when path coefficients are so strong that they explain slightly less than 25% of the variance. Bias in conditions with even stronger effects is so large that confidence intervals approach 40% coverage. Finally, the RMSE also shows that bias is induced when using the auxiliary variable. The fact that

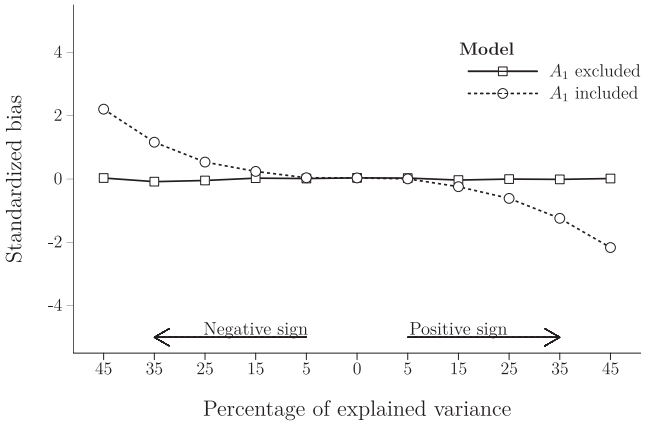

FIGURE 5 Partial results of Simulation Study 1. Standardized bias in the estimate of the mean across all conditions for both models. Arrows at the bottom of the graph display the sign of the path labeled with $\alpha*$.

RMSE rises comparatively slower than standardized bias is due to the fact that the inclusion of the auxiliary variable can reduce error variance and thus uncertainty. However, in this example this only means that the certainty around the wrong parameter estimate increases when using A_1 as an auxiliary variable. Examining average standard errors, standardized bias, and RMSE, we find little evidence that it would be beneficial to include a bias-inducing auxiliary variable just for the sake of increasing precision of the estimate. Also, as expected, the direction of bias changes with the sign of the coefficient α^*. In conditions in which the sign is negative, positive bias is induced due to the inclusion of the auxiliary variable, and negative bias is induced when the path coefficient has a positive sign, respectively. In addition, Table C1 also shows that biases in standard deviations are smaller than means, which is an expected pattern under missingness that was generated due to linear functions of covariates (Collins et al., 2001). The results of this simulation clearly show that

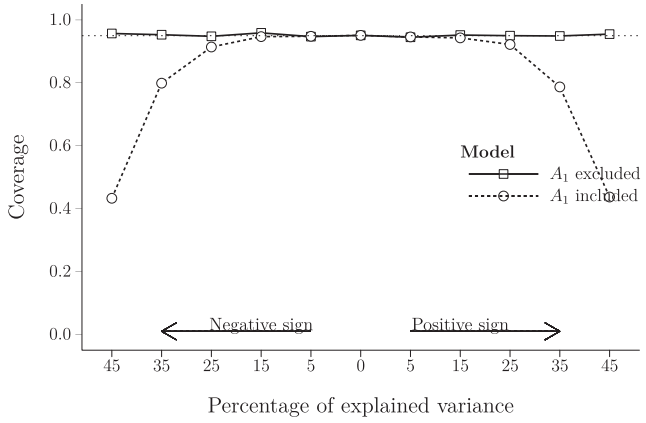

FIGURE 6 Partial results of Simulation Study 1. Coverage in the estimate of the mean across all conditions for both listwise and FIML. Arrows at the bottom of the graph display the sign of the path labeled with a $*$.

an auxiliary variable, even though it exhibits strong correlations with missingness and analysis variables, can *increase* bias.

BIAS-ENHANCING AUXILIARY VARIABLES IN THE PRESENCE OF MNAR

In our previous simulation, we examined situations in which an auxiliary variable A_1 induces MNAR. Now we consider the situation in which MNAR is already present before A_1 or any other auxiliary variable is introduced. As we shall see, in situations in which MNAR holds to begin with, bias due to auxiliary variables becomes more complicated. Recall that under MNAR Y and R_Y are dependent on each other, usually due to unobserved variables. In typical situations this dependence is of the form that some values of Y have a higher chance to be missing than others (unlike, e.g., MCAR in which every value of Y has an equal chance to be missing). Any other variable (including potential auxiliary variables) that also has a relation to Y and R_Y can also induce dependencies between these two variables. In some cases this induced dependency may offset the existing relationship between Y and R_Y, in others it may increase it. Whether or not an auxiliary variable increases or decreases bias is therefore a function of the strength and direction of the relationship of the auxiliary variable to Y and R_Y and likewise the strength and direction of the relationship of the unobserved variables that made this an MNAR situation to begin with.

To make this a bit more concrete, consider the two graphs in Figure 7. Again, these models are very simple to highlight the underlying mechanisms. In both graphs labeled (a) and (b), the inclusion of an auxiliary variable could lead to decreased or increased bias in the estimation of Y, dependent on the exact values of the path coefficients. Using Wright's tracing rules we can infer that the covariance between Y and R_Y in Figure 7(a) is a function of both products $\gamma_1\gamma_2$ and $\beta_1\beta_2$. In this example, the induced relationship due to U_1 cannot be eliminated because we cannot use the unobserved variable U_1 as an auxiliary variable. What, however, happens when

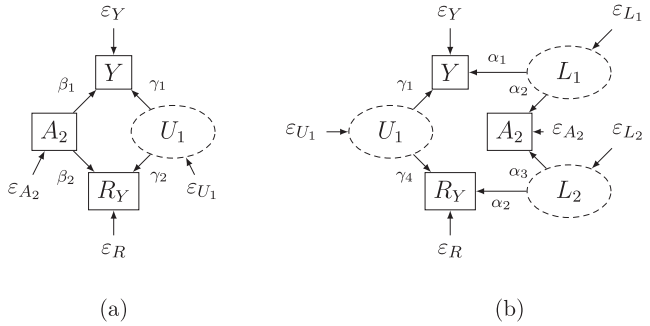

FIGURE 7 Data-generating models that are missing not at random (MNAR) and contain potential bias-inducing auxiliary variables.

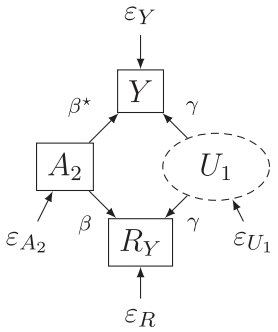

FIGURE 8 Data-generating model for Simulation 2.

we condition on A_2, or differently said, use it as an auxiliary variable? Now, the induced covariance is only due to U_1 and the associated product of paths $\gamma_1\gamma_2$. To judge whether such an inclusion of A_2 is helpful or increases bias, we need to compare which induced covariance (and thus bias) is stronger in absolute terms: either the induced covariance that is only due to $|\gamma_1\gamma_2|$ (induced covariance due to U_1 only, while using A_2 as an auxiliary variable) or due to $|\gamma_1\gamma_2 + \beta_1\beta_2|$ (induced covariance due to U_1, while A_2 is not used as an auxiliary variable). If the former term is larger than the latter, then inclusion of A_2 is expected to increase bias. This can in fact happen under a large number of situations, as every one of the four path coefficients can be positive or negative in sign. As an example, imagine that the paths labeled γ_1 and γ_2 in Figure 7(a) are both of positive sign. This implies that the unobserved variable U_1 induces an MNAR situation because participants with high values on U_1 are more likely to have high values on Y but are also more likely to be missing. Now consider that the path labeled β_1 is positive in sign, and the path β_2 is negative in sign. This implies that participants who are high on A_2 are more likely to be high on Y but less likely to be missing, thus inducing a marginal negative relationship between Y and R_Y. In such a situation the exclusion of A_2 and thus allowing the induction of a negative dependency would offset the dependency of variable U_1 and therefore reduce existing bias. In previously published simulation studies that examined behavior of auxiliary variables under MNAR, the sign of path coefficients was not varied in this way, therefore bias-enhancing properties did not become apparent. The same principles of increasing bias apply to Figure 7(b).

Simulation Study 2

In our second simulation study we explore the relationship between bias and auxiliary variables in MNAR situations. Our data-generating model is displayed in Figure 8. All aspects of the data generation, including sample size, magnitude of path coefficients, and amount of missingness, were identical to Simulation Study 1. The notable difference was that data were simulated under an MNAR scheme (indicated through the direct paths labeled γ from an unobserved variable U_1 to both Y and R_Y). The direct effects γ were always positive in sign and held constant at 20% explained variance, thus indicating a moderately strong degree of MNAR missingness. The strength of the paths labeled β was varied over the same levels as in Simulation 1, including the changing of signs of the path labeled β^*. The total number of conditions in the simulation was again 11.

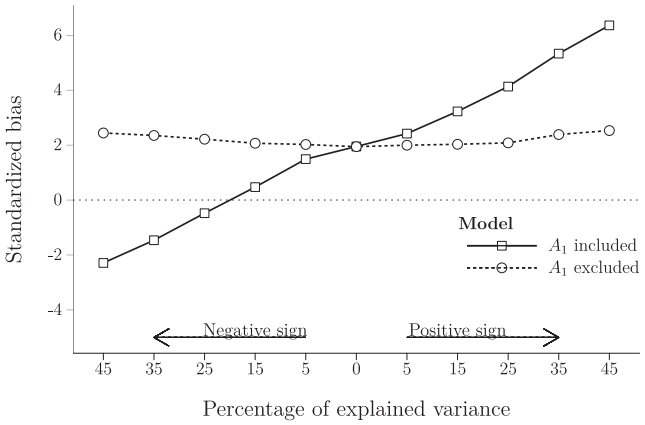

FIGURE 9 Partial results of Simulation Study 2. Standardized bias in the estimate of the mean across all conditions for both models. Arrows at the bottom of the graph display the sign of the path labeled with a $*$.

Results of Simulation Study 2. The results of Simulation Study 2 are given in full in Table C2 in Appendix C. We again present the main results in Figure 9 and Figure 10, displaying standardized bias of the estimate of the mean and coverage rates, respectively.

We observe that the model that excluded A_2 showed a pattern that consisted of regions of extreme positive bias, no bias, and some negative bias, depending on the size and magnitude of the path coefficients of the auxiliary variable A_2. In conditions in which the explained variance of the auxiliary variable was 0%, standardized bias was around 1.95. This

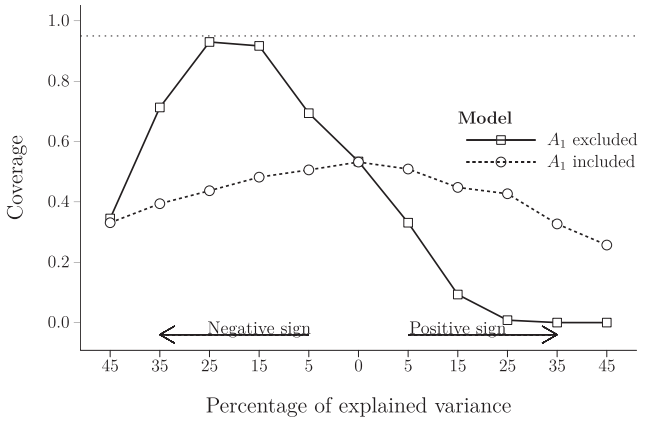

FIGURE 10 Partial results of Simulation Study 2. Coverage in the estimate of the mean across all conditions for both models. Arrows at the bottom of the graph display the sign of the path labeled with a $*$.

bias was entirely due to the presence of the unobserved variable U_1 and the resulting induced dependency between Y and R_Y. When the strength of the relationship to missingness and outcomes to A_2 was increased, the amount of bias changed, however, again dependent on the sign of the coefficient. If the sign of the β^* coefficient was positive (and therefore the induced relationship was positive) bias increased to very high levels (standardized bias larger than 6 in the most extreme conditions). If on the other hand the sign of the path coefficient β^* was negative, bias decreased, going toward zero, and then with even stronger path coefficients increased again but in the opposite direction. This pattern is also visualized in Figure 10, which shows coverage rates. We further observed that the model that excluded A_2 had coverage of around 50% in the condition in which path coefficients were set to 0. Increases in the magnitude of path coefficients while having a positive sign deteriorated coverage quickly all the way to 0%. Increasing path coefficients in the presence of a negative sign first decreased bias, and coverage rates approached the unbiased ideal of 95%, at about 25% of explained variance. In this region the bias due to the unobserved variable U_1 and the omitted auxiliary variable A_2 canceled each other out. When path coefficients became even stronger in the negative direction, the bias from omitting A_2 dominated and and we observed standardized biases of up to -2.

The model that included A_2 showed a somewhat stable amount of bias and coverage, however, with the interesting observation that bias increased in more extreme regions of explained variance. This can be seen in Figure 9 in which the line for bias for the model that included A_2 slopes slightly upward at both ends. In Figure 10 we see this pattern even clearer, as coverage rates drop from 50% at the center of the graph to around 30% at the extreme regions.

This phenomenon of residual bias amplification has been described previously in the context of instrumental variable models (Myers et al., 2011a; Myers et al., 2011b; Pearl, 2011) and it is clearly visible here in the context of missing data problems as well. In the context of instrumental variables it was shown that any bias of a relationship between two variables (in our case Y and R_Y) is amplified as soon as variables are introduced that explain variance in the putative cause . Pearl (2011) showed that bias amplification is equal to a factor of $\frac{1}{1-R^2}$, where R^2 is the explained variance of Y in our case. In our example, the inclusion of A_2 explains variance in Y and therefore any bias that is due to U_1 gets amplified monotonically, as the explained variance in Y due to A_2 increases.

Intermediate Summary of Simulation Studies

We have shown in both simulation studies that there are data situations in which auxiliary variables can increase bias when they are used in FIML or multiple imputation. These results should not give the impression that auxiliary variables are generally harmful and should be avoided. Rather, we simply showed that under certain data situations, auxiliary variables can indeed be bias inducing, which is a fact that the literature on auxiliary variables in missing data has so far ignored. This bias can emerge in situations that would be MCAR or MAR before inclusion of the bias-inducing auxiliary variable. We have highlighted when such situations could occur.

In MNAR situations the bias-increasing properties are more complex and depend on both the structure and the magnitude and direction of path coefficients. We have shown that in the presence of MNAR bias, auxiliary variables can be either bias reducing or bias inducing, depending on the sign of path coefficients connecting observed and observed variables with Y and R_Y. We have also shown that auxiliary variables that explain variance (are good predictors of the variable with missing data) can amplify existing MNAR bias.

A Simulated Case Study of Bias in Regression Coefficients

To augment our simulation studies that focused on bias in estimates of the mean, we also provide a brief simulated case study that explores biasing properties of auxiliary variables on regression coefficients. We preface this case study with references to Collins, Schafer, and Kam (2001), who showed that bias in regression coefficients tends to be smaller (when compared with bias in means) and is usually only induced by so-called sinister missing data patterns, meaning missing data patterns that are not simple (curvi)-linear functions of observed or unobserved variables.

In our example we generated data based on a model shown in Figure 11. In this example a variable X (with standard normal distribution) has an effect α on variable Y (also normally distributed with mean 0 and total variance fixed to 1). Both variables have missing data indicated by R_X and R_Y, respectively. The effect of X on Y is moderated by an unobserved variable L_1. This interactive effect is indicated by the solid circle in the diagram. It is important to note that L_1 is not a confounder, but only a moderator, and the average effect α is recovered without bias in the complete sample by the simple bivariate relationship between X and Y. Both missingness on R_X and R_Y are caused by an unobserved variable L_2. However, the presence of L_2 does not induce any bias in either means or regression coefficients as L_2 is unconditionally independent of both X and Y. Finally, there exists an auxiliary variable A_3, fully observed, caused by both L_1 and L_2.

Initially, we simulated data from this model using purely linear associations (except the interaction effect), but we realized that induced biases were quite small, which is in line with research by Collins, Schafer, and Kam (2001). Therefore, we decided for demonstration purposes to use binary variables for L_1, L_2, and A_3 and model relationships in a deterministic fashion. All deterministic paths are indicated in the graph using dashed lines. In particular, when L_1 was 0, the relationship α between X and Y was set to .5, and whenever L_1 was 1, the relationship α between X and Y was set to

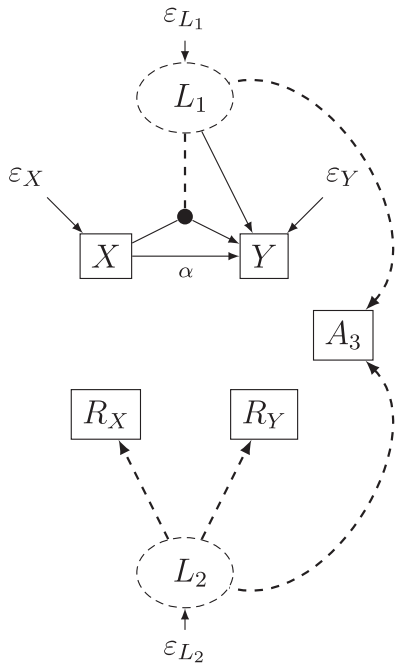

FIGURE 11 Data-generating model for simulated case study.

0. Because L_1 had a probability of .5 to be either 0 or 1, this resulted in an average effect α of .25 in the population. Further, whenever L_2 was set to 0, both R_X and R_Y were set to 1, indicating missing values on X and Y, respectively. Because L_2 had a probability of .3 to be 1, we observed on average 30% missingness on both X and Y. Finally, A_3 was set to 0 if and only if L_1 was equal to 0 and L_2 was equal to 1. In all other cases A_3 was set to 1. This particular pattern resulted in the following situation: when A_3 was not used as an auxiliary variable, an MCAR situation emerged because missingness and its causes (L_2) were completely unrelated to any other observed variable. When A_3 was used as an auxiliary variable, L_1 and L_2 became strongly related within strata of A_3, resulting in an MNAR situation. For our demonstration, we imputed values on both X and Y in strata of A_3. We chose an imputation model here because we found it easier to set up the multiple groups in this framework.

We generated samples of size 500 and performed 1,000 replications. In each replication, we analyzed data using an imputation model that excluded A_3 and an imputation model that included A_3. We multiply imputed data using the mice software in R, generating five imputations. For each replication we recorded the parameter estimate α, the standard error of α, and the confidence interval. Based on these estimates, we computed standardized bias, precision, RMSE, and coverage, as defined in the previous section on performance measures. The standardized bias for the model that excluded the auxiliary variable was .01, whereas inclusion of the auxiliary variable A_3 increased bias to -1.07, clearly indicating strong biasing effects. Precision (defined as the average standard error) was .04 in both models. RMSE was .06 and .10 for models without and with the auxiliary variable, respectively. Finally, coverage rates were somewhat below expectations at 85% for the model without the auxiliary variable [2] and poor at 57% for the model with the auxiliary variable. In summary, inclusion of the auxiliary variable induced large biases in regression coefficients. Of course, the deterministic relationships in our models represent extremely strong associations. Weaker, probabilistic relationships will yield much smaller biases that may not be substantial in any practical sense. We leave a more detailed examination of bias in regression coefficients to future studies.

APPLIED EXAMPLE

In this example we estimate a parameter of interest from an actual data set that included missing data. Our data are from the longitudinal study TOSCA (Trautwein et al., 2010) that assessed personality characteristics and many other academic variables on a sample of German youth. One of those variables was neuroticism, which we use in our applied example. For demonstration purposes we only consider the estimation of the mean and standard deviation of neuroticism at the last time point (that had the most missing data). We employ several different imputation models described here. However, the analytic model is simply an estimate of the mean and standard deviation and is identical across all imputation models.

We compare parameter estimates under a total of five different imputation models that were implemented using the mice package in R (van Buuren & Groothuis-Oudshoorn, 2011). We always generated five multiple imputations and used the predictive mean matching algorithm to impute missing values. All models differ in their assumptions about which auxiliary variables are needed to satisfy MAR and consequently use different sets of auxiliary variables.

In Table 1 we present all auxiliary variables that we consider for this example. We provide information on number of available cases, means, standard deviations, and display (pairwise-deleted) correlations of auxiliary variables with (a) our outcome of interest (neuroticiscm) and (b) the missingness vector of our outcome of interest. The first imputation model serves as a baseline for comparison and includes no auxiliary variables. This model assumes that MAR holds without any auxiliary variables—usually an assumption that is not plausible. The second model includes two demographic variables (age, socioeconomic status). We can see in Table 1 that both variables are correlated with the variable that has missing data and with missingness itself—although correlations are relatively small in magnitude. The assumption is that MAR holds with this sparse set of loosely related

[2]We suspect that chance correlations between L_1 and L_2 were disruptive of coverage because deterministic relationships emanating from L_2 and L_1 were very strong.

TABLE 1
Means, Standard Deviations, and Correlations of Auxiliary Variables. Correlations for the Analytic Variable of Interest (Neurotiscism) and Missingness on Neuroticism Are Shown

	n	M	SD	r_Y	r_{RY}
Age	9,626	19.5	.77	.04*	.05***
SES	9,419	58.3	15.7	−.04*	−.03***
Neuroticism Wave 1	6,641	2.27	.30	.59***	−.03*
Neuroticism Wave 2	3,708	2.16	.50	.70***	−.04*
Extraversion Wave 1	6,641	2.85	.39	−.28***	.00
Openness Wave 1	6,629	2.73	.43	.08***	−.02
Agreeableness Wave 1	6,631	2.92	.36	−.13***	−.07***
Conscientiousness Wave 1	6,637	2.87	.45	−.17***	−.08***

Note. SES = Socioeconomic status.
*$p < .05$. **$p < .01$. ***$p < .001$.

variables—again, this is not very plausible. The third model includes two pretest measures of neuroticism (assessed at two previous time points). Not surprisingly, both pretests are highly correlated with the outcome of interest and weakly correlated with missingness. The MAR assumption seems somewhat more plausible if we believe that no other unobserved variables affect neuroticism and missingness on neuroticism, over and above the already collected pretests. The fourth model uses both the two pretests and the two demographic variables as auxiliary variables. Finally, the fifth model contains all variables from previous models but also adds four additional personality characteristics (extraversion, openness, conscientiousness, agreeableness), all measured at the first time point. Some of the personality characteristics were moderately related to our outcome of interest (neuroticscm), but relationships to missingess were generally weak (although often statistically significant due to large sample size). One may think that this larger model makes MAR more plausible, but as we have shown in our simulations, it could also be that variables exhibit bias-inducing properties, especially if they are assumed to be only correlated to analysis variables via additional unobserved variables. In this final imputation model we might assume that the included variables are all needed to induce conditional independence and that no bias-inducing auxiliary variables are present.

A very important consideration for this example is that we do *not* know the true mean value of neuroticism if it had been fully observed, and it is *impossible* to claim that any estimate of our five models is closer to the truth than another. The only claim that can be made by this example is that different models may potentially yield different results and that someone might be more convinced by one set of auxiliary variables to fulfill conditional independence assumptions over another set.

Our analytic sample consisted of a total of 3,062 participants with observed data on our focal variable "neuroticism" and 7,850 participants who had missing data on this variable, thus a total sample of 10, 912. The missing data rate was quite high at 72%. The auxiliary variables at our disposal also had partially missing data. Participants who were missing on every single variable were excluded from the analysis. We estimated the mean and standard deviation for the focal variable "neuroticism" at Wave 3 in the multiple imputed data sets using Rubin's rules (Rubin, 2009). The pooled parameter estimates are displayed in Table 2. In this example, estimates from the listwise model and estimates from the model that only included demographics were virtually identical. This may not be very surprising as demographic variables were only very weakly related to neuroticism. The model that included pretests changed the mean slightly, but estimated standard deviations remained unchanged. The model that included both demographics and pretests, not surprisingly yielded means that were virtually identical to the model that only included pretests. Finally, the large model that included all previous variables, but also several other personality characteristics yielded means that were again very similar to previous models but had very slightly smaller standard deviations.

TABLE 2
Results of Applied Example Showing Mean and Standard Deviation of Focal Variable

	n	M	SD
No auxiliary variables	10,912	2.12	.53
Demographics only	10,912	2.12	.53
Pretests only	10,912	2.10	.53
Demographics and pretests	10,912	2.11	.53
Inclusive model	10,912	2.11	.52

Note. Sample size is constant because data has been imputed.

Overall, a pattern emerged that suggested that across different imputation models with different choices of auxiliary variables the results were very stable and similar to each other. Differently said, in this particular data set, there was a certain robustness in results with regard to the specification of auxiliary variables. This robustness of results with regard to choice of auxiliary variables has been observed by others in applied contexts (Mustillo, 2012). Given that the amount of missing data was quite large, this is a somewhat reassuring result. An example in which means differed drastically would have been arguably more interesting but probably also less typical in an applied context. Furthermore, one also has to consider that *standardized* mean differences between the models were as large as .05, which is still very modest but not unusual to be seen published in articles on changes in personality characteristics. As mentioned earlier, it is impossible to tell from this example which of the models is actually superior in terms of being closer to the true value of neuroticism.

DISCUSSION

The overarching picture that emerged from our study is that there are situations in which auxiliary variables can induce

bias or increase existing bias. We have demonstrated through focused simulations that auxiliary variables can increase biases under different missingness mechanisms, in some conditions substantially so. We showed that bias can be induced in both means and regression coefficients.

At the same time it needs to be acknowledged that in conditions in which associations of auxiliary collider variables with other variables in the model were small, resulting bias due to using an auxiliary collider variable was likewise small. Especially, when considering regression coefficients, only strong associations in the data-generating model induced sizeable amounts of bias.

The fact that in some conditions bias was rather small implies that even though we can theoretically show that bias emerges when using a particular kind of auxiliary variable, we are less certain as to how large biases in real-world settings will be. This raises the question whether applied researchers need to worry at all about bias due to using an auxiliary collider variable, as the resulting bias may be small in practice. We find this a difficult question to answer, as it is often unclear in an applied setting how much variance in variables of interest is explained by variables that are part of the collider structure. In our simulations we find that the bias due to using a collider auxiliary variable becomes large once explained variance is at around 25%, a rather high value. At the same time, if *many* variables with collider properties are considered, one may approach such large amounts of explained variance. Finally, we refer to the literature in the domain of causal inference that has identified collider bias as worthy of consideration; e.g., Asendorpf et al. (2012); Morgan and Winship (2007); Pearl (2009).

What do these results imply for the theory and practice of choosing auxiliary variables in applied settings? First, it encourages applied researchers to think more carefully about which auxiliary variables should be entered in a multiple imputation or FIML algorithm. At the same time it discourages blind reliance on simply using all available auxiliary variables or checks on correlations that do not necessarily lead to the best possible choice of auxiliary variables. Both approaches may pick auxiliary variables that we have shown to be bias inducing. But how can the practice of choosing auxiliary variables be improved? One possible attempt is to carefully think about the underlying structural relationships of auxiliary variables and make some assumptions about them. Any variable that is suspected to exhibit the bias-inducing collider properties that we outlined earlier should be regarded with caution. It may in fact be helpful for applied researchers to draw a graphical model in which the best available knowledge and assumptions are laid out and then—from this graph—determine if a variable will induce a covariance between Y and R_Y and thus induce bias when used as an auxiliary variable. We quickly caution, however, that if the assumptions that went into a graphical model are incorrect, the conclusion regarding the choice of auxiliary variables will likewise be incorrect in most circumstances.

We also note that there is a certain parallel of selection of auxiliary variables to questions of variable selection in causal models. The nonrandom and unobserved selection of participants into treatment or control groups is similar to the nonrandom and unobserved selection of participants into being either observed or unobserved. Neither of those situations are under the control of the researcher. There is a wide literature and debate on which variables should ideally be selected to estimate a causal effect if participants self-select into treatment conditions (Pearl, 2009; Rubin, 2009). A minimal consensus in terms of selection of variables in causal models is that their choice should be theoretically motivated and that some considerations as to which processes lead participants into one of the two groups should be considered before data collection. Ideally, variables that model this nonrandom selection process should then be assessed. Applying this idea to the missing data context means that researchers should deliberately select covariates (even before data collection) that might help explain the mechanisms through which participants become either observed or unobserved. This differs from common practice, where auxiliary variables are chosen based on statistical evidence when data are already collected. In other words, the findings of our study suggest that the careful and theoretically driven covariate selection should not be limited to the analytic model but should include the missing data model as well. This can strengthen the tenability of MAR assumptions especially in studies that typically suffer from large amounts of missing data, such as low-stakes assessments.

Our main goal was to add a note of caution to theoretical and applied researchers that situations exist in which auxiliary variables can have the surprising property of increasing bias due to missing data. It was not our goal, and can in fact not be inferred from our results, whether any given auxiliary variable in any particular data situation should be included or excluded. To make such a decision would require knowledge about the structure of relationships and magnitude and sign of path coefficients—knowledge that will hardly ever be at the disposal of applied researchers.

Limitations

Several objections might be raised to the concept of bias-inducing auxiliary variables that we have presented here. First, one might question whether these simple structures that we have displayed accurately capture more complex patterns in data. We completely agree that our simulations could be potentially expanded to include many more auxiliary variables that are interconnected in complex ways. However, our goal was simply to present that biases due to inclusion of auxiliary variables exist. Even if data constellations were more complex, as soon as we found within this complex set of variables an auxiliary variable that has the same structure as the variables we considered (and fulfills the conditions in Equation 6), bias would emerge. The same holds true for

the MNAR situation that we considered. Even if hundreds of observed and unobserved variables were simulated, the observed bias would still be a function of the totality of all inducing relationships between Y and R_Y—in some instances they may add up to produce even larger biases; in other cases they may cancel out and produce smaller biases.

Two astute reviewers also pointed out that bias may be reduced as soon as one uses additional auxiliary variables that are correlated with the collider auxiliary variable. This can in fact often be true but must not be necessarily true. Correlation with the collider auxiliary variable is not a sufficient condition to reduce bias (Pearl, 2000).

Other limitations include the fact that we only considered a single case study when exploring bias in regression coefficients. Clearly, the simulations could be extended to pinpoint exactly when bias in regression coefficients becomes large; however, we conjecture that bias in regression coefficients tends to be smaller unless missingness follows "sinister" patterns with strong associations.

Future Directions

The Limitations section points in the direction of future research. It will be interesting to consider more complex situations of missing data and auxiliary variables. Ultimately, it would be very valuable to generate data that is so complex and includes so many diverse causes of missingness that it potentially approximates a real data situation—even though this might be challenging. Such a simulation could be used in conjunction with a rigorous test of the different approaches to select auxiliary variables to judge which one performs better. Ideally, it would help to inform which approach to auxiliary variable selection might be preferred under which circumstances. Our conjecture is that there will always be situations in which an inclusive or data-driven approach will select bias-inducing auxiliary variables, as they cannot be distinguished from helpful variables based on correlations alone. At the same time there will be a large number of circumstances in which the "optimal" selection of auxiliary variables (based on the actual data-generating mechanism) will produce very similar results as an inclusive or data-driven approach.

In summary, we believe that there is still a lot to be learned about the selection of auxiliary variables in missing data and we hope that we have provided a glimpse into some surprising aspects of auxiliary variables.

ACKNOWLEDGMENTS

We thank the participants of the colloquium of the Methodology Center at the Pennsylvania State University and Judea Pearl for helpful discussions.

REFERENCES

Asendorpf, J. B., Rindermann, H., Woodley, M. A., Stratford, J., Rabaglia, C., Marcus, G., & Lane, S. (2012). Bias due to controlling a collider: A potentially important issue for personality research. *European Journal of Personality*, *26*, 391–413.

Berkson, J. (1946). Limitations of the application of fourfold tables to hospital data. *Biometrics Bulletin*, *2*, 47–53.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–253.

Collins, L., Schafer, J., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351.

Daniel, R., Kenward, M., Cousens, S., & De Stavola, B. (2011). Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, *21*, 243–256.

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

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (Vol. 1). New York, NY: Cambridge University Press.

Graham, J. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 80–100.

Hallquist, M. (2012). *Mplusautomation: Automating Mplus model estimation and interpretation* [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=MplusAutomation

Hayduk, L., Cummings, G., Stratkotter, R., Nimmo, M., Grygoryev, K., Dosman, D.,. .. Boadu, K. (2003). Pearl's d-separation: One more step into causal thinking. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 289–311.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, *95*, 49–70.

Lee, J. J. (2012). Correlation and causation in the study of personality. *European Journal of Personality*, *26*, 372–390.

Little, R. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198–1202.

MacCallum, R., Zhang, S., Preacher, K., & Rucker, D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40.

Mohan, K., Pearl, J., & Tian, J. (2013). Graphical models for inference with missing data. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in neutral information processing systems 26* (pp. 1277–1285). Red Hook, NY: Curran Associates, Inc.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, *26*, 67–82.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. New York, NY: Cambridge University Press.

Mustillo, S. (2012). The effects of auxiliary variables on coefficient bias and efficiency in multiple imputation. *Sociological Methods & Research*, *41*, 335–361.

Myers, J., Rassen, J., Gagne, J., Huybrechts, K., Schneeweiss, S., Rothman, K., . . . Glynn, R. (2011a). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, *174*, 1213–1222.

Myers, J., Rassen, J., Gagne, J., Huybrechts, K., Schneeweiss, S., Rothman, K., & Glynn, R. (2011b). Myers et al. respond to "Understanding Bias Amplification." *American Journal of Epidemiology*, *174*, 1228–1229.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York, NY: Cambridge University Press.

Pearl, J. (2009). Letter to the editor: Remarks on the method of propensity score. *Statistics in Medicine*, *28*, 1415–1424.
Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology*, *40*, 75–149. doi:10.1111/j.1467-9531.2010.01228.x
Pearl, J. (2011). Invited commentary: Understanding bias amplification. *American Journal of Epidemiology*, *174*, 1223–1227.
Pearl, J. (2013). Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference*, *1*, 155–170.
Pearl, J. (2014). Understanding Simpson's paradox. *The American Statistician*, *68*, 8–13.
Peugh, J., & Enders, C. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*, 525–556.
R Development Core Team. (2011). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org
Raghunathan, T., Lepkowski, J., Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*, 85–95.
Raykov, T. (2011). On testability of missing data mechanisms in incomplete data sets. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 419–429.
Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.
Rubin, D. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*, 473–489.
Rubin, D. (2009). Author's reply: Should observational studies be designed to allow a lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, *28*, 1420–1423.
Saris, W., Satorra, A., & Van der Veld, W. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 561–582.
Schafer, J. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman & Hall/CRC.
Schafer, J. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*, 3–15.
Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147.
Sharpsteen, C., & Bracken, C. (2012). *tikzdevice: A device for R graphics output in pgf/tikz format* [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=tikzDevice (R package version 0.6.2)
Thoemmes, F., & Mohan, K. (in press). Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal*.
Thoemmes, F., & Rose, N. (2013). Selection of auxiliary variables in missing data problems: Not all auxiliary variables are created equal (Tech. Rep. No. R-002). Cornell University.
Trautwein, U., Neumann, M., Nagy, G., Lüdtke, O., & Maaz, K. (2010). *Schulleistungen von Abiturienten: Die neu geordnete gymnasiale Oberstufe auf dem prüfstand* [School performance of high-school graduates: Testing the new ordered upper school]. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.
van Buuren, S., Boshuizen, H., & Knook, D. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*, 681–694.
van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. Retrieved from http://www.jstatsoft.org/v45/i03/
Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4
Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer. Retrieved from http://had.co.nz/ggplot2/book
Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*, 1–29. Retrieved from http://www.jstatsoft.org/v40/i01/
Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist*, *56*, 330–338.

APPENDIX A

Detailed Example With Application of Path-Tracing and d-Separation Rules

We present a worked-out applied example in which we also demonstrate how to use the rules of d-separation and path-tracing rules to derive model-implied relationships between variables. Readers interested in more examples applying these rules to linear models are referred to Pearl (2013) and in the context of missing data to Daniel, Kenward, Cousens, and De Stavola (2011); Mohan, Pearl, and Tian (2013); and Thoemmes and Mohan (in press).

For this example, we consider the model from Figure 3(a), with the only exception that the unobserved variables L_1 and L_2 are considered as observed variables. This does not change any of the results but only broadens the potential set of auxiliary variables. In the example it is of interest to estimate parameters of the partially observed variable Y. All other variables in the model are potential auxiliary variables that are at the disposal of the researcher. The data-generating model is displayed in Figure A1.

As outlined in the article, any bias due to missing data can be traced back to violations of independence between Y and R_Y. Consequently, one may use tracing rules to determine the model-implied covariance between Y and R_Y. If Y and R_Y are independent of each other, no bias will emerge. If dependencies exist, bias will be observed. The magnitude of the bias is a direct function of the magnitude of the induced covariance.

Wright's tracing rules state that in order to obtain the model-implied covariance between two variables one must sum up all routes between these two variables satisfy three criteria: (a) one may not trace a loop (no passing through the same variable twice), (b) one may not go forward and then backward (only going backward and then forward is allowed), and (c) one may not trace through more than one bidirected arrow. The more general rules of d-separation advise researchers to list all routes that are not loops and then decide whether they are open or closed, meaning whether or not they contribute to the model-implied covariance. All paths that can be identified with Wright's tracing rules are open and contribute to the model-implied covariance—all other paths are closed and do not contribute to the model-implied covariance. However, d-separation further tells us what happens to

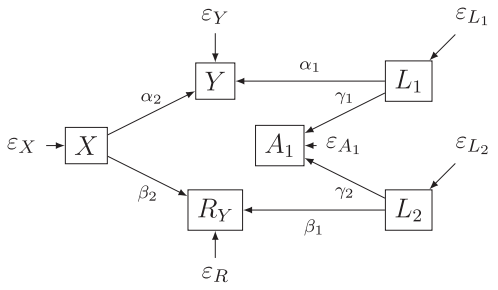

FIGURE A1 Data-generating model for detailed example.

the model-implied covariance when we condition on a particular variable in a route. An open route becomes closed once we condition on any variable in the route that is not a collider (i.e., has two arrows pointing in it). On the other hand, a route becomes open when we condition on a collider (and there are no other colliders in the route). Again, a more complete description of d-separation is given by Hayduk et al. (2003).

We begin by considering the unconditional covariance between Y and R_Y. The only open route between these two variables is via $Y \leftarrow X \rightarrow R_Y$. We therefore know that using no auxiliary variables will result in biased results. Specifically the bias is a function of the paths that connect Y and R_Y. This can be expressed as $B(\emptyset) = f(\alpha_2\beta_2)$, where B stands for bias due to induced dependency, followed by the set of conditioning variable (in this case none, indicated by $\emptyset$). Finally, f simply stands for function. The actual derivation of the numerical amount of bias is more complex as at least one of the coefficients is a logistic regression coefficient and bias is further amplified as a function of explained variance in Y.

Imagine now that an applied researcher considers using any set of variables X, A_1, L_1, L_2 as possible auxiliary variables when estimating parameters of the partially observed Y. For some reason, she chooses A_1 and X as auxiliary variable. The sequence of paths $Y \leftarrow X \rightarrow R_Y$ that was previously open and inducing bias is now closed due to using X as an auxiliary variable; however, the sequence $Y \leftarrow L_1 \rightarrow A_1 \leftarrow L_2 \rightarrow R_Y$ is now open due to using A_1 as an auxiliary variable. Therefore expected bias now becomes $B(A_1, X) = f(-\alpha_1\gamma_1\gamma_2\beta_1)$. The negative sign in the aforementioned equation is based on the fact that when conditioning on a collider, the sign of the induced covariance is reversed in sign (Pearl, 2010).

Using these rules reveals that bias due to missingness emerges in the presence of some sets of auxiliary variables but not in others. One can enumerate all possible combinations:

$$\begin{aligned}
B(\emptyset) &= f(\alpha_2\beta_2)\\
B(X) &= 0\\
B(L_1) &= f(\alpha_2\beta_2)\\
B(L_2) &= f(\alpha_2\beta_2)\\
B(A_1) &= f(\alpha_2\beta_2 - (\alpha_1\gamma_1\gamma_2\beta_1))\\
B(X, L_1) &= 0\\
B(X, L_2) &= 0\\
B(X, A_1) &= f(-\alpha_1\gamma_1\gamma_2\beta_1)\\
B(L_1, L_2) &= f(\alpha_2\beta_2)\\
B(L_1, A_1) &= f(\alpha_2\beta_2)\\
B(L_2, A_1) &= f(\alpha_2\beta_2)\\
B(X, L_1, L_2) &= 0\\
B(X, L_1, A_1) &= 0\\
B(X, L_2, A_1) &= 0\\
B(A_1, L_1, L_2) &= f(\alpha_2\beta_2)\\
B(X, L_1, L_2, A_1) &= 0.
\end{aligned} \tag{7}$$

As evident from the aforementioned equations, in this simple example bias will emerge when (a) no auxiliary variable is used, (b) X is not part of the set of auxiliary variables, and (c) A_1 is part of the set of auxiliary variables *and* at the same time neither L_1 nor L_2 are part of the set of auxiliary variables. In all other instances, no bias will emerge. Of course, different data-generating models will yield different outcomes; for example, Pearl (2014) presents an example in which sequential conditioning can induce, then reduce, and then induce bias again. Finally, this example also shows nicely that knowledge of correlations alone is not sufficient to find a set of auxiliary variables that minimizes bias. In the aforementioned example all four auxiliary variables were correlated with both Y and R_Y. However, only some sets of auxiliary variables remove bias (in this example the complete set of variables removes bias, but we could have easily generated data in which only a smaller subset removes bias). Further, this example also shows that if a bias-inducing auxiliary variable like A_1 has been selected, it is again insufficient to know which other variables are correlated with it to determine which of these variables might nullify the bias due to a variable like A_1. L_1 and L_2 can nullify bias due to using A_1 as an auxiliary variable, whereas X leaves the bias that is due to A_1 ($f(-\alpha_1\gamma_1\gamma_2\beta_1)$) completely unchanged.

APPENDIX B

Generation of Missing Values and Explained Variance in R_Y

In all of our data generations we modeled the relationship between predictor variables and missingness by modeling a latent, continuous variable that expresses the likelihood of being missing, given values on variables that predict missingness. This allowed us to use the same path coefficients and model the same amount of explained variance. This latent variable is not displayed in our graphs to make the visualization of the underlying missingness mechanism clearer. Paths going into the latent, continuous variable had the same magnitude and explanatory power as paths from variables going into the variable with missing data, hence they are also displayed with the same letter α in our graphs. To generate missing data, we created a binary indicator based on the latent missingness propensity by performing a cut at the 30th percentile of the underlying continuous variable. We fully acknowledge that this dichotomization results in amounts of explained variance in the binary variable that are nominally lower than the ones that were specified with regard to the latent continuous variable. We examined this attenuation and found in line with previous research (Cohen 1983; MacCallum, Zhang, Preacher, & Rucker, 2002) that the attenuation factor is constant across our simulation conditions as long as the dichotomization always occurs at the same percentile. We also reran our simulations and modeled the

binary missingness indicator directly, choosing logistic regression coefficients that map on to the same values on the McKelvey-Zavoina Pseudo-R^2. Results from these studies showed a very similar pattern, with biases across all conditions being slightly higher due to the absence of any attenuation in path coefficients due to a median split. The only reason we did not employ the approach of directly modeling the binary response was that it becomes exceedingly hard to get the exact desired Pseudo-R^2 in models with several, potentially correlated predictors. This is a result of the fact that logistic regression coefficients are affected by other predictors even if the are completely uncorrelated with each other. This lesser known point about logistic regression is explained in more detail by Mood (2010).

APPENDIX C

Detailed Tables

TABLE C1
Results of Simulation Study 1

	Sign of Coefficient		Negative 45%	35%	25%	Unique Explained Variance in Each Path α 15%	5%	0%	Positive 5%	15%	25%	35%	45%
		Std. bias	0.03	−0.08	−0.05	0.03	0.02	0.04	0.03	−0.03	0.00	−0.01	0.02
	μ_y	SE	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
		RMSE	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.05
A_1 excluded		Coverage	0.96	0.95	0.95	0.96	0.95	0.95	0.94	0.95	0.95	0.95	0.96
		Std. bias	−0.02	−0.07	−0.03	−0.03	−0.02	−0.04	−0.13	−0.02	−0.04	0.02	−0.10
	σ_y	SE	0.08	0.08	0.08	0.08	0.08	0.08	0.07	0.08	0.08	0.08	0.08
		RMSE	0.08	0.07	0.07	0.08	0.07	0.08	0.08	0.07	0.08	0.08	0.07
		Coverage	0.95	0.96	0.95	0.94	0.95	0.93	0.94	0.96	0.95	0.95	0.95
		Std. bias	2.21	1.16	0.54	0.24	0.04	0.03	0.01	−0.24	−0.61	−1.24	−2.17
	μ_y	SE	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
		RMSE	0.12	0.08	0.06	0.05	0.05	0.05	0.06	0.06	0.06	0.08	0.12
A_1 included		Coverage	0.43	0.80	0.91	0.95	0.95	0.95	0.95	0.94	0.92	0.79	0.44
		Std. bias	0.31	0.05	0.00	−0.02	−0.02	−0.04	−0.13	−0.02	−0.01	0.13	0.24
	σ_y	SE	0.08	0.08	0.08	0.08	0.08	0.08	0.07	0.08	0.08	0.08	0.08
		RMSE	0.08	0.07	0.07	0.08	0.07	0.08	0.08	0.07	0.08	0.08	0.08
		Coverage	0.95	0.96	0.95	0.94	0.95	0.93	0.94	0.95	0.95	0.96	0.95

Note. Results are broken up by model (A_1 excluded, A_1 included), parameter estimate (mean μ or variance σ), type of performance measure (standardized bias [Std. bias], standard error [SE], root-mean square error [RMSE], coverage), and across columns, the sign and magnitude of the relationships on the R^2 metric.

TABLE C2
Results of Simulation Study 2

	Sign of Coefficient		Negative 45%	35%	25%	Unique Explained Variance in Each Path α 15%	5%	0%	Positive 5%	15%	25%	35%	45%
		Std. bias	−2.29	−1.46	−0.48	0.47	1.49	1.95	2.42	3.23	4.13	5.33	6.36
	μ_y	SE	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
		RMSE	0.14	0.09	0.06	0.06	0.09	0.11	0.14	0.18	0.23	0.28	0.33
A_1 excluded		Coverage	0.34	0.71	0.93	0.92	0.69	0.53	0.33	0.09	0.01	0.00	0.00
		Std. bias	−0.47	−0.16	−0.05	−0.02	−0.17	−0.26	−0.49	−0.94	−1.49	−2.34	−3.68
	σ_y	SE	0.07	0.07	0.08	0.08	0.07	0.07	0.07	0.07	0.07	0.06	0.06
		RMSE	0.08	0.07	0.08	0.08	0.08	0.08	0.08	0.10	0.12	0.17	0.23
		Coverage	0.90	0.94	0.94	0.95	0.94	0.93	0.90	0.81	0.64	0.33	0.08
		Std. bias	2.44	2.35	2.21	2.07	2.02	1.94	2.00	2.03	2.08	2.38	2.53
	μ_y	SE	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
		RMSE	0.14	0.13	0.12	0.12	0.12	0.11	0.12	0.12	0.12	0.14	0.14
A_1 included		Coverage	0.33	0.39	0.44	0.48	0.51	0.53	0.51	0.45	0.43	0.33	0.26
		Std. bias	1.14	0.86	0.43	0.16	−0.15	−0.27	−0.47	−0.79	−1.07	−1.49	−2.02
	σ_y	SE	0.09	0.08	0.08	0.08	0.07	0.07	0.07	0.07	0.07	0.07	0.07
		RMSE	0.13	0.11	0.09	0.08	0.08	0.08	0.08	0.09	0.11	0.13	0.16
		Coverage	0.84	0.91	0.93	0.95	0.94	0.93	0.90	0.85	0.76	0.62	0.44

Note. Results are broken up by model (A_1 excluded, A_1 included), parameter estimate (μ mean or variance σ), type of performance measure (standardized bias [Std. bias], standard error [SE], root-mean square error [RMSE], coverage), and across columns, the sign and magnitude of the relationships on the R^2 metric.