2023 ■ Vol. 19 ■ no. 2

How to generate missing data for simulation studies

Xijuan Zhang [a]

[a]York University

Abstract ■ Missing data are common in psychological and educational research. With the improvement in computing technology in recent decades, more researchers have begun developing missing data techniques. In their research, they often conduct Monte Carlo simulation studies to compare the performances of different missing data techniques. During such simulation studies, researchers must generate missing data in the simulated dataset by deciding which data values to delete. However, in the current literature, there are limited guidelines on how to generate missing data for simulation studies. Our paper is one of the first that examines ways of generating missing data for simulation studies. I emphasize the importance of specifying missing data rules which are statistical models for generating missing data. I begin the paper by reviewing the types of missing data mechanisms and missing data patterns. I then explain how to specify missing data rules to generate missing data with different mechanisms and patterns. I emphasize the advantages and disadvantages of using different missing data rules and algorithms to generate missing data for simulation studies. Next, I discuss other important aspects of simulation studies involving missing data. I end the paper by offering recommendations for generating missing data for simulation studies.

Keywords ■ Missing Data; Incomplete Data; Simulation Studies; Creating Missing Data; Generating Missing Data. **Tools** ■ R.

✉ cathyxijuan@gmail.com

10.20982/tqmp.19.2.p100

Acting Editor ■ Roland Pfister (Universität Würzburg)

Reviewers
■ Peida Zhan (Zhejiang Normal University)

■ and one anonymous reviewer.

Introduction

Missing data are prevalent in many psychological and educational research studies, particularly those where questionnaires are used to collect data and where participants are studied over a period of time. Historically, statistical analysis methods are developed assuming no missing data, and statistical techniques for handling missing data are hard to implement due to intensive computation. However, with the advance of computing technology, beginning in the late 1980s, the problem of missing data began to receive a lot of attention. In recent decades, more and more research articles have studied statistical techniques for dealing with missing data. Two of the most commonly studied modern missing data techniques are the full information maximum likelihood (FIML; Arbuckle, 1999; Schafer & Graham, 2002b) and multiple imputation (MI; Little & Rubin, 2019); a relatively less popular method is the two-stage (TS) method (Savalei & Bentler, 2005; Yuan & Bentler, 2000).

In addition, due to the increase in computing power, Monte Carlo simulation studies have become routinely used by researchers to evaluate different statistical techniques. In typical simulation studies, researchers first specify the population parameters and distribution, then generate sample data from the population distribution they specified, and finally, analyze the data using different statistical techniques (Morris et al., 2019). With simulation studies, researchers can compare the effectiveness of different statistical techniques since they know the true population parameters from which they generate the sample data; therefore, simulation studies have become a valuable tool for comparing different existing statistical techniques or for studying new statistical techniques (Morris et al., 2019). Due to the importance of simulation studies, how to generate data for simulation studies and how to design a good simulation study have become research topics of their own. For example, many researchers (e.g., Fleishman, 1978; Foldnes & Olsson, 2016; Mattson, 1997; Olvera Astivia & Zumbo, 2015; Reinartz et al., 2002) have examined different ways of generating normal and non-normal data, and have provided recommendations for conducting simulation studies involving normal and non-normal data.

In the context of conducting simulation studies for studying missing data techniques such as FIML and MI, researchers not only have to generate sample data based on the specified population parameters but also need to generate missing data in the sample data. In other words, researchers must decide which values in the sample data should be deleted in order to create missingness in the dataset; after generating missing data, researchers can then analyze the incomplete data to compare different missing data techniques such as FIML and MI. [1] However, unlike the research on generating complete normal or non-normal data, there has been almost no research that examines ways of generating missing data and offers recommendations for simulation studies. The current paper is one of the first papers that address this gap of research.

The main purpose of our paper is to explain the different ways of generating missing data with different properties that are important for simulation studies, with a focus on explaining the statistical modelling behind generating missing data. To design a good simulation study involving missing data, researchers must first understand the modelling behind missing data generation and systematically manipulate different properties of the missing data. However, in the current missing data literature, most simulation studies' designs were done haphazardly, usually based on what past simulation studies had done. Particularly, most simulation studies were not designed with the missing data generation modelling in mind and do not systematically vary the properties of the missing data, thus creating confounds in the results of the simulation studies. In addition, researchers are usually unaware of how different computer algorithms for generating missing data affect the properties of the missing data, making it hard for them to design a good simulation study involving missing data. In short, when researchers want to design a simulation study involving missing data, they are faced with a variety of decisions and challenges, often unsure how to vary one missing data property while holding the other properties constant. As a quantitative psychology researcher who has published several papers involving simulation studies with missing data, I have experienced many of these challenges myself. The current paper will address these challenges of generating missing data for simulation studies. The current paper mainly targets those researchers who wish to conduct simulation studies with missing data, including those who plan to do simulation studies with missing data for the first time and those who have some prior experience with it. Furthermore, my paper can also help any students or researchers who are interested in learning about missing data because understanding the modelling behind generating missing data will help gain a deeper understanding of important concepts (e.g., missing data mechanism) in the missing data literature.

The rest of the paper is organized as follows. I first provide background information related to missing data. Specifically, I review important missing data concepts such as missing data rules, missing data mechanisms, and missing data patterns. I also introduce the concept of "missing data rule", which is used to describe the statistical model for generating missing data. Understanding these concepts is essential for generating missing data, especially for researchers who are doing their first simulation study involving missing data. I then explain how to use different missing data rules and their associated algorithms to generate missing data with different properties. Here, I focus on describing how to generate missing data with different missing data mechanisms because the type of missing data mechanism is one of the most important properties that affect the performance of modern missing data techniques, and is almost always manipulated in simulation studies involving missing data. I also explain the advantages and disadvantages of using different missing data rules and algorithms to generate missing data in simulation studies. Next, I discuss several other important considerations for conducting simulation studies with missing data. Finally, I conclude with recommendations for generating missing data for simulation studies.

Preliminaries

What Are Missing Data Rules?

In this paper, I use the term "missing data rule(s)" to mean a statistical model for generating missing data. This model allows researchers to calculate the probability of being missing for each subject and each variable. An example of a missing data rule is *each subject has* 20% *probability of being missing from the variable* Y. In statistical terms, this missing data rule is $P(M = 1) = 0.2$ where M is a random indicator variable with $M = 1$ indicating a missing value in Y. This missing data rule assumes that the chance of one subject being missing is independent of the chance of another subject being missing, a common assumption made by researchers when generating missing data (Graham, 2010). A given dataset can have a set of missing data rules, one for each variable, or a single missing data rule for multiple variables.

Like other kinds of statistical models, a missing data rule has one or more parameters associated with it. Specifically, these parameters are associated with the distribution of the missing data indicator M. When researchers gen-

[1] Generating missing data for simulation studies is different from imputing missing data in MI. For the former, researchers need to create missing data in simulated complete datasets; however, for the latter, they need to fill in the missing data based on the best estimates of the parameters.

erate missing data, they have to specify these parameters associated with the missing data rule. In the above example of the missing data rule, the parameter associated with the missing data rule is the 20% probability of being missing. This parameter pertains to the population. With sample data, the parameters associated with the missing data rule can only be estimated. Although the average of the estimated parameter values over repeated samples is equal to the true parameter value, the estimated parameter in a specific sample data is usually different from the true parameter value. In conclusion, when generating missing data for simulation studies, it is important to explicitly state the missing data rule and the parameters associated with it. As I will show later, knowing the missing data rule makes it easy for researchers to figure out many properties of the missing data being created.

Missing Data Mechanisms

In the missing data literature, missing data mechanism is usually defined as the statistical relationship between subjects (or variables) and the probability of missing data (Nakagawa, 2015). In this paper, I note that missing data mechanisms are equivalent to missing data rules. More precisely, a missing data rule is a specific missing data mechanism that describes how missing values are generated in the data. An introductory course on missing data usually explains the three types of missing data (i.e., three types of missing data rules) described in Rubin, 1976: 1) missing completely at random (MCAR), 2) missing at random (MAR), and 3) missing not at random (MNAR). In this section, I review these three types of missing data mechanisms in both informal and formal terms.

Let us consider a dataset with n subjects and p variables denoted as $Y_1, \ldots, Y_p$. When researchers do not have missing data, their dataset should look like a matrix with n rows and p columns. When researchers have missing data, they can consider the missing data as unobserved values that create holes in the data matrix. Suppose only Y_1 has missing values. If Y_1 is MCAR, then the probability of a subject having a missing value of Y_1 does not depend on its unobserved value in Y_1 nor its observed values of other variables. This means that knowing the subject's values on any of the variables does not give you any information about its probability of being missing. An example of MCAR data is when the paper-form questionnaire data are missing because a house cat spilled coffee on the table. In this case, there are no observed or missing data that can predict the probability of being missing. If Y_1 is MAR, then the probability of a subject being missing depends on its observed values of other variables but does not depend on its value of Y_1. In other words, MAR means *conditionally missing at random*: conditional on the observed values of other variables, the probability of being missing does not depend on the value of Y_1. An example of MAR data is when shy participants are less willing to answer questions regarding their sexuality, thus creating missing values on a sexuality question in a survey. In this case, if researchers have measured participants' shyness, they can predict the probability of missing data on the sexuality question. If Y_1 is neither MCAR nor MAR, then Y_1 is MNAR, where the probability of a subject having a missing value on Y_1 depends on its value of Y_1. A classical example of MNAR data is when participants with high incomes avoid answering questions about income. In this case, the probability of missing the income data is related to the participants' own income.

To define the types of missing data mechanisms formally, let $Y = (Y_1, \ldots, Y_p)^T$ be a random vector representing the p variables in the dataset and $y = (y_1, \ldots, y_p)^T$ represent the realizations of Y. Same as above, suppose Y_1 is the only random variable with missing data. Let M be a random indicator variable with $M = 1$ representing a missing value in Y_1; for the rest of the paper, I will call M the *missing data indicator*. MCAR occurs when the distribution of M does not depend on y:

$$P(M = 1|y) = P(M = 1) \text{ and } P(M = 0|y) = P(M = 0).$$

To define MAR and MNAR, we have to break down y into the observed (y_{obs}) and the unobserved or missing (y_{mis}) parts of y; that is $y = (y_{mis}, y_{obs})^T$. In this case, since Y_1 is the only variable with missing data, $y_{mis} = y_1$ and $y_{obs} = (y_2, \ldots, y_p)^T$, MAR occurs when the distribution of M depends on y_{obs} but not y_{mis}:

$$P(M = 1|(y_{mis}, y_{obs})^T) = P(M = 1|y_{obs})$$

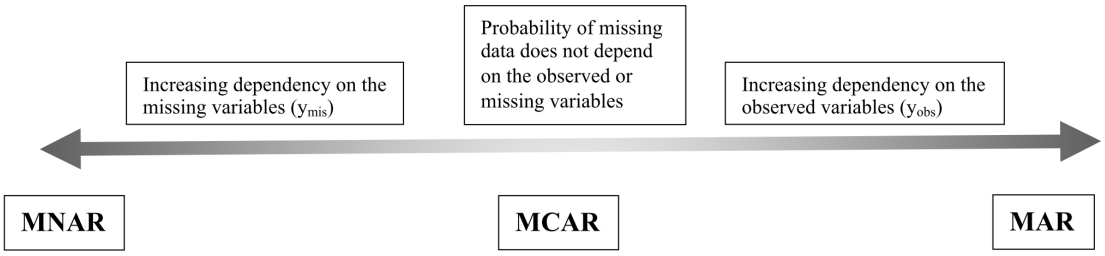
and

$$P(M = 0|(y_{mis}, y_{obs})^T) = P(M = 0|y_{obs}).$$

Lastly, MNAR occurs when the distribution of M depends on y_{obs}; that is when $P(M = 1|(y_{mis}, y_{obs})^T)$ and $P(M = 0|(y_{mis}, y_{obs})^T)$ can not be simplified further.

From the above definitions of MCAR, MAR, and MNAR data, we can see that MCAR can be viewed as a special case of MAR data or MNAR data. Specifically, MAR data becomes MCAR data when M's dependency on y_{obs} is zero; similarly, MNAR becomes MCAR data when M's dependency on y_{mis} is zero. In fact, the distinction between MCAR and MAR lies along a continuum where the missing data indicator M increases its dependency on y_{obs}; similarly, the distinction between MCAR and MNAR lies along a continuum where M increases its dependency on y_{mis} (see Figure 1 for a graphic representation of the relationships between MCAR, MAR, and MNAR data). In other words, some data can be more or less MAR depending on how strong M is related on y_{obs};

Figure 1 ■ Relationships between different missing data mechanisms. The differences between MCAR, MAR, and MNAR data fall on a continuum. As the probability of missing data becomes more dependent on the values of the observed variables, MCAR becomes MAR data. On the other hand, as the probability of missing data becomes more dependent on the values of the missing variables, MCAR becomes MNAR data.



and some data can be more or less MNAR depending how strong M is related on y_{mis}. In the later sections, I will focus on explaining how to generate missing data with varying degrees of being MAR.

Another important concept related to the types of missing data mechanisms is *ignorability*. Ignorable data are the types of missing data that can be *effectively* handled by modern missing data techniques such as FIML, MI and TS. Missing data needs to satisfy two conditions to become ignorable missing data: 1) the missing data are either MCAR or MAR; 2) parameters associated with the specific missing data rule are distinct from the parameters associated with the distribution of the variables in the dataset (Rubin, 1976). The second condition means that the parameters associated with the distribution of M are distinct from the parameters associated with the distribution of Y. To explain why these conditions are needed, let θ and ϕ are the parameters associated with Y and M, respectively, and let $f(y, m; \theta, \phi)$ denote the joint density of Y and M. Because θ and ϕ are distinct, when the data are incomplete, the observed data likelihood can be obtained via the marginal of y_{obs} as follows:

$$f(y_{obs}, m, \theta, \phi) = \int f(y_{obs}, y_{mis}; \theta) f(m|y_{obs}, y_{miss}; \phi) dy_{mis} \quad (1)$$

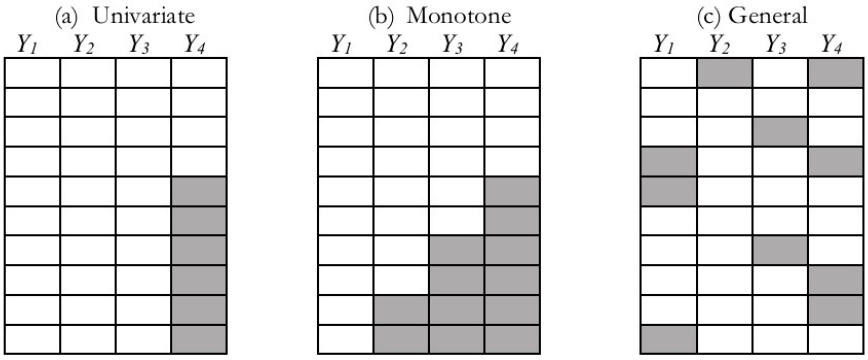
When the data are MCAR, $f(m|y_{obs}, y_{miss}; \phi) = f(m; \phi)$; when the data are MAR, $f(m|y_{obs}, y_{mis}; \phi) = f(m|y_{obs}; \phi)$. Since neither $f(m; \phi)$ nor $f(m|y_{obs}; \phi)$ involves y_{mis}, we can take $f(m; \phi)$ or $f(m|y_{obs}; \phi)$ out of the integral. In other words, for MCAR or MAR data, it is sufficient to maximize $\int f(y_{obs}, y_{mis}; \theta) dy_{mis}$ with respect to θ if we only want estimate θ. There are MAR data that violate the second assumption for ignorable missing data (i.e.,θ and ϕ are not distinct); in such cases, statistical methods assuming ignorability are not optimal but may still be good. Therefore, in practice, ignorable missing data stand for MCAR or MAR data and non-ignorable missing data implies MNAR data. The advantage of ignorable data and their relationship with types of missing mechanisms motivate researchers to generate missing data with different missing mechanisms when studying methods for handling missing data.

Missing Data Patterns

Missing data pattern refers to the arrangement of observed and missing values in a dataset (Graham, 2010). It is often confused with missing data mechanism (e.g. Grigsby & McLawhorn, 2019). The distinction is that a specific missing data mechanism is a missing data rule that describes the relationship between subjects and the probability of missing whereas a specific missing data pattern is a data configuration that describes the location of the missing values in the data.

There are generally three kinds of missing data patterns. The univariate pattern occurs when missing values are on one variable or a group of variables that is either entirely observed or entirely missing for each case, but all other variables are completely observed (Schafer & Graham, 2002a) (see Figure 2a). The univariate pattern has the lowest number of missing data patterns; in other words, it has two missing data patterns, one pattern where subjects have complete data and the other pattern where subjects have missing data. Another type of missing pattern is the monotone pattern (e.g. Newman, 2003; Schafer & Graham, 2002a; Strike et al., 2001). In the monotone missing pattern, a group of variables $Y_1, \ldots, Y_p$ can be ordered in such a way that if Y_j is missing for a subject, then $Y_{j+1}, \ldots, Y_p$ is also missing (see Figure 2b). Notice that the univariate pattern can be viewed as a special case of monotone pattern. Monotone patterns can be seen in longitudinal studies with

Figure 2 ■ Types of missing data patterns. Rows represent subjects; columns represent variables. The shaded cells represent the location of missing values.



attrition, where Y_j representing a variable or a group of variables collected at time j. Last, the general missing data pattern occurs when a group of variables may be missing for any subject, creating a dataset with missing values dispersed throughout the data matrix in a haphazard fashion (Graham, 2010) (see Figure 2c).

Although missing data pattern and missing data mechanism are distinct concepts, they do affect each other. Given a specific missing data rule with a certain type of missing data mechanism, the number and the type of missing data pattern will be determined. For example, suppose a dataset has $Y_1, \ldots, Y_p$ variables, if the missing data rule is *each subject has* 20% *probability of being missing from the variable* Y_1, then the missing data pattern is univariate, implying two missing patterns.

When designing simulation studies examining missing data techniques, researchers often consider missing data patterns less important than missing data mechanisms, probably because missing data patterns are not directly related to the ignorability property of missing data. However, some researchers have found that the number of missing data patterns can affect the performance of missing data techniques (Savalei & Bentler, 2005; Zhang & Savalei, 2020). Therefore, when designing simulation studies, researchers should manipulate both missing data mechanisms and patterns; this requires researchers to know how to generate missing data with different missing data mechanisms and patterns, which I will explain in detail in the following section.

Generating Missing Data for Simulation Studies

In this section, I will explain how to specify missing data rules to generate data with different missing data mechanisms and patterns and how to implement these missing data rules with computer algorithms. For each missing data rule, I describe the parameters and various properties associated and discuss the advantages and disadvantages relative to other types of missing data rules. I mainly focus on MCAR and MAR data with the univariate missing data pattern because they are the most commonly studied missing data in the missing data literature, but I will also briefly discuss generating MNAR data as well as generating missing data with a large number of missing data patterns. Table 4 at the end of the article presents a summary of different missing data rules and algorithms, highlighting their advantages and disadvantages. Sample R code for implementing different missing data rules is posted on the Open Science Frame (OSF) website. [2]

Generating MCAR Missing Data

General Missing Data Rules for MCAR Data. Missing data rules for MCAR data always involve each subject's probability of being missing from one or more variables. The probability of being missing is the parameter value associated with the missing data rule, denoted as θ earlier. This parameter value affects the expected percentage of missing and the expected number of missing data patterns in a sample dataset.

For MCAR data with univariate pattern, the missing data rule is that *each subject has* π *probability of being missing from the variable(s) with missing data.* Putting it in statistical terms, this missing data rule is $P(M = 1) = \pi$ where M is the missing data indicator, and π is the parameter associated with the missing data rule. Given this missing data rule, researchers can determine various properties associated with the MCAR data, including the expected

[2] Link to the OSF website: https://osf.io/pmn9z/

percentage of missing values and the expected number of missing data patterns in the MCAR data.

To explain how the missing data rule affects the missing data properties, let n be the number of subjects in the data, and K be the random variable indicating the number of subjects with missing values in the data. Given this missing data rule and assuming the chance of one subject being missing is independent of the chance of another being missing, K follows a binomial distribution: $K \sim \text{Bin}(n, \pi)$, where $0 \leq \pi \leq 1$. Since $E(K) = n\pi$ and $\text{Var}(K) = n\pi(1-\pi)$, the expected percentage of missing values is

$$E(\Pi) = E(\frac{K}{n}) = \frac{1}{n}E(K) = \pi, \tag{2}$$

where $\Pi = K/n$ is the random variable denoting the estimated percentage of missing values in a sample. The variance for this estimated percentage of missing values is

$$\text{Var}(\Pi) = \text{Var}(\frac{K}{n}) = \frac{1}{n^2}\text{Var}(K) = \frac{\pi(1-\pi)}{n}. \tag{3}$$

This variance shows that given our missing data rule, researchers may not always obtain the exact π percentage of missing values in a sample dataset.

Researchers can also determine the expected number of distinct missing data patterns in a sample, given a MCAR missing data rule. Consider two possible missing data patterns: pattern 1 includes subjects with complete data; pattern 2 includes subjects with missing values. For $j \in \{1, 2\}$, let I_j be the indicator variable of the event that pattern j is present in at least one subject in the sample. The probability that pattern 1 is present in at least one subject is $P(I_1 = 1) = E(I_1) = 1 - \pi^n$. The probability that pattern 2 is present in at least one subject is $P(I_2 = 1) = E(I_2) = 1 - (1-\pi)^n$. Let D be the number of distinct missing data patterns: $D = \sum_{j=1}^{2} I_j$. The expected number of distinct missing data patterns is

$$\begin{aligned} E(D) &= E(\sum_{j=1}^{2} I_j) = \sum_{j=1}^{2} E(I_j) \\ &= (1-\pi^n) + (1-(1-\pi)^n) \\ &= 2 - \pi^n - (1-\pi)^n, \end{aligned} \tag{4}$$

which shows that as the sample size increases, the expected number of patterns converges very quickly to 2, which is the maximum number of patterns for this missing data rule.

MCAR Missing Data Rules for Creating More Missing Data Patterns. To generate MCAR data with more missing data patterns, researchers can allow each subject's chance of being missing on one variable to be independent of their chance of being missing on another variable. For example, if the variables $Y_1, \ldots, Y_l$ have missing values, then the missing data rule that can create the maximum number of missing data patterns is *each subject has π_i probability of being missing on variable Y_i where $i \in \{1, \ldots, l\}$*; in other words, the probability of being missing on Y_1 is independent of the probability of being missing on $Y_2, Y_3, \ldots, Y_l$. In this case, there is a total of l parameters: $\pi_1, \ldots, \pi_l$. For each variable Y_i with missing data, the expected percentage of missing values and the variance associated with the estimated missing percentage are the same as those shown in (2) and (3), respectively $\big($i.e., $E(\Pi_1) = \pi_1$ and $\text{Var}(\Pi_i) = \frac{\pi_i(1-\pi_i)}{n}\big)$.

As mentioned before, this missing data rule can create the maximum number of missing data patterns with l number of variables (i.e., $m = 2^l$ number of patterns). However, in a given sample, some of the missing data patterns may not be realized. The expected number of distinct missing data patterns in a sample is

$$E(D) = m - \sum_{j=1}^{m}(1-\eta_j)^n, \tag{5}$$

where $\eta_1, \ldots, \eta_m$ are the corresponding probabilities for patterns $1, \ldots, m$. Equation (5) is just a generalized formula for Equation (4). Similar to (4), Equation (5) shows that as the sample size increases, the expected number of patterns converges to m, the maximum number of possible patterns. This makes sense because as the sample size increases, all possible patterns will eventually be realized. In addition, since n appears as an exponent in (5), $E(D)$ converges to m at an exponential rate; therefore, for a dataset with a relatively large sample size, this missing data rule should create a large number of missing data patterns.

Algorithms for Implementing MCAR Missing Data Rules. In the missing data literature, there are generally two methods for implementing MCAR missing data rules. The first method is randomly deleting the desired percentage of missing values (e.g., Enders, 2001b; Yuan & Bentler, 2000; Savalei & Bentler, 2005; Strike et al., 2001; Savalei & Yuan, 2009). The deletion can be accomplished by deleting every ith subject (e.g., deleting every second subject to create 50% missing data) (e.g., Yuan & Bentler, 2000) or deleting randomly until the desired percentage is reached (e.g., Savalei & Bentler, 2005; Savalei & Yuan, 2009; Strike et al., 2001). One problem with this method is that the estimated probability of being missing is equal to the expected probability of being missing across different datasets; however, as shown in (3), there is sampling variability associated with the estimated percentage of missing data. [3] Whether this problem matters for simulation studies depends on the

[3] I note that in planned missing design, the percentage of missing data is held constant across samples.

purpose of the simulation study. For example, if the purpose of the simulation is to examine the average performance of a missing data technique across samples with a large sample size, then this issue does not matter because the sampling variability is very small for a large sample and does not really affect the computation of statistics that are aggregated across samples. However, if the purpose is to examine the performance of missing data techniques under small sample sizes, it may be better to incorporate the sampling variability when implementing the missing data rule; this will make the simulation more realistic.

The second method involves comparing the values of a variable that has missing data with the corresponding values of a uniform random variable ranging between zero and one (e.g., Enders, 2001a, 2004; Enders, 2010; De Raadt et al., 2019; Kim & Bentler, 2002; Jamshidian & Siavash, 2010). Taking a concrete example, suppose that there are 200 subjects in the data and the missing data rule is that *each subject has 20% probability of being missing from the variable* Y. Given that the data of 200 subjects for variable Y are already generated, to create missing values, researchers first draw 200 subjects from a uniform random variable U ranging from zero to one. Then they pair 200 subjects for Y with the 200 corresponding subjects for U. If the ith subject in U is less than 0.2, then the ith subject in Y should be removed. This method is equivalent to implementing the missing data rule directly by allowing each subject in Y to have a 20% chance of being missing. In fact, researchers can create a missing value indicator M from U by letting $M = 1$ whenever $U \leq 0.2$. In other words, instead of drawing cases from a uniform variable, researchers can draw 200 subjects from an indicator random variable M that has a 20% chance of being one, and then delete subjects for Y when M equals one. I recommend this way of implementing the missing data rule because it is the most direct and straightforward way of implementing MCAR missing data rules. For sample R code for generating MCAR data, please refer to the OSF website in Footnote 2.

Advantages and Disadvantages of MCAR Missing Data Rules and Algorithms. The main advantage of the MCAR missing data rules is that they are very easy and intuitive to understand and implement. However, as explained previously, MCAR data are just special cases of MAR or MNAR data where the probability of missing data does not depend on any the observed or the missing variables. As the probability of missing becomes more dependent on the observed or missing variable, the missing data mechanism changes from MCAR to MAR or MNAR. Therefore, for a simulation study investigating the effect of different missing data mechanisms on missing data techniques, it is insufficient to solely generate MCAR data. It is also necessary to generate MAR and MNAR data, which I will explain in the following sections.

Generating MAR Missing Data

For MAR data, the probability of a subject having a missing value depends on the observed values of other variables. In other words, researchers can predict the probability of missing values from the observed values of other variables. For the rest of the paper, I call the variable that can predict the probability of missing values the *missing data predictor*. A missing data predictor can be one single variable in the dataset or it can be a new variable that is a linear combination of several variables in the dataset.

Generating MAR data is more complicated than generating MCAR data in two ways. First, the missing data rules for MAR data are more complicated than those for MCAR data. The missing data rules for MAR data can be organized into several categories: 1) single cutoff method; 2) multiple cutoff method; 3) percentile method; 4) logistic regression method. The most commonly used MAR missing data rule in psychological research is the single cutoff method (e.g., Yuan & Bentler, 2000; Allison, 2000; Enders, 2004; Musil et al., 2002; Yuan & Savalei, 2014). Second, researchers can vary the strength and shape of the dependency between the missing data indicator and the missing data predictor. The strength of the dependency can be weak or strong; the shape of the dependency can be linear or curvilinear. The dependency commonly used in simulation studies is strong and linear (e.g., Yuan et al., 2015; Yuan & Savalei, 2014). In addition, like MCAR data, MAR data can vary in the number and type of missing data patterns, with the univariate pattern being the most commonly studied pattern (e.g., Enders, 2001b, 2010; Yuan & Bentler, 2000; Jia & Wu, 2019).

In the following sections, I will explain the different types of missing data rules and algorithms for generating MAR missing data. For each type of MAR missing data rule, I explain how to create different kinds of patterns and strengths of dependency. I focus more on the MAR data generated using the single cutoff method with univariate pattern and linear dependency because this kind of MAR data is more commonly used in simulation studies involving missing data. Finally, I compare and contrast the relative advantages and disadvantages of the different types of MAR missing data rules and algorithms.

Single Cutoff Method

Missing Data Rules for the Single Cutoff Method. Missing data rules associated with the single cutoff method involves specifying one cutoff point in each missing data predictor. Consider a MAR dataset where Y_1 is the variable with missing data and Y_2 is the missing data predictor with a cutoff point a. In this case, the MAR dataset has a univariate pattern; the missing data rule is *if a subject has* $Y_2 \geq a$,

Table 1 ■ Contingency table for generating MAR data using the single cutoff method.

	M	
U	1	0
1	$A = P(M = 1\|U = 1)P(U = 1) = \pi_1\pi_0$	$B = P(M = 0\|U = 1)P(U = 1) = (1 - \pi_1)\pi_0$
0	$C = P(M = 1\|U = 0)P(U = 0) = \pi_2(1 - \pi_0)$	$D = P(M = 0\|U = 0)P(U = 0) = (1 - \pi_2)(1 - \pi_0)$

Note. M is an indicator variable indicating whether Y_1 is missing: $M = 1$ when Y_1 is missing, and $M = 0$ when Y_2 is not missing. U is the indicator variable indicating whether Y_2 is equal to and greater than a: $U = 1$ when $Y_2 \geq a$, and $U = 0$ when $Y_2 < a$.

then its probability of being missing on Y_1 is π_1, and if it has $Y_2 < a$, then its probability of being missing on Y_1 is π_2. To define the missing data rule in statistical terms, let M be the missing data indicator for missing data in Y_1, and U be the indicator denoting whether Y_2 is above a (i.e., $U = 1$ when $Y_2 \geq a$ and $U = 0$ when $Y < a$). Notice since U is a direct function of Y_2, U is also a missing data predictor. Therefore, the missing data rule can be written as $P(M = 1|U = 1) = \pi_1$ *and* $P(M = 1|U = 0) = \pi_2$. Since this missing data rule just involves the two indicators M and U, it can be best illustrated using a contingency table for the two indicators; this contingency table is shown in Table 1.

There are three parameters associated with this missing data rule. Two of the parameters are π_1 and π_2, the conditional probabilities of being missing on Y_1. The third parameter is the probability that Y_2 is equal to or greater than a: $P(Y_2 \geq a) = P(U = 1) = \pi_0$. Notice that the third parameter is directly related to the cutoff point a; this means that to set a value for this parameter, researchers only need to specify the value for a. For each parameter, researchers can calculate the variance associated with the estimated value (see Equation (3) for derivation). If n is the total number of subjects and $n_1 = n\pi_0$ is the number of subjects with Y_2 values above a, then the respective variances for the estimated π_0, π_1, and π_2 are

$$\text{Var}(\Pi_0) = \frac{\pi_0(1 - \pi_0)}{n}, \tag{6a}$$

$$\text{Var}(\Pi_1) = \frac{\pi_1(1 - \pi_1)}{n_1} \tag{6b}$$

and

$$\text{Var}(\Pi_2) = \frac{\pi_2(1 - \pi_2)}{n - n_1}. \tag{6c}$$

Now, I explain how to use these parameter values to determine the expected percentage of missing values in Y_1 and the strength of dependency between the missing data indicator M and the missing data predictor U. To find the expected percentage of missing values, we first calculate the unconditional probability of a subject being missing from Y_1:

$$\begin{aligned}\pi_{miss} &= P(M = 1) \\ &= P(M = 1|U = 1)P(U = 1) \\ &+ P(M = 1|U = 0)(1 - P(U = 1)) \\ &= \pi_1\pi_0 + \pi_2(1 - \pi_0).\end{aligned}$$

Then let K be a random variable indicating the number of subjects with missing data in a sample dataset. We know $K \sim \text{Bin}(n, \pi_{miss})$. Thus, the expected percentage of missing data across samples is

$$E(\frac{K}{n}) = \pi_{miss} = \pi_1\pi_0 + \pi_2(1 - \pi_0), \tag{7}$$

and the variance for this estimated percentage of missing date is

$$\text{Var}(\frac{K}{n}) = \frac{\pi_{miss}(1 - \pi_{miss})}{n}. \tag{8}$$

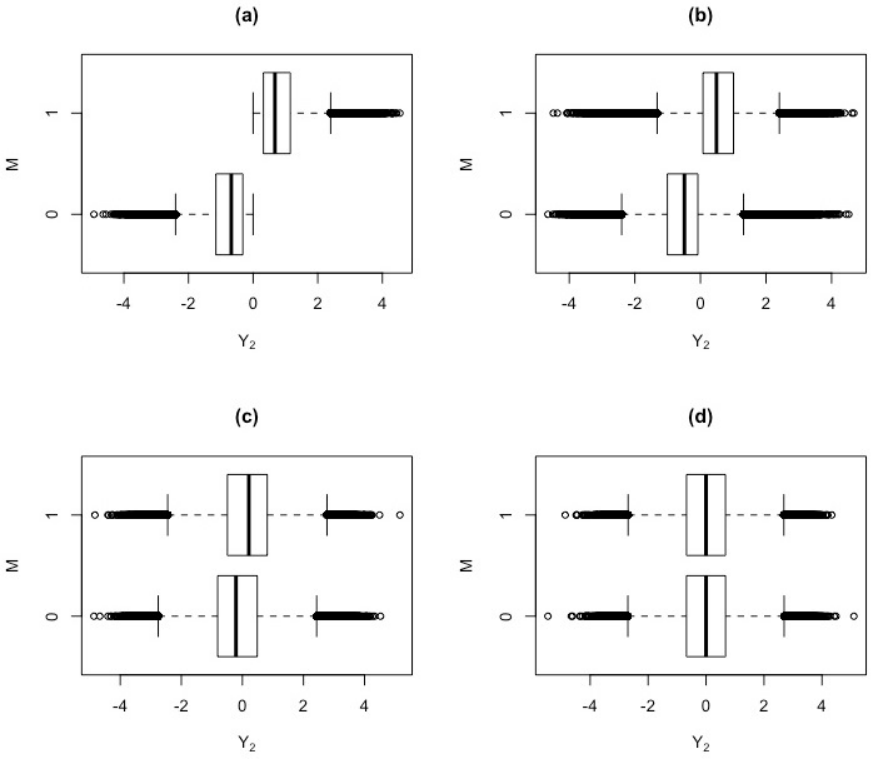
Notice Equations (7) and (8) are the same as Equations (2) and (3), expect that π in (2) and (3) is replaced by π_{miss} in (7) and (8). Similarly, by replacing π in (4) to π_{miss}, researchers can find the expected number of patterns for the MAR missing data rule: $E(D) = 2 - \pi_{miss}^n - (1 - \pi_{miss})^n$.

Measuring the Strength of Dependency Under the Single Cutoff Method. As mentioned in the preliminaries section, the difference between MCAR and MAR data can be viewed as a continuum where the missing data indicator M increases its dependency on the missing data predictor U. When there is no dependency between M and U, the missing data are MCAR, but as M and U become more dependent, the missing data become *more* MAR. As a result, when conducting simulation studies involving missing data, it is important to know how to measure the strength of dependency between M and U in order to manipulate the *degree* of MAR in the missing data.

Since M and U are two binary variables, researchers can measure the strength of dependency between M and U using the absolute risk difference (ARD) or odd ratio (OR), which are standard association measures for binary variables. The respective equations for ARD and OR are

$$\text{ARD} = |\pi_1 - \pi_2|, \tag{9}$$

Figure 3 ■ MAR data created by the single cutoff method, varying in the strength of dependency. M is the missing data indicator with $M = 1$ indicating Y_1 is missing, and $M = 0$ indicating Y_1 is not missing. Y_2 is the missing data predictor, which follows the standard normal distribution. Since M is a binary variable while Y_2 is a continuous variable, boxplots can be used to show the strength of dependency between M and Y_2. The strength of dependency decreases as the boxplox for $M = 0$ overlaps more with the one for $M = 1$. In other words, the strength of dependency decreases as the graph goes from (a) to (d). Each graph is based on a large simulated dataset.



and

$$\begin{aligned} \text{OR} &= \frac{P(M=1|U=1)/(1-P(M=1|U=1))}{P(M=1|U=0)/(1-P(M=1|U=0))} \\ &= \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}. \end{aligned} \tag{10}$$

Large ARD values indicate strong dependency; OR values farther away from one indicate strong dependency. Notice that OR is not defined when $1-\pi_1 = 0$ or $1-\pi_2 = 0$; therefore, if any of these cases occurs, ARD should be used to measure the strength of dependency.

Equations (9)-(10) measure the strength of dependency between M and U at the population level. At the sample level, the estimated strength of dependency may vary from sample to sample. The variances associated with the estimated ARD and estimated $\log(\text{OR})$ are as follows (see Agresti & Kateri, 2011, for derivation):

$$\begin{aligned} \text{Var}(\Pi_1 - \Pi_2) &= \text{Var}(\Pi_1) + \text{Var}(\Pi_2) \\ &= \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n-n_1}. \end{aligned} \tag{11}$$

$$\text{Var}(\log \text{OR}) = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}, \tag{12}$$

where A, B, C and D are defined in Table 1.

Since OR is closely related to the logistic regression model, we can also define the relationship between M and U using the logistic regression framework. In other words, the log-odds of M can be predicted by U:

$$\log\left(\frac{P(M=1)}{1-P(M=1)}\right) = \beta_0 + \beta_1 U, \tag{13}$$

where β_0 is the log-odds of M given $U = 0$:

$$\beta_0 = \log\left(\frac{P(M=1|U=0)}{1-P(M=1|U=0)}\right) = \log\left(\frac{\pi_2}{1-\pi_2}\right),$$

and β_1 is the log of the OR:

$$\beta_1 = \log(\text{OR}) = \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right).$$

In the logistic regression, the higher the β_1 value, the stronger the dependency is. Note that Equation (13) shows that the missing data rule specified using the single cutoff method is actually equivalent to a logistic regression model, which can be directly used to generate MAR data using the logistic regression method. I will explain more about the connections between these two methods in the section about the logistic regression method.

Examples of Missing Data Rules with Different Strengths of Dependency Under the Single Cutoff Method. Since the strength of dependency is one important property for MAR data, in this section, I show with a few examples how to specify missing data rules that vary in the strength of dependency between the missing data indicator (M) and the missing data predictor (U or Y_2). I begin by explaining the missing data rules with the strongest dependency. If a missing data rule specifies the strongest dependency between M and U, then the value of M can always be accurately predicted given the value of U, a case that occurs when $P(M = 1|U = 1) = \pi_1 = 1$ and $P(M = 1|U = 0) = \pi_2 = 0$ or when $P(M = 1|U = 0) = \pi_2 = 1$ and $P(M = 1|U = 1) = \pi_1 = 0$. In this case, ARD equals one, but OR and the logistic regression model are undefined.

An example of a missing data rule with the strongest dependency is *if a subject has $Y_2 \geq 0$, then its Y_1 value is always missing* (see Table 2a for the contingency table for this missing data rule). The cutoff point used in this missing data rule is $Y_2 = 0$. The three parameters associated with this rule are $\pi_0 = P(U = 1) = P(Y_2 \geq 0) = 0.5$, $\pi_1 = 1$, and $\pi_2 = 0$, assuming Y_2 follows the standard normal distribution. To demonstrate the property of this missing data rule, I have simulated a large sample dataset ($n = 1,000,000$) and then generated missing values in the dataset according to the above missing data rule. Since Y_2 is a continuous variable while M is a categorical random variable, I used boxplots to show the assication between Y_2 and M. Figure 3a shows the association between Y_2 and M based on this missing data rule. Figure 3a shows a complete separation of the boxplot for $M = 0$ from the one for $M = 1$ along the $Y_2 = 0$ cutoff point; this indicates that researchers can accurately predict the value of M based on the value of Y_2; in other words, M and Y_2 are highly associated with each other. In addition, another interesting property for missing data rules with the strongest dependency is that the percentage of missing values only depends on the parameter π_0. For the missing data rule in our example, the expected percentage of missing values calculated using Equation (7) is $\pi_{miss} = \pi_0 = 0.5$.

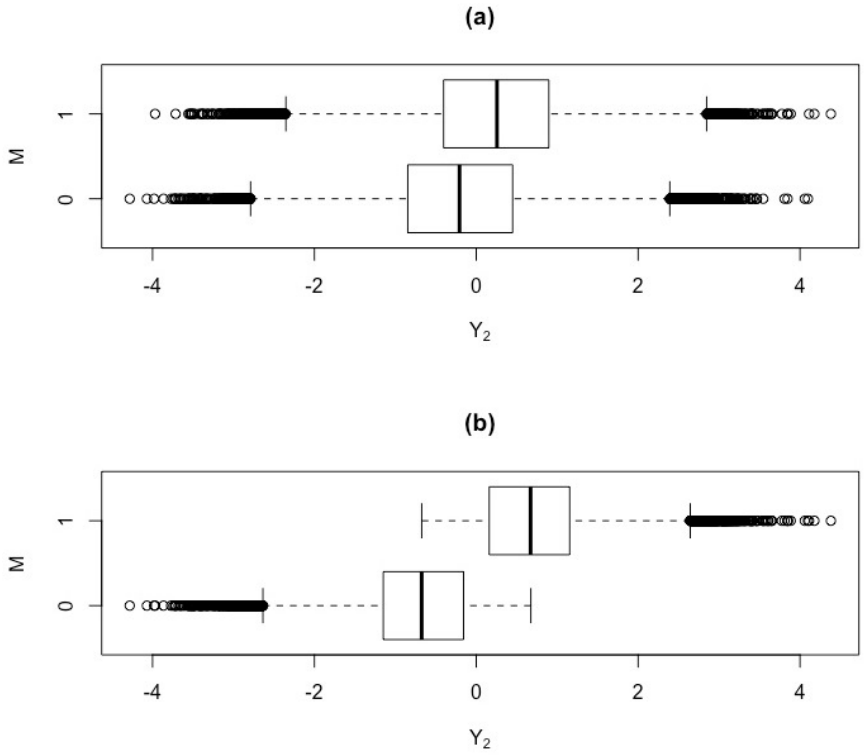
As the strength of dependency decreases, it is harder to predict the missing data indicator from the missing value predictor; ARD value decreases, but OR and the logistic regression model are no longer undefined. An example of a missing data rule with a weaker dependency is *if a subject has $Y_2 \geq 0$, then its Y_1 value has 80% probability of being missing; otherwise, its Y_1 has 20% probability of being missing* (see Table 2b for the contingency table for this rule). The three parameters for this rule are $\pi_0 = 0.5$, $\pi_1 = 0.8$ and $\pi_2 = 0.2$. In this case, the ARD and OR are 0.5 and 16, respectively. The logistic regression model is $\log\left(\frac{P(M=1)}{1-P(M=1)}\right) = \log(0.25) + \log(16)U$. Figure 3b shows the boxplots of Y_2 values for this missing data rule. With this missing data rule, the boxplot of Y_2 for subjects with $M = 1$ overlaps with the boxplot for $M = 0$, making it impossible to accurately predict M from Y_2.

An example missing data rule with an even weaker dependency is *if a subject has $Y_2 \geq 0$, then its Y_1 value has 60% probability of being missing; otherwise, its Y_1 has 40% probability of being missing* (see Table 2c for contingency table). The parameters are $\pi_0 = 0.5$, $\pi_1 = 0.6$ and $\pi_2 = 0.4$. The ARD and OR are 0.2 and 2.25. The logistic equation is $\log\left(\frac{P(M=1)}{1-P(M=1)}\right) = \log(0.67) + \log(2.25)U$. Figure 3c shows that the boxplot of Y_2 for subjects with $M = 1$ almost completely overlaps with the one for $M = 0$, making the prediction of M based on Y_2 only slighter better than chance.

The weakest dependency occurs when π_1 equals π_2 (see Table 2d as an example). In this case, the data become MCAR; the ARD value is 0 and OR is 1. With MCAR data, as shown in Figure 3d, the boxplot for $M = 1$ completely overlaps with the one for $M = 0$, making the prediction of M based on Y_2 no better than chance. In summary, researchers can generate MAR data with different strengths of dependency by varying the parameters π_1 and π_2 in the missing data rule. ARD and OR can be used to measure the strength of dependency. The strongest dependency has an ARD value of 1; as the strength of dependency decreases, the ARD value decreases and the OR values get closer to 1; the weakest dependency occurs when the data become MCAR, in which case ARD is 0 and OR is 1.

Creating More Missing Data Patterns Under the Single Cutoff Method. To generate MAR data with many missing data patterns, I can let the different missing data indicators depend on different missing data predictors. For example, if two variables, Y_1 and Y_2, have MAR missing data, I can let the probabilities of missing values for Y_1 and Y_2 depend on the observed values of Y_3 and Y_4, respectively. In fact, this way of creating missing data patterns can be used in combination with the single-cutoff, multiple-cutoff, percentile, or logistic regression method for generating MAR data. In the case of the single-cutoff method, an example missing

Figure 4 ■ MAR data created by the multiple cutoff method, varying in the strength of dependency. M is the missing data indicator with $M = 1$ indicating Y_1 is missing, and $M = 0$ indicating Y_1 is not missing. Y_2 is the missing data predictor, which follows the standard normal distribution. Since M is a binary variable while Y_2 is a continuous variable, boxplots can be used to show the strength of dependency between M and Y_2. The strength of dependency decreases as the boxplox for $M = 0$ overlaps more with the one for $M = 1$. In other words, the strength of dependency in graph (b) (AARD=0.33) is stronger than that in graph (a) (AARD=0.10). Each graph is based on large a simulated dataset.



data rule that can create the maximum number of patterns (i.e., four patterns) for two variables Y_1 and Y_2 with missing data is *if the subject has $Y_3 \geq a_1$, then its Y_1 has π_1 probability of being missing, otherwise, Y_1 has π_2 probability of being missing; if the subject has $Y_4 \geq a_2$, then its Y_2 has π_3 probability of being missing; otherwise Y_2 has π_4 probability of being missing.* With this missing data rule, we can still use Equation (5) (with $m = 4$) to calculate the expected number of patterns in a sample dataset. However, for MAR data, the probability of each missing data pattern (i.e., $\eta_1, \ldots, \eta_4$ in Equation (5)) also depends on the correlation between the missing data predictors Y_3 and Y_4. In the most extreme case, if Y_3 and Y_4 have a correlation of one and a_1 equals a_2, then this missing data rule creates data with only two patterns (i.e., univariate pattern); in other words, the probabilities of the other two patterns are both zero. Therefore, to maximize the number of patterns in a sample, I suggest generating missing data predictors that are moderately correlated so that the probability of each pattern is greater than zero, making Equation (5) quickly converges to m, the maximum number of patterns, as $n \to \infty$.

Algorithms for Implementing Missing Data Rules Using the Single Cutoff Method. In the missing data literature, there are three different algorithms to implement missing data rules associated with the single-cutoff method, each with some drawbacks. The first method involves setting a missing data rule, and then applying this rule subject by subject until the desired percentage of missing data is reached (e.g., Enders, 2004). This method is highly problematic because it violates the assumption that each subject is coming from the same population. Assuming that each subject comes from the same population and thus follows the same missing data rule, then it does not make sense to apply the missing data rule to some subjects but not to others. A consequence of this method is that the percentage of missing data in a sample may be very different from the expected percentage of missing data given by the missing data rule (see Equation (7)). Another problem with this method

Table 2 ■ Contingency tables for MAR data with different strengths of dependency

(a)	M	
U	1	0
1	$A = 0.50$	$B = 0$
0	$C = 0$	$D = 0.50$

(b)	M	
U	1	0
1	$A = 0.40$	$B = 0.10$
0	$C = 0.10$	$D = 0.40$

(c)	M	
U	1	0
1	$A = 0.30$	$B = 0.20$
0	$C = 0.20$	$D = 0.30$

(d)	M	
U	1	0
1	$A = 0.25$	$B = 0.25$
0	$C = 0.25$	$D = 0.25$

Note. M is the missing data indicator with $M = 1$ indicating Y_1 is missing, and $M = 0$ indicating Y_1 is not missing. U is the missing data predictor with $U = 1$ when $Y_2 \geq 0$, and $U = 0$ when $Y_2 < 0$. Suppose Y_2 follows the standard normal distribution. The missing data rule for (a) is *if a subject has* $Y_2 \geq 0$*, then its* Y_1 *value is always missing*; the rule for (b) *if a subject* $Y_2 \geq 0$*, then its* Y_1 *has 80% probability of being missing; otherwise,* Y_1 *has 20% probability of being missing*; the rule for (c) is *if a subject has* $Y_2 \geq 0$*, then its* Y_1 *has 60% probability of being missing; otherwise, its* Y_1 *has 40% probability of being missing*; the rule for (d) is *each subject has 50% probability of being missing from* Y_1. As the table goes from (a) to (d), the strength of dependency goes from the strongest to the weakest.

is that it is impossible to determine the strength of dependency between the missing data indicator and the missing data predictor since the missing data rule is not applied to every subject.

The second method involves deleting a subject whenever its percentile ranking in a sample is higher than the desired missing data percentage (e.g., Savalei & Yuan, 2009; Enders, 2001b, 2010). For example, suppose that each subject's probability of being missing from Y_1 depends on its Y_2 value, and researchers want k percent of subjects in a sample to have missing values in Y_1. Using this method, subjects whose Y_2 values are in the top k percent will have their Y_1 values deleted. This is equivalent to the missing data rule that sets a cutoff point corresponding to the quantile point for the top k percent values of Y_2 and that says *if a subject's* Y_2 *value is greater than the cutoff point, then its* Y_1*'s probability of being missing is one, otherwise,* Y_1*'s probability of being missing is zero.* Notice that this missing data rule is the one with the strongest dependency between the missing data indicator and the missing data predictor. Therefore, one disadvantage of this method is that it does not allow researchers to vary the strength of dependency. Another problem with this method is that the cutoff point may vary across datasets. In other words, the estimated percentage of missing values is forced to be the same across datasets by shifting the cutoff point. Shifting the cutoff point violates the assumption that the same missing data rule should be applied to datasets that come from the same population; thus, I suggest setting a cutoff point and holding it constant across datasets when generating MAR data.

The third method involves deleting the desired percentage of subjects that are above or below a specific cutoff point (e.g., Savalei & Bentler, 2005; Yuan & Bentler, 2000). This method allows researchers to generate MAR data with different strengths of dependency and is almost equivalent to implementing the missing data rule directly to all subjects. However, one problem with this method is that the estimated values for the parameters (i.e., π_1 and π_2 in Equation (6)) associated with the missing data rule are held constant across datasets. However, as shown in (6), there should be variances associated with the estimated values across samples. This problem is trivial if researchers are only interested in large-sample simulations, but if researchers want to study small samples with missing data, it may be important to incorporate the variances of parameter estimates.

Since each of the three methods mentioned above has drawbacks, I do not recommend any of these methods. I recommend researchers to explicitly specify a missing data rule, and then apply this missing data rule to every subject in the dataset. If researchers have the desired percentage of missing data, they should manipulate the parameters associated with the missing data rule so that the expected percentage of missing equals the desired percentage of missing values. Specifically, researchers can manipulate the parameters π_0, π_1 and π_2 in Equation (7) so that π_{miss} in (7) equals to the desired percentage of missing values. Similarly, if researchers have the desired strength of dependency between missing data indicator and predictor, they can manipulate the parameter values so that the AR and OR in Equations (9)-(10) show the desired strength of dependency (see the OSF website in Footnote 2 for sample R code).

Multiple Cutoff Method.

When using the multiple cutoff method to generate MAR data, researchers need to specify multiple cutoff points in a missing data predictor. One advantage of the multiple cutoff method is that it can be used to create a nonlinear relationship between the missing data indicator and the missing data predictor (e.g., Collins et al., 2001; Graham, 2010). A nonlinear relationship occurs when subjects with extreme values on the missing data predictor have a higher or lower probability of being missing than subjects with mid-range values on the predictor. In contrast, a linear relationship occurs when the probability of being missing gradually increases or decreases as the value of the missing data predictor increases. In the following subsections, I explain how to specify the missing data rules associated with the multiple cutoff method to create a linear and nonlinear relationship between the missing data indicator and the missing data predictor.

Missing Data Rules for the Multiple Cutoff Method. When using the multiple cutoff method to create a nonlinear relationship between the missing data indicator and the missing data predictor, I need to specify an upper cutoff and a lower cutoff. Suppose the probability of missing values on Y_1 depends on two cutoff points, a and $-a$, in the variable Y_2. Let M be the missing data indicator, and U be the indicator denoting whether Y_2 value is between the two cutoff points: $U = 1$ when $Y_2 \geq a$ or $Y_2 \leq -a$, and $U = 0$ when $-a < X < a$. In statistical terms, the missing data rule is $P(M = 1|U = 1) = \pi_1$ *and* $P(M = 1|U = 0) = \pi_2$. Notice that this missing data rule is the same as the one for the single cutoff method. In other words, in the case of a nonlinear relationship, a missing data rule associated with the multiple cutoff method can be framed to be the same as the missing data rule associated with the single cutoff method. As a result, in this case, all equations for the single cutoff method can be used for the multiple cutoff method.

On the other hand, to create a linear relationship between the missing data indicator and the missing data predictor, researchers need to specify at least two cutoff points in the missing data predictor. Most of the times, researchers specify three or four cutoff points, which are usually the quartile or quantile points of the missing data predictors (e.g., Strike et al., 2001; Graham, 2010). In other words, researchers can use the quartile or quantile points to divide the values of the missing data predictor into four or five groups, and each subject's value on the missing data predictor has an equal chance to be in any of the groups. Going from the group with the lowest values to the group with the highest values, the probability of being missing usually increases or decreases at a constant rate (e.g., Strike et al., 2001; Graham, 2010).

To give an example, suppose the probability of a subject being missing from Y_1 depends on the three quartile points in the missing data predictor Y_2 (i.e., Y_2 values is divided into *four* groups). One possible missing data rule is that *if a subject's Y_2 value falls into the 1st, 2nd, 3rd or 4th group of Y_2 values (i.e., ordered from the lowest Y_2 value to the highest), then its probability of being missing from Y_1 is 0.3, 0.4, 0.5 or 0.6, respectively.* To define the missing data rule in statistical terms, let M be the missing data indicator, and V be a discrete uniform random variable created based on the values of Y_2 (i.e., V can be considered a missing data predictor):

$$V = \begin{cases} 1 & \text{if } Y_2 < Q_1 \\ 2 & \text{if } Q_1 \leq Y_2 < Q_2 \\ 3 & \text{if } Q_2 \leq Y_2 < Q_3 \\ 4 & \text{if } Y_2 \geq Q_3, \end{cases} \tag{14}$$

where Q_1, Q_2, and Q_3 are the quartile points in Y_2. The missing data rule is that

$$\begin{aligned} P(M = 1|V = 1) &= \pi_1, \\ P(M = 1|V = 2) &= \pi_2, \\ P(M = 1|V = 3) &= \pi_3, \\ &\text{and} \\ P(M = 1|V = 4) &= \pi_4, \end{aligned} \tag{15}$$

where $\pi_1 = 0.3$, $\pi_2 = 0.4$, $\pi_3 = 0.5$ and $\pi_4 = 0.6$ in this example. There are five parameters associated with this missing data rule. Four of them, of course, are π_1, π_2, π_3 and π_4. The fifth parameter is the one related to the probability of V: $P(V = i) = \pi_0 = 0.25$ where $i \in \{1, 2, 3, 4\}$. Note that the value for π_0 is set when researchers decide to use the quartile cutoff points. For each parameter, researchers can calculate the variances associated with the estimates of the parameters. Let n be the total number of subjects, and $n_0 = 0.25n$ be the number of subjects in each quartile group. The variance for the estimated π_0 is

$$\text{Var}(\Pi_0) = \frac{\pi_0(1 - \pi_0)}{n}. \tag{16}$$

The variance of the estimated π_j where $j \in \{1, 2, 3, 4\}$ is

$$\text{Var}(\Pi_j) = \frac{\pi_j(1 - \pi_j)}{n_0}. \tag{17}$$

Table 3 shows contingency table for M and V. Using this contingency table, researchers can calculate each subject's probability of being missing by calculating the

Table 3 ■ Contingency table for generating MAR data using the multiple cutoff method

	M	
V	1	0
1	$\pi_1\pi_0 = (0.3)(0.25) = 0.075$	$(1-\pi_1)(\pi_0) = (1-0.3)(0.25) = 0.175$
2	$\pi_2\pi_0 = (0.4)(0.25) = 0.100$	$(1-\pi_2)(\pi_0) = (1-0.4)(0.25) = 0.150$
3	$\pi_3\pi_0 = (0.5)(0.25) = 0.125$	$(1-\pi_3)(\pi_0) = (1-0.5)(0.25) = 0.125$
4	$\pi_4\pi_0 = (0.6)(0.25) = 0.150$	$(1-\pi_4)(\pi_0) = (1-0.6)(0.25) = 0.100$

Note. M is the missing data indicator with $M = 1$ indicating Y_1 is missing, and $M = 0$ indicating Y_1 is not missing. V be a discrete uniform random variable indicating which quartile the Y_2 value is in: $V = 1$ when $Y_2 < Q_1$; $V = 2$ when $Q_1 \leq Y_2 < Q_2$; $V = 3$ when $Q_2 \leq Y_2 < Q_3$; and $V = 4$ when $Y_2 \geq Q_3$, where Q_1, Q_2 and Q_3 are the quartile points in Y_2.

marginal probability of $M = 1$:

$$\begin{aligned}\pi_{miss} &= P(M = 1)\\ &= \pi_1\pi_0 + \pi_2\pi_0 + \pi_3\pi_0 + \pi_4\pi_0\\ &= (0.3)(0.25) + (0.4)(0.25) + (0.5)(0.25) + (0.6)(0.25)\\ &= 0.45.\end{aligned} \tag{18}$$

Let n be sample size and K be the number of subjects with missing data. We know that $K \sim \text{Binomial}(n, \pi_{miss})$. Therefore, the expected percentage of missing values is

$$\begin{aligned}E\left(\frac{K}{n}\right) = \pi_{miss} &= \pi_1\pi_0 + \pi_2\pi_0 + \pi_3\pi_0 + \pi_4\pi_0\\ &= 0.45.\end{aligned} \tag{19}$$

The variance of this estimated percentage over repeated samples is

$$\text{Var}\left(\frac{K}{n}\right) = \frac{\pi_{miss}(1-\pi_{miss})}{n}. \tag{20}$$

The expected number of distinct missing patterns can be calculated by Equation (4) by setting $\pi = \pi_{miss}$. Overall, the multiple cutoff method is very similar to the single cutoff method. The main difference is with the single cutoff method, the missing data predictor only has one cutoff point, whereas with the multiple cutoff method, the missing data predictor usually has three or four cutoff points.

Measuring Strength of Dependency Under the Multiple Cutoff Method. One problem with the multiple cutoff method is that there is no straightforward way to measure the strength of dependency between the missing data indicator and the missing data predictor. I propose two possible ways to measure the strength of dependency for missing data rules with the same number of cutoff points. One way is to calculate the average change in the probability of missing values as the missing data predictor V increases. I call this the average absolute risk difference (AARD), which is analogous to the ARD for the single cutoff method. For missing data rules that use quartile points (i.e., three cutoff points), AARD is

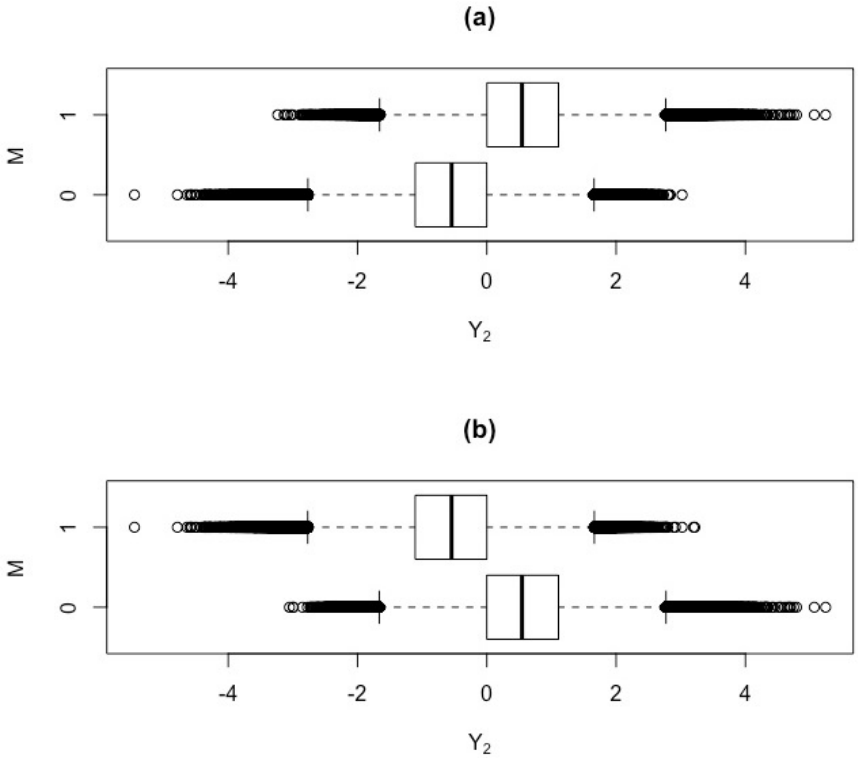
$$\text{AARD} = \frac{|\pi_4 - \pi_1|}{3}, \tag{21}$$

where π_4 and π_1 are defined in (15).

Similar to the single cutoff method, as the AARD increases, the strength of dependency increases. With quartile points, the maximum AARD is $1/3 = 0.33$. In this case of maximum AARD, the parameters need to be set as $\pi_1 = 0$, $\pi_2 = 0.33$, $\pi_3 = 0.67$ and $\pi_4 = 1$. Figure 4a shows the relationship between the missing data predictor Y_2 and the missing data indicator M for our previous example with AARD $= 0.1$, and Figure 4b shows the relationship between Y_2 and M for the example with the maximum AARD (i.e., AARD $= 0.33$). As expected, as the strength of dependency increases (i.e., comparing Figure 4a and 4b), the boxplot for $M = 1$ overlaps less with the one for $M = 0$. However, with the multiple cutoff method, it is no longer possible to obtain the case where the boxplot for $M = 1$ is completely separate from the boxplot for $M = 0$ (as shown in Figure 3a); this means that with the multiple cutoff method, researchers can never achieve the strongest dependency which they can do with the single cutoff method. In fact, as the number of cutoff points increases, the maximum strength of dependency researchers can create decreases. The reason is that the possible range of the probability of missing values is from 0 to 1, and as the number of cutoff points increases, researchers need to divide this range into smaller and smaller pieces, thus the maximum AARD decreases.

Another way to measure the strength of dependency is to build a logistic regression model using the missing data predictor V to predict the missing data indicator M, and then use the regression coefficient from the model to determine the strength of dependency. When researchers use this logistic regression model, they need to make two assumptions: 1) V is a continuous variable, and 2) the re-

Figure 5 ■ MAR data that are generated using the percentile method. M is the missing data indicator with $M = 1$ indicating Y_1 is missing, and $M = 0$ indicating Y_1 is not missing. Y_2 is the missing data predictor, which follows the standard normal distribution. Since M is a binary variable while Y_2 is a continuous variable, boxplots can be used to show the dependency between M and Y_2. In graph (a), the relationship between the probability of being missing from Y_1 and the percentile rank on Y_2 is a direct relationship; in graph (b), this relationship is an inverse relationship. Each graph is based on a large simulated dataset.



lationship between V and log-odds of M is linear. Since V is really an ordinal variable that follows a uniform distribution, these two assumptions are violated. As a result, researchers can only *approximate* the relationship using a logistic regression model. For the missing data rule in (15), the approximate logistic regression model is

$$\log\left(\frac{P(M=1)}{1-P(M=1)}\right) = -1.25 + 0.42V, \qquad (22)$$

where the regression coefficients are obtained by fitting a straight line describing the relationship between V and log-odds of M. As the the regression coefficient for V increases, the strength of dependency increases. However, similar to the single cutoff method, in the case of the maximum strength of dependency (i.e., when AARD $= 0.33$), the logistic regression model cannot be estimated because the log-odds of M for $V = 4$ (or for $V = 1$) is not defined.[4]

In conclusion, researchers can use AARD or the coefficient from the logistic regression model to measure the strength of dependency when using the multiple cutoff method to specify missing data rules. A higher AARD or regression coefficient value indicates a higher strength of dependency; however, in the case of the maximum dependency, researchers can only calculate AARD as the regression coefficient is undefined.

Implementing Missing Data Rules for the Multiple Cutoff Method. In the missing data literature, to implement missing data rules associated with the multiple cutoff method, researchers usually just delete the desired percentage of subjects that are below the lowest cutoff or above the highest cutoff or between two cutoffs (e.g., Strike et al., 2001; Graham, 2010). For example, to implement the

[4] If the parameters in (15) are set as $\pi_1 = 0$, $\pi_2 = 0.33$, $\pi_3 = 0.67$ and $\pi_4 = 1$, when $V = 4$, the log-odds is $\log\left(\frac{P(M=1)}{1-P(M=1)}\right) = \log(\frac{1}{0})$, which is undefined. If the parameters are set as $\pi_1 = 1$, $\pi_2 = 0.67$, $\pi_3 = 0.33$ and $\pi_4 = 0$, then the log-odds is undefined for $V = 1$.

missing data rule in (15), researchers will delete 30% of subjects with $Y_2 < Q_1$, 40% of subjects with $Q_1 \leq Y_2 < Q_2$, and so on. With this method, the estimated values for the parameters π_1, π_2, π_3 and π_4 are held constant across the datasets; in other words, there is no sampling variability for the parameter estimates. This issue may be a problem if researchers want to study small samples. Once again, I recommend researchers specify a missing data rule and then apply this missing data rule to every subject in the dataset.

Percentile Method

Missing Data Rules for the Percentile Method. The percentile method is an extension of the multiple cutoff method. In the percentile method, each subject's probability of being missing depends on its percentile rank in the missing data predictor, therefore, it can be viewed as the multiple cutoff method where each subject has its own cutoff point based on its percentile rank.

To define the missing data rule formally, suppose that a subject's probability of being missing from Y_1 is related to their percentile rank on the missing data predictor Y_2. Again, let M be the missing data indicator. If there is a direct relationship between the missing data indicator and the missing data predictor, then the missing data rule is *if a subject is at k^{th} percentile on Y_2, then it has k% probability of being missing from* Y_1 or $P(M = 1|Y_2 = q_k) = k/100$ where q_k is the Y_2 value corresponding to its kth percentile. If there is an indirect relationship, then the missing data rule is *if a subject is at k^{th} percentile on Y_2, then it has $(100 - k)$% probability of being missing from* Y_2 or $P(M = 1|Y_2 = q_k) = 1 - k/100$. These two missing data rules are the only possible missing data rules associated with the percentile method. Since the percentile method only involves these two missing data rules, there are no parameter values researchers need to consider when generating MAR data using the percentile method.

To calculate the probability of missing data, researchers need to determine the distribution of the percentile ranks of Y_2. According to the universality of the uniform, when plugging any continuous random variable into its own cumulative distribution function (CDF), I get a standard uniform distribution:

$$F(Y_2) \sim \text{Unif}(0, 1). \tag{23}$$

Since CDF is a function that maps a value of a random variable to its percentile rank, this means the percentile ranks of all possible Y_2 values are distributed as the standard uniform distribution. As a result, the expected percentile rank of a subject is the 50th percentile; thus, the probability of missing data is always 50% or $P(M = 1) = 0.5$. Let n be sample size and K be the number of subjects with missing data. We know $K \sim \text{Bin}(n, 0.5)$. Therefore, the expected percentage of missing values is

$$E(\frac{K}{n}) = 0.5. \tag{24}$$

The variance of the estimated percentage over repeated samples is

$$\text{Var}(\frac{K}{n}) = \frac{0.5(1 - 0.5)}{n} = \frac{0.25}{n}. \tag{25}$$

The expected number of distinct missing patterns can be calculated by Equation (4) by setting $\pi = 0.5$.

With the percentile method, researchers cannot vary the strength of dependency. The reason is that the two missing data rules associated with the percentile method only vary in the direction of dependency between the missing data indicator and the missing data predictor, and do not vary in the strength of dependency. Since the percentile method can be viewed as the multiple cutoff method with a large number of cutoffs, the strength of dependency created by the percentile method is less than the maximum strength created by the single cutoff method (see Figure 3a) or by the multiple cutoff method with quartile cutoffs (see Figure 4b). Figure 5 shows the relationship between Y and M for the two missing data rules under the percentile method. As expected, relative to the boxplots in Figure 3a and 4b, the boxplots for $M = 0$ and $M = 1$ in Figure 5a or 5b have more overlap with each other.

Perhaps, one way to quantify the strength of dependency created by the percentile method is to find a logistic regression model that approximates the missing data rule. Based on a large sample simulation ($n = 1,000,000$) where Y_2 follows the standard normal distribution, if the probability of being missing from Y_1 is directly related to the percentile rank of Y_2, an approximate logistic regression model is

$$\log\left(\frac{P(M = 1)}{1 - P(M = 1)}\right) = 1.70Y_2. \tag{26}$$

If the probability of being missing from Y_1 is inversely related to the percentile rank of Y_2, then the logistic regression is the same as the above except that the coefficient 1.70 is replaced with -1.70. Equation (26) shows that the missing data rule specified using the percentile method can also be specified using the logistic regression method (which will be explained in the next section). Therefore, the percentile method can also be viewed as a part of the logistic method.

The advantage of using the percentile method is that the probability of missing values gradually increases or decreases as the value of the missing data predictor Y_2 increases. This gradual change in probability as Y_2 is more realistic than the sudden change in probability as Y_2 passes a certain cutoff, which is used in the single or multiple cutoff method. However, I do not recommend the percentile

method to generate MAR data because this method does not allow researchers to vary the strength of dependency and the expected percentage of missing data. Alternatively, according to Equation (26), researchers can use the logistic regression method to generate MAR data equivalent to those created by the percentile method. With the logistic regression method, researchers can vary the strength of dependency and the percentage of missing values (see the next section for details).

Implementing Missing Data Rules for the Percentile Method. In the missing data literature, to implement the missing data rule associated with the percentile method, researchers usually apply the missing data rule in an ascending order according to the value of the missing data predictor (i.e., from the lowest Y_2 value to the highest Y_2 value) until the desired percentage of missing data is reached (e.g., Enders, 2001a). This implementation is very problematic. If researchers believe that each subject comes from the same population and thus follows the same missing data rule, it does not make sense that they apply the missing data rule to only a fraction of the subjects. As I have mentioned above, the expected percentage of missing data is 50% when the percentile method is used. However, with this implementation, the percentage of missing data in a dataset is commonly set to 5% or 15%, which is highly unlikely given this missing data rule.

If researchers want to use the percentile method, they should apply the missing data rule to each subject. In addition, they should calculate each subject's percentile rank on Y_2 based on the population distribution of Y_2, not based on the sample distribution of Y_2 values (see our OSF website for sample R code).

Logistic Regression Method

As shown before, the missing data rules associated with the single cutoff, multiple cutoff, and percentile methods can be reframed as logistic regression models (see Equations (13), (22), and (26)). In other words, the single cutoff, multiple cutoff, and percentile methods are all related to the logistic regression method for generating MAR data. In this section, I explain how to directly use logistic regression models to generate MAR data.

Missing Data Rules for the Logistic Regression Method. When using the logistic regression method to generate MAR data, researchers can view the logistic regression model as the missing data rule, and the population regression coefficients associated with the model as the parameters of the missing data rule. For example, if each subject's probability of being missing from Y_1 is related to the missing data predictor Y_2, then the logistic regression model for subject i is

$$\log\left(\frac{P(M_i=1|y_{2,i})}{1-P(M_i=1|y_{2,i})}\right)=\beta_0+\beta_1 y_{2,i}, \tag{27}$$

where M is the missing data indicator and $y_{2,i}$ is subject i's value on Y_2. The parameters associated with the missing data rule are β_0 and β_1.[5] Conditional on the value of Y_2, each subject's (or subject i's) probability of being missing is given by

$$P(M_i=1|y_{2,i})=\frac{1}{1+e^{-\beta_0-\beta_1 y_{2,i}}}. \tag{28}$$

Because the above function is continuous, it means the probability of being missing for Y_1 gradually increases or decreases as the value of Y_2 increases, an advantage shared with the percentile method.

With the logistic regression, there is no simple formula for calculating the expected percentage of missing data.[6] Researchers can use computer simulation to estimate the expected percentage of missing by calculating the mean of the probabilities in a sample with a large sample size (e.g., $n=100,000$):

$$\pi_{miss}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+e^{-\beta_0-\beta_1 y_{2,i}}}. \tag{29}$$

In terms of the strength of dependency, higher β_1 values indicate stronger dependency between Y_2 and M. However, the logistic regression model cannot be estimated when $P(M=1|y_2)=1$ or $P(M=0|y_2)=1$ because the log-odds of M is undefined or equals to infinity in those cases. This means researchers cannot generate MAR data with the strongest dependency with the logistic regression method.

In conclusion, researchers can generate MAR data by specifying a logistic regression model that predicts the probability of missing values given the value of the missing data predictor. The advantage of this method is that the probability of missing values gradually changes as the value of the missing data predictor changes, creating a more realistic situation relative to the single cutoff and multiple cutoff methods. However, the disadvantage of the logistic regression method is that it does not allow researchers to set a very strong dependency between the missing data indicator and predictor.

Implementing Missing Data Rules for the Logistic Regression Method. In the missing data literature in psychological sciences, researchers rarely generate missing

[5] With sample data, the regression coefficients and the variances associated with the coefficients can be estimated using the maximum likelihood method. More details can be found in any textbook on logistic regression (e.g., Hilbe, 2009).

[6] The reason is that it is hard to solve $P(M=1)=E(\frac{1}{1+e^{-\beta_0-\beta_1 Y_{2,i}}})$ analytically since it involves finding the expected value of a nonlinear transformation of a random variable.

data directly from a logistic regression model. Researchers in the statistics area are more likely to use logistic regression models to generate missing data (e.g., White & Carlin, 2010; Preisser et al., 2002; Miao et al., 2016). In addition, the *mice* package in the computer software R has a function called *ampute* that generates missing data using the logistic regression model (van Buuren & Groothuis-Oudshoorn, 2011). However, this package does not provide much information regarding its algorithm for generating missing data. To directly use a logistic regression model to generate missing data, I recommend researchers calculate each subject's probability of being missing based on the logistic regression model they have specified, and then apply the corresponding probability of being missing to each subject (see our OSF website for sample R code).

Advantages and Disadvantages of Different Types of MAR Missing Data Rules and Algorithms. Comparing the different types of MAR missing data rules and algorithms, there are several important advantages and disadvantages to highlight. Table 4 summarizes these advantages and disadvantages. First, the main advantage of using the single cutoff method is the only method that can produce the strongest dependency between the missing data indicator and the missing data predictor; in other words, it is the only method that can produce the *most* MAR data, enabling researchers to maximize the difference between MCAR and MAR data, creating the strongest manipulation of missing data mechanisms. This advantage of the single cutoff method is one of the reasons why it is the most popular method for generating MAR data. Another reason for its popularity is that the single cutoff method is very easy to understand and implement.

However, the main disadvantage of using the single cutoff method is that the MAR data it generates are not very realistic because it is hard to imagine that with real-life data, the probability of missing data depends on *one* single cutoff point of the missing data predictor. In contrast, the percentile and the logistic regression methods, although cannot generate the *most* MAR data, can generate more realistic MAR data, where the probability of missing data gradually increases or decreases as the value of the missing data predictor increases. As mentioned previously, one main disadvantage of the percentile method is that it can only create 50% missing data; therefore, if researchers want to create more realistic MAR data, I recommend using the logistic regression method, which can also approximate the percentile method according to Equation 26.

Finally, the multiple cutoff method has advantages and disadvantages that fall between the single cutoff and logistic regression methods. Since the multiple cutoff method involves specifying two or more cutoff points and allows the probability of missing data to change across the multiple cutoff points, it generates MAR data that are more realistic than those by the single cutoff method; on the other hand, it creates less realistic MAR data generated by the logistic regression method, where the probability of missing data varies gradually with the value of the missing data predictor. Furthermore, another unique advantage of the multiple cutoff method is that it makes it easier to create nonlinear relationship between the missing data indicator and the missing data predictor, usually a curvilinear relationship where the probability of missing data depends on the missing data predictor in one way for the high and low values of the missing data predictor but in another way for the middle values of the missing data predictor.

In conclusion, each type of missing data rule and algorithm for generating MAR data comes with its own advantages and disadvantages. If researchers wish to maximize the difference between MCAR and MAR data, I recommend using the single cutoff method; if researchers wish to create more realistic MAR data, I recommend using the logistic regression method.

Generating MNAR Missing Data

MNAR data are less studied in the missing data literature relative to the MCAR and MAR data because most statistical methods for handling missing data are unable to handle MNAR data. Generating MNAR data is very similar to generating MAR data. Recall that the only difference between MAR and MNAR data is that in MAR data, the probability of missing values for one variable depends on the observed values of another variable, but in MNAR data, the probability of missing values *depends on the variable's own underlying missing value*. Therefore, when generating MNAR missing data for simulation studies, researchers can change the missing data predictor to the variable with missing values, and then use one of the methods for generating MAR data to generate MNAR data. For example, suppose a missing data rule that generates MAR data says *when* Y_2 *value is above a cutoff point* a, Y_1 *has* π_1 *probability of being missing, otherwise,* Y_1 *has* π_2 *probability of being missing.* To change this MAR missing data rule to one that generates MNAR data, researchers simply have to change the variable Y_2 to Y_1; therefore, the corresponding missing data rule for generating MNAR data is *when* Y_1 *value is above a cutoff point* a, Y_1 *has* π_1 *probability of being missing, otherwise,* Y_1 *has* π_2 *probability of being missing.* In summary, by changing the missing data predictor to the variable with missing data, researchers can change all the MAR missing data rules to MNAR missing data rules and then generate MNAR missing data accordingly. This implies that the advantages and disadvantages for each type of MAR missing data rules also apply to for MNAR missing data (see Table 4).

Other Important Considerations for Conducting Simulation Studies Involving Missing Data

Beyond applying missing data rules to generate missing data, conducting simulation studies with missing data involves many other important considerations. Morris et al., 2019 provided a comprehensive tutorial on how to conduct simulation studies to evaluate statistical methods. The general procedure and important considerations explained in Morris et al., 2019 are also relevant to simulation studies involving missing data; therefore, readers who wish to gain more understanding of simulation studies are encouraged to refer to Morris et al., 2019. In this section, I highlight a few important considerations specifically relevant to simulation studies involving missing data.

First, the correlations among the variables in the dataset may affect how well missing data techniques (e.g., FIML and MI) handle MAR data. On the one hand, as mentioned previously, the correlations among variables may affect the number of missing data patterns in a MAR dataset. Specifically, if researchers want to create more missing data patterns by letting the different missing data indicators depend on different missing data predictors, then the more correlated the missing data predictors become, the fewer the number of missing data patterns will be. The number of missing data patterns, in turn, may affect the performance of missing data techniques (Savalei & Bentler, 2005; Zhang & Savalei, 2020).

On the other hand, the correlation between the missing data predictor and the variable with missing data may affect how well missing data techniques such as MI predict the values of the missing data in a MAR dataset. In the special case of uncorrelated MAR data, the correlation between the missing data predictor and the variable with missing data is zero, but the probability of missing values is related to the values of the missing data predictor. For example, suppose Y_2 is the missing data predictor such that for subjects with $Y_2 \geq 0$, their Y_1 values are missing (i.e., single cutoff method with the strongest dependency), but Y_2 and Y_1 has a correlation of zero. In this case, subjects with missing values on Y_1 have high values on Y_2 but had researchers observed their values on Y_1, the distribution of their Y_1 values would be the same as the one for the subjects without missing values. In other words, given the Y_2 values, it is possible to predict which subjects have missing values on Y_1 but not their missing values of Y_1. Although relative to MCAR data, which cannot even predict which subjects have missing data, uncorrelated MAR data provide researchers with slightly more information about the variable with missing data relative to MCAR data, they provide less information about the missing data relative to MAR data where the missing data predictor and the variable with missing data are moderately or strongly correlated. Therefore, if researchers wish to generate MAR data that are more different from MCAR data, I recommend researchers generate correlated MAR data.

The second factor that may affect the performance of missing data techniques is the location of the variables with certain properties (e.g., variables with model misfit or variables with nonnormality) relative to the location of the variables with missing data. When there are missing data, researchers loss information about the features of the data that have missing values.[7] Therefore, the location of the variables with certain properties may interact with the location of the variables with missing data to affect the performance of missing data techniques. For example, Zhang and Savalei, 2020 showed that when the variables that are misspecified are the same as those with missing data (i.e., the location of the model misfit overlaps with the location of missing data), the model fit improves relative to the fit for data without missing values because some of the information regarding the model misfit is lost due to missing data. In contrast, when the variables that are misspecified are different from the ones with missing data, the model fit does not change much because the information regarding model misfit is not affected by the missing data. Of course, depending on the purpose of the simulation study, researchers may only be interested in a small number of properties of the data; nonetheless, when designing the study, they should think carefully about how the location of these properties may interact with the location of missing data.

Third, when conducting simulation studies with missing data, it is important to decide the percentage of missing data in each variable. Common percentages of missing data used in simulation studies are 5%, 15%, 20%, 25%, 30%, and 50% per variable with missing data (e.g., Zhang & Savalei, 2023; Savalei & Bentler, 2009; Yuan & Savalei, 2014; Enders, 2001b). According to Peugh and Enders, 2004, in psychological and educational research, the percentage of missing data can range from 1% to 67%, with a mean of 7.60% ($SD = 8.07\%$), therefore, it makes sense that almost all previous simulation studies' percentages of missing data ranged from 5% to 50%.

Fourth, throughout the paper, I have emphasized that missing data rules are the underlying statistical models used to generate missing data; this implies that if the missing data rule is applied to the population level, researchers can theoretically envision a population dataset with missing values. Therefore, when conducting simulation stud-

[7] I am not using the term "information" in a technical sense (e.g., it does not mean Fisher information). I use 'information" in a loose sense to mean things about the dataset (e.g., covariance structure of the data) that will allow us to predict values of missing data.

ies to evaluate the effectiveness of missing data techniques, researchers may consider first conducting at the population level so that they can show a proof of concept that the missing data technique is at least effective at the population level. It is important to first demonstrate the effectiveness of the missing data technique at the population level because if the missing data technique does not work at the population level, it definitely will not work at the sample level. In previous simulation studies that include a population-level simulation (e.g., Zhang & Savalei, 2023, 2020; Savalei & Bentler, 2005), researchers usually first generate a dataset with a very large sample (e.g., $n = 1,000,000$) to mimic the population data, then create missing values within this large sample, and finally evaluate the missing data technique.

Finally, conducting simulation studies involving missing data also requires researchers to make decisions regarding the number of simulation runs, sample sizes needed for each missing data condition, how to handle nonconvergency issues, etc. Decisions regarding these issues should be handled similarly to those with complete data (see Morris et al., 2019, for a detailed tutorial). In terms of the simulation runs, if researchers conduct the simulation study at the population level, only one simulation run with a very large sample is needed (e.g., Zhang & Savalei, 2023). If researchers conduct the simulation study at the sample level, the number of simulation runs is usually around 1000 in previous studies (e.g., Zhang & Savalei, 2023), which is the same for simulation studies with complete data (e.g., Morris et al., 2019). For sample-level simulation studies (e.g., Enders, 2001b; Savalei & Bentler, 2009; Zhang & Savalei, 2023), the sample size was usually manipulated to vary from 200 to 500, which are the common sample sizes seen in psychological and educational research and used in simulation studies with complete data. In conclusion, in this section, I have discussed several important considerations for conducting simulation studies involving missing data that are not related to generating missing data. Of course, the factors that affect the results of simulation studies are not limited to those discussed in this section, but the main message is that researchers should also pay careful attention to aspects of simulation studies unrelated related to missing data generation when they design their simulation studies.

Summary and Final Recommendations

Simulation studies play a crucial part in the development and evaluation of many statistical methods, including statistical techniques for handling missing data (e.g., FIML or MI). To conduct simulation studies involving missing data, researchers must sample data from a known population distribution and then generate missing data in the sample data (i.e., deciding which values to delete in the data). The main purpose of the current paper is to provide guidelines on generating missing data for simulation studies, which have never been done in past research. Specifically, I have provided detailed explanations regarding the statistical models, also known as "missing data rules", for generating missing data with different missing data mechanisms and patterns. For each type of missing data rules, I have also explained the computer algorithm that can implement the rules and provided R code for algorithms. I conclude the paper by providing the following summary of recommendations for generating missing data for simulation studies.

- Researchers should always specify the missing data rule and identify the parameters associated with the rule before generating missing data on the computer. Knowing the specific missing data rule makes it easier for researchers to figure out and understand the missing data properties, such as the expected percentage of missing values, the type of missing mechanism, and the number of missing data patterns.
- Researchers should apply the missing data rule subject by subject when generating missing data on the computer. It is the easiest and most straightforward way to apply the missing data rule to generate missing data.
- Researchers should maximize the difference between MCAR and MAR data to achieve a strong manipulation of the type of missing data mechanism. To maximize the difference between MCAR data and MAR data, I suggest that researchers include a MAR dataset with the strongest dependency (between the missing data indicator and the missing data predictor) using the single cutoff method, and make sure that for all MAR data, there is a moderate correlation between the missing data predictor and the variable with missing data (i.e., avoid uncorrelated MAR data).
- If researchers wish to include more realistic MAR data that do not involve sudden changes in the probability of missing values as the value of the missing data predictor increases, I suggest that they generate MAR data using the logistic regression method rather than the percentile method because the percentile method does not allow researchers to manipulate the strength of dependency between the missing data indicator and the missing data predictor.
- If researchers want to manipulate the type of missing data mechanism, they should control for the number of missing data patterns between conditions with different missing data mechanisms. In other words, they should compare MCAR and MAR data with approximately the same number of missing data patterns.
- When conducting a simulation study involving miss-

ing data, researchers also need to take into considerations others aspects of the simulation study beyond generating the missing data. In this paper, I highlighted a few important considerations including correlations among variables, locations of variables with certain properties, percentages of missing data, population- versus sample-level simulation, etc. Many aspects of simulation studies with missing data are also similar to those with complete data. Readers who want to get more guidance on how to conduct a simulation study are recommended to refer to Morris et al., 2019.

References

Agresti, A., & Kateri, M. (2011). *Categorical data analysis*. Springer.

Allison, P. D. (2000). Multiple imputation for missing data: A cautious tale. *Sociological Methods & Research*, *28*, 301–309. doi: 10.1177/0049124100028003003.

Arbuckle. (1999). *Full information estimation in the presence of incomplete data*. Lawrence Erlbaum.

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

De Raadt, A., Warrens, M. J., Bosker, R. J., & Kiers, H. A. (2019). Kappa coefficients for missing data. *Educational and Psychological Measurement*, *Advanced Online Publication*. doi: 10.1177/0013164418823249.

Enders, C. K. (2001a). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, *6*, 352–370. doi: 10.1037/1082-989X.6.4.352.

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psychological Measurement*, *64*, 419–436. doi: 10.1177/0013164403261050.

Enders, C. K. (2001b). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, *61*(5), 713–740. doi: 10.1177/00131640121971482.

Enders, C. K. (2010). *Applied missing data analysis*. The Guilford Press.

Fleishman, A. I. (1978). A method for simulating nonnormal distributions. *Psychometrika*, *43*(4), 521–532. doi: 10.1007/BF02293811.

Foldnes, N., & Olsson, U. H. (2016). A simple simulation technique for nonnormal data with prespecified skewness, kurtosis, and covariance matrix. *Multivariate behavioral research*, *51*(2-3), 207–219. doi: 10.1080/00273171.2015.1133274.

Graham, J. W. (2010). *Missing data: Analysis and design*. Springer.

Grigsby, T. J., & McLawhorn, J. (2019). Missing data techniques and the statistical conclusion validity of survey-based alcohol and drug use research studies: A review and comment on reproducibility. *Journal of Drug Issues*, *49*(1), 44–56. doi: 10.1177/0022042618795878.

Hilbe, J. M. (2009). *Logistic regression models*. Chapman; Hall/CRC.

Jamshidian, M., & Siavash, J. (2010). Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika*, *75*, 649–674. doi: 10.1007/s11336-010-9175-3.

Jia, F., & Wu, W. (2019). Evaluating methods for handling missing ordinal data in structural equation modeling. *Behavior research methods*, *Advanced Online Publication*. doi: 10.3758/s13428-018-1187-4.

Kim, K. H., & Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, *67*, 609–624. doi: 10.1007/BF02295134.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Mattson, S. (1997). How to generate non-normal data for simulation of structural equation models. *Multivariate Behavioral Research*, *32*(4), 355–373. doi: 10.1207/s15327906mbr3204_3.

Miao, W., Ding, P., & Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, *111*(516), 1673–1683. doi: 10.1080/01621459.2015.1105808.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. doi: 10.1002/sim.8086.

Musil, C. M., Warner, C. B., Yobas, P. K., & Jones, S. L. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Western Journal of Nursing Research*, *24*, 815–829. doi: 10.1177/019394502762477004.

Nakagawa, S. (2015). Missing data: Mechanisms, methods, and messages. In G. A. Fox, S. Negrete-Yankelevich, & V. J. Sosa (Eds.), *Ecological statistics:contemporary theory and application* (pp. 81–105). Oxford Scholarship Online. doi: 10.1093/acprof:oso/9780199672547.003.0005.

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation

of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, *6*, 328–362. doi: 10.1177/1094428103254673.

Olvera Astivia, O. L., & Zumbo, B. D. (2015). A cautionary note on the use of the vale and maurelli method to generate multivariate, nonnormal data for simulation purposes. *Educational and Psychological Measurement*, *75*(4), 541–567. doi: 10.1177/0013164414548894.

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*(4), 525–556. doi: 10 . 3102 / 00346543074004525.

Preisser, J. S., Lohman, K. L., & Rathouz, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, *21*, 3035–3054. doi: 10.1002/sim.1241.

Reinartz, W. J., Echambadi, R., & Chin, W. W. (2002). Generating non-normal data for simulation of structural equation models using Mattson's method. *Multivariate Behavioral Research*, *37*(2), 227–244. doi: 10.1207/S15327906MBR3702_03.

Rubin, J. B. (1976). Inference and missing data. *Biometrika Trust*, *63*, 581–592. doi: 10.1093/biomet/63.3.581.

Savalei, V., & Bentler, P. M. (2005). A statistically justified pairwise ml method for incomplete nonnormal data a comparison with direct ml and pairwise adf. *Structural Equation Modeling*, *12*, 183–214. doi: 10 . 1207 / s15328007sem1202_1.

Savalei, V., & Yuan, K. H. (2009). On the model-based bootstrap with missing data: Obtaining a p-value for a test of exact fit. *Multivariate Behavioral Research*, *44*, 741–763. doi: 10.1080/00273170903333590.

Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 477–497. doi: 10 . 1080 / 10705510903008238.

Schafer, J. L., & Graham, J. W. (2002a). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. doi: 10.1037//1082-989X.7.2.147.

Schafer, J. L., & Graham, J. W. (2002b). Missing data: Our view of the state of the art. *Psychological Methods*, *8*, 147–77. doi: 10.1037/1082-989X.7.2.147.

Strike, K., Emam, K. E., & Madhavji, N. (2001). Software cost estimation with incomplete data. *IEEE Transactions on Software Engineering*, *27*, 890–908. doi: 10.1109/32.962560.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. doi: 10 . 18637/jss.v045.i03.

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, *28*, 2920–2931. doi: 10.1002/sim.3944.

Yuan, K. .-., & Savalei, V. (2014). Consistency, bias and efficiency of the normal-distribution-based mle: The role of auxiliary variables. *Journal of Multivariate Analysis*, *6*, 353–370. doi: 10.1002/wics.1287.

Yuan, K. .-., Tong, X., & Zhang, Z. (2015). Bias and efficiency for sem with missing data and auxiliary variables: Two-stage robust method versus two-stage ml. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 178–192. doi: 110.1080/10705511.2014.935750.

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165–200. doi: 10.1111/0081-1750.00078.

Zhang, X., & Savalei, V. (2020). Examining the effect of missing data on the population RMSEA and CFI under full information maximum likelihood. *Structural Equation Modeling*, *27*(2), 219–239. doi: 10.1080/00273171.2015.1133274.

Zhang, X., & Savalei, V. (2023). New computations for rmsea and cfi following fiml and ts estimation with missing data. *Psychological Methods*, (2), 263–283. doi: 10.1037/met0000445.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on osf.io/pmn9z/

Citation

Zhang, X. (2023). How to generate missing data for simulation studies. *The Quantitative Methods for Psychology*, *19*(2), 100–122. doi: 10.20982/tqmp.19.2.p100.

Table 4 ■ Summary of missing data rules and algorithms for generating MCAR, MAR, MNAR data

Type of Missing Data Rule and Algorithm	Example	Advantages	Disadvantages
		Generating MCAR Data	
General MCAR: involves specifying a certain percentage of missing data in a variable.	Each subject has 20% probability of being missing on Y_1, where Y_1 is the variable with missing data.	● Easy to understand and implement	● MCAR data are special cases of MAR or MNAR data; thus, researchers cannot only generate MCAR data if they want to study the effect of missing data mechanisms.
		Generating MAR Data	
Single Cutoff: involves specifying one cutoff point in each missing data predictor.	If a subject has $Y_2 \geq 1$, then their probability of being missing on Y_1 is 80%, otherwise, the probability of being missing is 20%, where Y_2 is the missing data predictor.	● Easy to understand and implement relative to other types of MAR missing data rules. ● Can create the strongest dependency between the missing data indicator and predictor. ● Easy to quantify the strength of dependency between missing data indicator and predictor using indices such as ARD or OR.	● Create unrealistic MAR data; real-life MAR data probably do not involve one single cutoff point.
Multiple Cutoffs: involves specifying multiple cutoff points in each missing data predictor.	If a subject's Y_2 value falls into the 1st, 2nd, 3rd or 4th quartiles of Y_2 values, then its probability of being missing from Y_1 is 0.3, 0.4, 0.5 or 0.6, respectively.	● Can create nonlinear relationship between the missing data indicator and predictor. ● Relative to the single cutoff method, generates more realistic MAR data where the percentage of missing data gradually changes as the value of the missing data predictor increases.	● Cannot create the strongest dependency between the missing data indicator and predictor. ● Relative to the single cutoff method, harder to quantify the strength of dependency between the missing data indicator and predictor. ● Relative to the percentile and logistic regression methods, generate more unrealistic MAR data based on two or more cutoff points.
Percentile: involves specifying each subject's probability of being missing to be their percentile rank in the missing data predictor.	If a subject is at k% percentile on Y_2, then it has k% probability of being missing on Y_1.	● The probability of missing data changes as the value of the missing data predictor increases, creating more realistic missing data relative to the single and multiple cutoff methods.	● Can only produce 50% of missing data. ● Cannot vary the strength of dependency between the missing data indicator and predictor.
Logistic Regression: involves specifying a logistic regression that describes the relationship between the missing data indicator and predictor.	Each subject's probability of being missing from Y_1 is related to the Y_2 according to the logistic regression model: $\log\left(\frac{P(M_i=1\|Y_2)}{1-P(M_i=1\|Y_2)}\right) = 1.8Y_2$, where M is the missing data indicator.	● The probability of missing data changes as the value of the missing data predictor increases, creating more realistic missing data relative to the single cutoff and multiple cutoff methods. ●Can vary the strength of dependency between the missing data indicator and predictor, an advantage relative to the percentile method.	● Cannot create the strongest dependency between the missing data indicator and predictor.
		Generating MNAR Data	
Types for MNAR: the same as the types of MAR missing data rules. The only difference is that the missing data predictor is the same as the variable with missing data.	The same as the ones for MAR data, except replacing the missing data predictor Y_2 with the variable with missing data Y_1.	● Share the same advantages as the types of MAR missing data rules	● Share the same disadvantages as the types of MAR missing data rules ● May not be included in simulation studies that examine the effectiveness of missing data techniques because most missing data techniques (e.g., FIML and MI) cannot handle MNAR data theoretically.

Note: ARD: absolute risk difference; OR: odd ratio; FIML: Full Information Maximum Likelihood; MI: Multiple Imputation

Received: 06/11/2022 ∼ Accepted: 25/05/2023