Structural Equation Modeling: A Multidisciplinary Journal, 22: 631–642, 2015

ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2014.937378

Routledge
Taylor & Francis Group

Graphical Representation of Missing Data Problems

Felix Thoemmes[1] and Karthika Mohan[2]
[1]*Cornell University*
[2]*University of California, Los Angeles*

Rubin's classic missingness mechanisms are central to handling missing data and minimizing biases that can arise due to missingness. However, the formulaic expressions that posit certain independencies among missing and observed data are difficult to grasp. As a result, applied researchers often rely on informal translations of these assumptions. We present a graphical representation of missing data mechanism, formalized in Mohan, Pearl, and Tian (2013). We show that graphical models provide a tool for comprehending, encoding, and communicating assumptions about the missingness process. Furthermore, we demonstrate on several examples how graph-theoretical criteria can determine if biases due to missing data might emerge in some estimates of interests and which auxiliary variables are needed to control for such biases, given assumptions about the missingness process.

Keywords: auxiliary variables, full information, graphical models, maximum likelihood, missing data, multiple imputation

The classic missingness mechanisms by Rubin (1976) define how analysis variables and missingness relate to each other. Many researchers have an intuitive understanding about these mechanisms, but lack knowledge about the precise meaning of the conditional independencies that are expressed in Rubin's taxonomy. In this article, we first review classic missingness mechanisms and discuss how the conditional independencies that define those mechanisms can be encoded in a graphical model. Graphs have been used informally in popular texts and articles to aid understanding of the mechanisms (Enders, 2010; Schafer & Graham, 2002) and to illustrate how missingness relates to other variables in a model. However, in previous publications, graphs were used simply as illustrations, whereas we use formal graph theory (Pearl, 2009) to encode the assumptions that are critical for techniques such as multiple imputation (MI), or full-information maximum likelihood (FIML). The use of such formal graphs can aid in thinking about missing data problems and can help researchers to formalize what relations among the observed, partially observed, and unobserved causes of missingness are pertinent for bias removal.

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

MISSING DATA MECHANISM

We begin by reviewing the classic mechanisms defined by Rubin (1976): missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). We note that NMAR is also often called missing not at random (MNAR). In our overview, we use a slightly modified version of the notation employed by Schafer and Graham (2002). We also express the equalities of probabilities that are used to describe the missingness mechanisms using conditional independence statements (Dawid, 1979), because these will map onto the graphical concept of d-separation that we employ later.

We denote an $N \times K$ matrix by D. The rows of D represent the cases $n = 1, \ldots, N$ of the sample and the columns represent the variables $i = 1, \ldots, K$. D can be partitioned into an observed part, labeled D_{obs}, and a missing part D_{mis}, which yields $D = (D_{obs}, D_{mis})$. Further, we denote an indicator matrix of missingness, R, whose elements take on values of 0 or 1, for observed or missing values of D, respectively. Accordingly, R is also an $N \times K$ matrix. Each variable in D can therefore have both observed and unobserved values.

MCAR

MCAR is the most restrictive assumption. It states that the unconditional distribution of missingness $P(R)$ is equal to the

conditional distribution of missingness given D_{obs} and D_{mis}, or simply D.

$$P(R \mid D) = P(R \mid D_{obs}, D_{mis}) = P(R) \quad (1)$$

These equalities of probabilities can be expressed as conditional independence statements, here in particular

$$R \perp\!\!\!\perp (D_{obs}, D_{mis}). \quad (2)$$

The MCAR condition is therefore fulfilled when the missingness has no relationship with (is independent of) both the observed and unobserved parts of D.

In an applied research context, we could imagine MCAR being fulfilled if the missing data arose from a purely "accidental" (random) process. In such an instance, missingness R would be completely independent of every observed or unobserved variable, as expressed in Equation 2. As an example of MCAR, a single item from an online questionnaire might be missing because a participant accidentally hit a button to submit an answer twice and therefore accidentally skipped a question. The reason this item is missing is based on a presumably purely random accident and is unrelated to other observed or unobserved variables. Another example might be a missing behavioral observation; for example, the view of a camera that was recording a playground was temporarily obstructed by another object. MCAR is rare in applied research and usually does not hold, unless it has been planned by the researcher in so-called missingness by design studies (Graham, Taylor, Olchowski, & Cumsille, 2006). When MCAR holds, even simple techniques like listwise deletion will yield consistent estimates (Enders, 2010); however it is generally not advisable to use these simple methods due to loss in statistical power. The modern approaches of MI and FIML are preferred, because their estimates will also yield consistent estimates without this loss of statistical power (Enders, 2010).

MCAR cannot be empirically verified (Gelman & Hill, 2007; Raykov, 2011), but examination of homogeneity of means and variances can at least refute that MCAR holds. Little (1988) provided a multivariate test of homogeneity, and Raykov, Lichtenberg, and Paulson (2012) discussed individual testing of homogeneity of means and variances with Type I error correction. Mohan and Pearl (2014) also provided a complete characterization of the refutable implications of MCAR. The inability to directly test MCAR can also be seen by the fact that it posits independence assumptions about quantities that are by definition unobserved, here in particular D_{mis}.

MAR

MAR is a somewhat less restrictive condition than MCAR. MAR states that the conditional probability of missingness, given the observed part D_{obs} is equal to the conditional probability of missingness, given the observed and the unobserved part (D_{obs}, D_{mis}).

$$P(R \mid D) = P(R \mid D_{obs}, D_{mis}) = P(R \mid D_{obs}). \quad (3)$$

These equalities of probabilities can be expressed as conditional independence statements, here in particular

$$R \perp\!\!\!\perp D_{mis} \mid D_{obs}. \quad (4)$$

In words, MAR states that missingness is independent of the unobserved portion of D, given information about the observed portion of D. Dependencies between the observed portion and missingness are allowed.

In an applied research context, we could imagine that missingness is caused by certain observed variables that might also have an effect on important analysis variables. For example, missingness on an achievement measure could be caused by motivation (or lack thereof). Further we can assume that motivation also has an effect on achievement. As long as motivation is observed and conditioned on, there is no more dependence between R and D_{mis}; they are conditionally independent of each other, as expressed in Equation 3. For MAR to hold, we have to observe and condition on those covariates that affect the causal missingness mechanisms. This might not often be easy to achieve in an applied setting, as presumably many variables might exhibit such a structure. MI and FIML will yield consistent results if MAR holds (Allison, 2001). Just as MCAR, MAR cannot be verified empirically either, as it also posits conditional independence assumptions among quantities that are by definition unobserved, specifically, D_{mis}. Recently, a refutation test has been suggested that tests whether data follow a condition labeled MAR^{+}. MAR^{+} always implies MAR, but the reverse is not true. Failure to reject MAR^{+} thus lends ample evidence that MAR also cannot be rejected (because the occurrence of MAR without MAR^{+} is rare). For details on testing MAR^{+} see Potthoff, Tudor, Pieper, and Hasselblad (2006) and Pearl and Mohan (2014).

NMAR

Finally, NMAR is the most problematic case. NMAR is characterized by the absence of any of the aforementioned equalities of probabilities or conditional independencies. That is,

$$P(R \mid D_{obs}, D_{mis}) \neq P(R \mid D_{obs}). \quad (5)$$

No conditional independencies are implied by Equation 5.

We discuss two cases in which NMAR could emerge. The first case emerges when particular values of a variable are associated with higher probabilities of missingness on the same variable. A typical example for NMAR is a situation

in which participants with very high incomes are unwilling to answer survey questions about their income, and are thus missing. In this case missingness is directly related to the variable with missing data and they are thus dependent on each other, as expressed in Equation 5. A second example in which NMAR is present are situations in which an unobserved variable has an effect on both the variable with missing data and its missingness mechanism. This unobserved variable could induce a dependency between missingness and the variable with missing values. In an applied research context, we could again imagine that motivation has an effect on test scores and whether or not missing data are observed, but in this case motivation has not been measured and is therefore a fully unobserved variable. In the more general case, a set of fully and partially observed variables might induce a dependency between causes of missingness (R_X) and the variable with missing values (X). Note that observing proxies (variables that are either causes of the unobserved variables, or are caused by the unobserved variable) can help mitigate the bias that is due to not observing the variables that induce dependencies. The bias-reducing properties of such proxy variables in the context of causal inference were discussed by Pearl (2010b).

The reason it is important to distinguish among these three mechanisms is that they prescribe different treatments of the missing data problem. If MCAR holds, listwise deletion yields consistent results (even though FIML or MI will still outperform listwise deletion in terms of statistical power and are thus preferred). If MAR holds, FIML and MI will yield consistent estimates. If NMAR holds, other special techniques need to be used. Those include approaches that estimate a separate model for the probability of being missing, or examine individual subsamples of cases that share the same pattern of missing data. For details on these models see Enders (2011), or Muthén, Asparouhov, Hunter, and Leuchter (2011). However, none of these approaches is guaranteed to yield consistent estimates in all NMAR situations (Mohan, Pearl, & Tian, 2013).

An applied researcher therefore needs to think about which mechanism might be present. One method that can aid in this deliberation is to use graphical models to display assumed relationships between fully observed variables, partially observed variables, unobserved variables, and missingness. We now present how missingness mechanisms can be displayed in graphs, and then explain how applied researchers can encode their assumptions in these graphs and determine what data analytic treatment is likely to be effective.

GRAPHICAL DISPLAYS OF MISSINGNESS MECHANISMS

The graphs that we are going to use are sometimes referred to as nonparametric structural equation models (because the arrows in the graphs do not imply linear, but functional relationships with unknown form; Pearl, 2010a), directed acyclic graphs (DAGs), or in the context of missing data, *m*-graphs (Mohan et al., 2013). The idea to represent missing data problems using graph theory was (to our knowledge) first briefly mentioned by Glymour (2006), and has also been used by Daniel, Kenward, Cousens, and De Stavola (2011), and Martel García (2013).

An *m*-graph consists of nodes that represent fully observed variables, partially observed variables, unobserved variables, and missingness information. In our graphs, fully observed variables are represented as solid rectangles. Observed variables are sometimes endowed with disturbance terms that represent other unobserved variables that have direct effects on this variable. Disturbance terms are displayed using the letter ε. Often, they are omitted for simplicity, but we show them explicitly in our graphs for completeness. Whenever it is necessary to explicitly show a fully unobserved variable, we do so by displaying it with a dashed circle. Partially observed variables (i.e., variables with missing data) are displayed in the following manner: Any variable that has missing data is shown with a dashed rectangle. The actually observed portion of this variable, however, is displayed in a proxy of this variable and is drawn with a solid rectangle. This proxy is further signified with a star ($\star$) symbol in its variable name. The proxy variable takes on the values of the variable in the dashed circle when R indicates that data are observed, and has missing data whenever R indicates that data are in fact missing. Information about missingness deviates slightly from the common notation used earlier that simply uses R as an indicator for missingness in the data. In *m* graphs, the nodes labeled R represent causal mechanisms that are responsible for whether a datum ultimately becomes observed or unobserved. In addition, we consider such mechanisms for every single variable and hence add a subscript to the nodes labeled R that shows which variable this missingness indicator is associated with. We do not explicitly portray R variables corresponding to fully observed variables in the graph since they are constants. We still might refer to the nodes labeled R as missingness *indicators*, with the understanding that this also refers to the causal mechanism responsible for missingness. Missingness indicators R are also endowed with disturbance terms that represent all additional and unobserved causal influences on missingness.

Observed variables, unobserved variables, disturbance terms, and missingness indicators can be connected in the graph by unidirected or bidirected arrows. Unidirected arrows represent assumed causal relations between variables, whereas bidirected arrows are a shorthand to express that one or more unobserved variables have direct effects on the variables connected with bidirected arrows. We use *m*-graphs to express the process by which variables in the model, including missingness indicators, R, obtain their values. In other words, one should think about an *m*-graph

as a data-generating model in which the values of each variable are determined by the values of all variables that have direct arrows pointing into that variable. To determine the statistical properties of the variables in the graph, we rely on the so-called *d-separation* criterion (Pearl, 1988), which determines whether two variables in a graph are statistically independent of each other given a set of other variables. The d-separation criterion forms the link between the missingness mechanisms depicted in the graph and the statistical properties that are implied by those mechanisms.

The d-Separation Criterion

Conditional independence in graphs, or d-separation (Pearl, 2010), can be derived from a DAG using a set of relatively simple rules. Two variables X and Y, could be connected by any number of paths in a graph. A path is defined as any sequence of directed or bidirected arrows that connect X and Y. It is not of importance for the definition of a path whether the individual segments of it have arrows pointing in one or the other direction. A path is defined to be open if it does not contain a so-called collider variable that has two arrows pointing into it; for example, $X \rightarrow C \leftarrow Y$. Any path that contains at least one collider is said to be closed. An open path induces a dependency between two variables, whereas a closed path does not induce any such dependency. Conditioning on a variable in a path that is not a collider closes (blocks) this path. Importantly, conditioning on a collider (or any variable that is directly or indirectly caused by a collider), on the other hand, opens a previously closed path. Two variables in a graph are d-separated if there exists a set of variables Z in the graph that blocks every open path that connects them. This set Z may be empty (implying unconditional independence). Likewise, two variables are said to be d-connected conditional on Z if and only if Z does not block every path between the two. Being d-connected implies that the two variables are stochastically dependent on each other. One way to determine whether two variables are d-separated would be to list all paths that connect two variables and determine whether each path is open or closed, given a conditioning set Z. In large graphs this can become time-consuming, if done by hand. Programs like DAGitty (Textor, Hardt, & Knüppel, 2011), DAG program (Knüppel & Stang, 2010), TETRAD (Scheines, Spirtes, Glymour, Meek, & Richardson, 1998), or the R package dagr (Breitling, 2010) automate this task. Readers who need more detailed information on how to apply the d-separation criterion are referred to Appendix A, which provides a small worked-out example of determining paths and checking whether they are open or closed. In addition, readers can consult the article by Hayduk et al. (2003) or the chapter by Pearl (2009), entitled "d-separation without tears." This chapter can be accessed online under *bayes.cs.ucla.edu/BOOK – 2K/d – sep.html.*

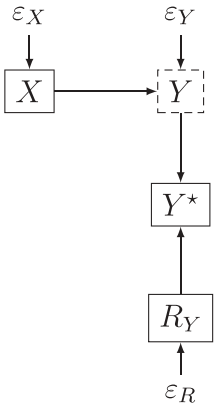

FIGURE 1 A simple missing completely at random model.

Graphical Display of MCAR

In Figure 1, we present a graphical display of MCAR for the simple case in which a single variable X has an effect on a single variable Y. In this simple case, X is completely observed and only Y suffers from missingness. We use a dashed rectangle to represent the variable Y that has missing data. Note that this should not be confused with a latent variable in structural equation modeling that is being estimated in a model. Whether data on Y are missing is determined by the variable R_Y in the graph. Note that the term ε_R denotes all possible causes of why the variable Y is missing. The proxy of Y is denoted as $Y^\star$ and is strictly a function of the underlying Y and the missingness indicator, and therefore has no ε term. We use an additional subscript for R to denote that this missingness indicator pertains only to variable Y. When R_Y takes on the value 0, $Y^\star$ is identical to Y, and if R_Y takes on the value 1, $Y^\star$ is missing. The graphical model allows that individual variables have different causes of missingness, meaning that we could consider a situation in which one variable has missingness that might be MCAR, whereas another variable might have missingness that would be considered NMAR.

In Figure 1, we can see that there is only a single arrow pointing to R_Y from the disturbance term ε_R, meaning that missingness arises only due to unobserved factors, contained in ε_R. Further, these unobserved factors have no association with any other variable or disturbance term in the model, as can be seen by the fact that ε_R is not connected to other parts of the model. We could also express this by stating that missingness is due to completely random and unobserved factors, all contained in ε_R. The single path that connects Y and R_Y via $Y^\star$ is blocked, because $Y^\star$ is a collider with two arrows pointing into it.

The important independence that we need to focus on is between variables that have missing data and their associated missingness indicators, in our example R_Y and Y. In this graph Y and R_Y (and X and R_Y) are said to be d-separated

without having to condition on any other variables, implying unconditional stochastic independence between the variables Y and R_Y. Note that this maps on the definition of MCAR as defined using conditional independence in Equation 2. To express this more generally, the missingness of a variable Y could be viewed as MCAR, whenever the missing data indicator R_Y is unconditionally d-separated from Y with missing data. If more than one variable exhibits missing data and we want to check whether MCAR holds for each of these variables as well, we simply need to check whether they are also unconditionally d-separated from their respective missingness indicators.

Graphical Display of MAR

To illustrate MAR, we employ the same example with two variables X and Y, in which only Y has missing data. The MAR condition (see Equation 4) implies the conditional independence $R_Y \perp\!\!\!\perp Y|X$. In Figure 2a we show the simple situation in which MAR holds, as long as X is observed and used in either FIML or MI. In Figure 2a, Y and R_Y are d-connected, via the open path $Y \leftarrow X \rightarrow R_Y$. However, if one conditions[1] on X, this path becomes blocked and Y and R_Y are now d-separated, implying conditional stochastic independence $R_Y \perp\!\!\!\perp Y|X$, as similarly defined in Equation 4, and therefore MAR holds.

In our second example in Figure 2b, we add an additional variable A. A represents a variable that might not be of substantive interest, but could aid in the estimation of missing data; for example, through virtue of making MAR more plausible, or by reducing variance and thus standard errors. Such a variable is usually referred to as an auxiliary variable. Auxiliary variables are typically correlated with the variable with missing data and missingness (Enders, 2010). In Figure 2b, Y and R_Y are d-connected via two paths, one traversing X, and the other one traversing A. Specifically, Y and R_Y are d-connected via the open path $Y \leftarrow A \rightarrow R_Y$ and via the path $Y \leftarrow X \rightarrow R_Y$. However, if one conditions on X, the second path becomes blocked, and if one conditions on A, the first path becomes blocked and Y and R_Y are now d-separated, implying conditional stochastic independence $R_Y \perp\!\!\!\perp Y|(A, X)$, and therefore MAR holds. We see here that using only X as a conditioning variable leaves Y and R_Y d-connected and thus MAR is violated. Only if variable A (even though it might not be of substantive interest) is also used to condition, Y and R_Y become d-separated and MAR holds. Expressed generally, whenever the set of missingness indicators R and the sets of partially observed and unobserved variables in the graph can be d-separated given the set of observed variables, MAR holds.

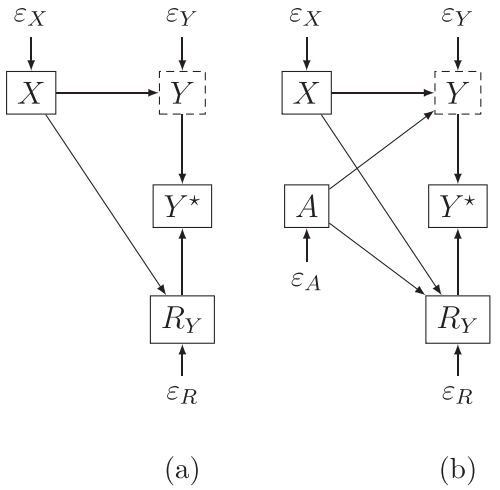

FIGURE 2 A simple missing at random model (a) without auxiliary variables and (b) with auxiliary variables.

[1]In the context of missing data, conditioning on a variable can refer to using this variable in the FIML estimation or alternatively as a predictor in an MI framework.

Graphical Display of NMAR

Finally, we consider graphs that are NMAR. The first example is given in Figure 3a, in which Y and R_Y are directly connected by a path. Y and R_Y are said to be d-connected through the direct path $Y \rightarrow R_Y$. Two adjacent, connected variables in a graph can never be d-separated. Hence, no conditional stochastic independence can arise, and NMAR is present.

Another situation that is also NMAR emerges whenever there is an omitted variable that has an effect on both the variable with missing data and the missingness on this variable. This omitted variable can be displayed as a latent,

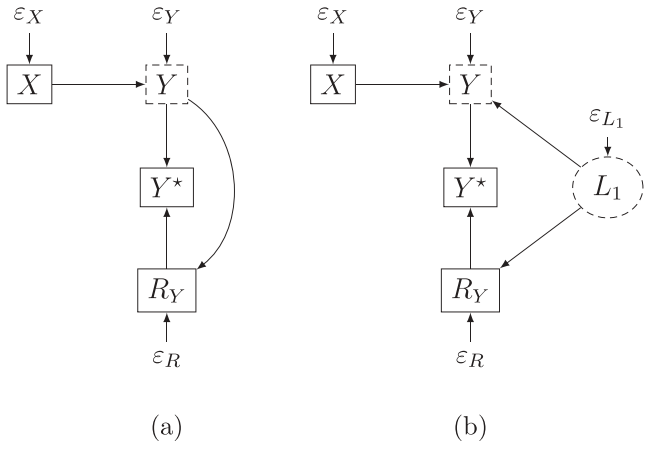

FIGURE 3 A simple not missing at random model with direct path between missingness and (a) variable with missing data and (b) unobserved variable related to both Y and R_Y.

unobserved variable in the graph, or simply as correlated disturbance terms. Figure 3b displays such a situation in which an omitted variable influences both Y and R_Y. Here, Y and R_Y are d-connected via the path $Y \leftarrow L_1 \rightarrow R_Y$. The variable L_1 in the graph should not be confused with a modeled, latent variable in a structural equation model, but rather is a simple depiction of an unobserved variable. The path between Y and R_Y cannot be blocked via conditioning, because no observed variables reside in the middle of the path. Again, no stochastic conditional independence can be achieved through conditioning and NMAR holds.

In the previous sections we showed how the classic missingness mechanisms can be expressed via graphs that encode conditional independencies and applied the graph-theoretic concept of d-separation. In summary, when a variable Y and its associated missingness indicator R_Y cannot be d-separated using any set of observed variables, NMAR holds. If Y and R_Y can be d-separated using any set of other observed variables then MAR holds, and parameters related to Y (e.g., means) can be consistently estimated, when using methods that assume MAR (FIML, MI). A special case arises when Y and R_Y are d-separated given no other variables (unconditionally independent), which maps onto the classic MCAR condition.

Differences Between *m*-graphs and Other Graphical Displays

After we have introduced *m*-graphs, it is informative to highlight some important differences from other graphical displays that are being used in the literature. Some readers might be familiar with graphs that have been used in the context of missing data; for example, in the seminal paper by Schafer (1999) or the widely used text by Enders (2010). A key difference is that in *m*-graphs, directed arrows specify causal relations among variables and hence permit us to infer conditional independencies. Other texts use either bidirected arrows or undirected arrows interchangeably. Enders (2010) described the relations in his graphs as "generic statistical associations," and specifically did not distinguish between two variables simply being correlated due to unobserved variables (spurious correlation), or two variables having a causal relationship with each other (e.g., A causing B).

We illustrate now why it is important to distinguish causal relationships from generic statistical associations to recover consistent parameter estimates from variables with missing data. Consider a simple example in which a single variable B has missing data, indicated by R_B and a variable A, fully observed is at the disposal of the researcher. This example mirrors one that is also used in Enders (2010) to describe the MAR mechanism. In Figure 4a, dashed lines are shown to display generic statistical associations. A generic statistical association might emerge because of direct effects as displayed in Figure 4b, but they could also emerge due to spurious associations due to unobserved variables L_1 and L_2

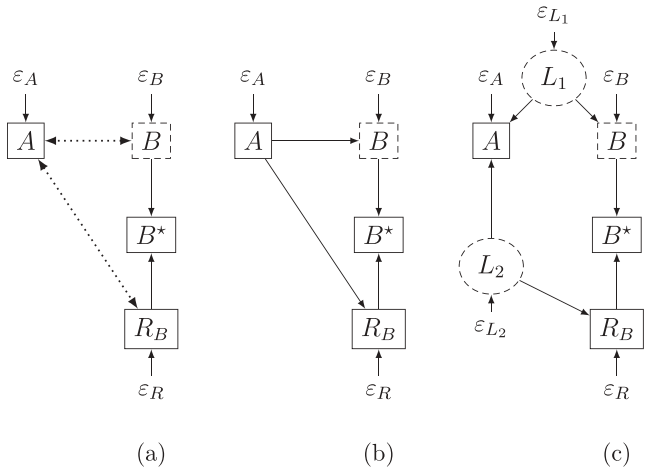

FIGURE 4 Differences in graphs comparing (a) generic statistical associations, and directed relationships in (b) and (c). Disturbance terms are omitted.

in Figure 4c. Both patterns in Figure 4b and 4c have the same generic statistical associations (i.e., correlational patterns), yet they have different underlying structures. Hence they require different treatments for missing data. If we were to rely solely on correlational patterns, Figure 4b and 4c would be treated the same and in both cases inclusion of A as an auxiliary variable would be recommended (because A is correlated with B and missingness on B). Further, it would be expected that inclusion of A as an auxiliary would eliminate bias in B. However, when applying graphical criteria, the two situations in Figure 4b and 4c require different treatments. In Figure 4b, we conclude that B and R_B can only be d-separated if one uses A as a conditioning variable. Ignoring A will yield biased results and inclusion of it eliminates bias. The exact opposite conclusion is yielded by Figure 4c. Here, B and R_B are unconditionally independent from each other, because A is a collider and no open path exists between B and R_B. Because conditioning on a collider opens a path (Pearl, 2010), inclusion of A as a conditioning auxiliary variable will induce dependencies between B and R_B that bias estimates of B.

To further convince readers that this pattern of bias reduction and induction holds, we simulated data based on models in Figure 4b and 4c, and estimated the mean of B using either listwise deletion or a FIML model that included A as an auxiliary variable. In this illustration all variables were completely standardized (true mean of 0 and unit variance) and all path coefficients were set to .7. The missing data rate on B was set to 30%. Sample size was fixed at 100,000 to minimize sampling error. All analyses were performed in R (R Development Core Team, 2011) and lavaan (Rosseel, 2012). Results of this data illustration are given in Table 1.

It can be easily seen in Table 1 that in the situation in which A is a direct cause of B and missingness, listwise deletion is heavily biased and inclusion of A as an auxiliary

TABLE 1
Illustrative Data Example: Means and Standard Deviation of Variable B

	Listwise	*FIML With A*
Model 4b	−.30 (.93)	−.01 (1.00)
Model 4c	.00 (1.00)	.18 (1.05)

Note. FIML = full-information maximum likelihood.

variable completely nullifies this bias. On the other hand, if A is only spuriously correlated due to unobserved third variables, listwise deletion is completely unbiased in this example, whereas inclusion of A induces strong biases. Interestingly, in both situations A is strongly correlated to both B and its missingness (R_B) and according to conventional wisdom should be included as an auxiliary variable. The graphs that only rely on generic statistical associations are unable to differentiate these two cases, even though they clearly have very different implications. The *m*-graphs that we present in this article immediately tell us that this type of bias behavior will emerge because in the graph in Figure 4b, A is needed to d-separate B from R_B, whereas in the graph in Figure 4c, A will induce a dependency in previously d-separated variables B and R_B. For simulation studies that further explore this phenomenon of bias-inducing auxiliary variables in more detail, the reader is referred to Thoemmes and Rose (2014).

CONSTRUCTION OF AN *M*-GRAPH

So far we have only discussed how missingness mechanisms would be displayed in a graph and how this could be a pedagogical tool to think about missing data. However, *m*-graphs could also be used in an applied context. Researchers can use *m*-graphs to graphically display their theoretical knowledge and assumptions about relationships among variables and missingness. In practice, one could start by constructing a graph of substantive variables of interest along with their missingness indicators, and the observed portion of the variables. Assumed causal relationships among substantive variables themselves and substantive variables and missingness could then be added to the graph. In the next step, the applied researcher can augment the graph with potentially observable variables that are not of substantive interest, but might be related to analysis variables or missingness (auxiliary variables). Finally, unobservable variables can be added to the graph along with their assumed relations between substantive variables and missingness. The resulting graph would then represent the best state of knowledge of the applied researcher. Differently said, this graph is a representation of all theoretical considerations and assumptions about relationships among variables and causes of missing data.

This is clearly no easy task for an applied researcher, but there are some advantages to doing it. First, the assumptions that the researcher is making are clearly laid out in the graphical model. If, for example, no unmeasured confounders are assumed between two particular variables in a graph (or between a missingness indicator and a variable), their disturbance terms will not be connected by a bidirected arrow. This assumption can immediately be seen (and challenged) by other researchers, something that is arguably harder if researchers only appeal to broader concepts, for example, claiming that data are MAR, without providing much evidence on why this should, in fact, hold. The graph thus replaces a general statement, usually provided as a narrative, that MCAR, MAR, or NMAR might hold, with specific assumptions about the relationships among variables. Note, for example, that the generic claim that MAR holds could be translated in a graphical model in which all disturbance terms of missingness indicators and variables with missing data are uncorrelated, and all variables that induce dependencies between a variable Y and its mechanism R_Y are in fact collected. Given that MAR is frequently invoked by applied researchers, it is actually informative to think about what strict assumptions are imposed on a graph that implies MAR. Second, if the researcher and critical peers agree about the assumptions in the graph, then there should also be agreement that a particular missingness mechanism holds and that therefore certain parameters in the model can be consistently estimated. Of course, disagreement about the assumptions will also lead to disagreement as to which mechanism actually holds. Either way it should foster a more critical discussion about missingness mechanisms among researchers, and encourage researchers and recipients of research to think about the necessary assumptions that need to be made to claim a particular mechanism.

Note that it is not necessary to have complete knowledge about all relationships between variables among each other and missingness. A lack of knowledge can always be represented by allowing directed and bidirected arrows to exist between two variables, representing an effect that is also confounded by unobserved variables. Likewise, an unwillingness to rule out certain relationships can be expressed by adding additional directed or bidirected arrows to the graph and to potentially connect many variables. This usually results in situations in which no (conditional) independence holds anymore and thus NMAR is present. We would also like to emphasize that graph structure is not estimated from data, but is based on qualitative assumptions based on theory that the applied researcher is able and willing to make. That means that we are not claiming that we can verify particular missingness mechanisms from data alone, but rather that certain assumptions, encoded in a graph, might imply a particular mechanism. Once assumptions have been postulated, it is possible to determine whether these sets of assumptions about missingness map onto any of the three mechanisms. Further it is possible to determine whether any particular

parameter estimate (e.g., mean, regression coefficient) can be consistently estimated from the observed data, given the spelled-out assumptions encoded in the graph.

HOW TO DETERMINE WHETHER PARAMETERS CAN BE RECOVERED

After a researcher has settled on a particular structure in an *m*-graph that represents his or her best theoretical knowledge along with all assumptions that the researcher is willing (and able) to make, it is possible to query the graph to determine if certain parameters can be recovered even in the presence of missing data. A researcher might ask whether a particular mean, or a regression coefficient, could be consistently estimated, and if so, which auxiliary variables should be used in the estimation. Mohan et al. (2013) provided a set of necessary and sufficient conditions under which parameter estimates can be recovered[2] in the presence of missing data. We do not present every condition here, but focus our attention on cases that we deem especially useful for applied researchers. In particular, we discuss conditions under which parameter estimates of means and regression coefficients can be recovered.

Graphical Criteria for Identification of Means

Mohan et al. (2013) provided criteria and associated proofs that can be used to determine the recoverability of means (or other univariate parameter estimates) in the presence of missing data, given the assumed structure of a particular *m*-graph. If it is possible to find a set of fully observed variables W that d-separate Y from R_Y, then univariate parameters of Y (e.g., the mean) can be recovered. W might include variables that are of substantive interest or are auxiliary variables. We could express this in a formula, which is quite similar to the expression of MAR, save for the fact that W might be a select subset of observed variables, as:

$$Y \perp\!\!\!\perp R_Y | W. \tag{6}$$

Note that this definition immediately precludes recoverability of means whenever Y has a direct effect on R_Y, or if there is an unobserved (latent) variable that affects both Y and R_Y (e.g., displayed in the form of a bidirected arrow that connects disturbance terms). Both of these situations are NMAR. If any variable in W is partially observed, the simple criterion in Equation 6 also does not generally hold anymore.[3] Note that the criterion in Equation 6 is a sufficient condition for recoverability of means in Y, meaning that there are cases in which it does not hold, but yet it would be possible to recover means and other parameters.

We briefly note that there is an alternative graphical criterion to check whether means of Y are recoverable. It considers all R indicators of all variables in an *m*-graph at once (Mohan et al., 2013). The joint distribution (and therefore means of variables) can be recovered if (a) there are no arrows connecting the R variables, (b) there are no unobserved (latent) variables that have direct effects on any R, and (c) no variable with missing data has a direct effect on its own associated missingness indicator.

Graphical Criteria for Identification of Regression Coefficients

If the estimation of a regression coefficient of Y on X is of interest, different graphical criteria need to be applied. Now Y needs to be d-separated not only from R_Y, but also from R_X, conditional on X.

$$Y \perp\!\!\!\perp \{R_Y, R_X\} | X. \tag{7}$$

Importantly, when applying this criterion, X, the predictor variables, could have missing data themselves. The recoverabilty of the mean of X itself is not of concern when probing the recoverability of the regression coefficient.

If this condition does not hold, we might attempt to find a set of covariates W in which this independence holds. However, we also need to ensure that Y is conditionally independent of R_W, as expressed here:

$$Y \perp\!\!\!\perp \{R_Y, R_X, R_W\} | X, W. \tag{8}$$

In the case where W is fully observed, Equation 8 is sufficient for recovery of the regression of Y on X. Otherwise, additional conditions are required, namely that we also need to show that the regression coefficients using X as predictors of W are recoverable as well. This means that we need to recursively apply Equations 7 and 8 to the regression of W on X.

For readers who are unfamiliar with graphical models and d-separation, some of these criteria might appear daunting. Next we provide a worked-out example that shows the application of these criteria. In addition, we provide computer code for the program DAGitty (Textor et al., 2011), that will automate queries of conditional independence (d-separation).

ILLUSTRATIVE EXAMPLE

In this illustrative example, we want to highlight how an applied researcher could use graphical methods to think

[2] "Recovered" is defined as the ability to asymptotically estimate a consistent parameter value in the presence of missing data.

[3] More complex criteria exist that allow the use of partially observed variables in the conditioning set W. Those rely on a concept called ordered factorization and can be found in Mohan et al. (2013).

about missing data. Our example is necessarily quite simple, but the general underlying process could be applied to larger problems. We consider a situation in which a single dependent variable Y is predicted by a total of two independent variables, X_1, and X_2. All of these variables have missing values. The substantive interest is on univariate measures of the Xs and Y (e.g., means), and on the regression coefficients in which the Xs are used as predictors of Y. In addition, there are a total of two auxiliary variables, A_1 and A_2, at the disposal of the researcher. Those are not of substantive interest, but have some assumed relationships with the substantive variables and with missingness. For simplicity, these auxiliary variables are all completely observed and have no missing data. Based on theoretical considerations, we construct the m-graph in Figure 5. This graph represents our theory and our assumptions about the causal processes that govern relationships among variables and missingness. Other researchers might challenge these assumptions encoded in the graph, but as soon as there is agreement about assumptions, it is well defined which effects can be consistently estimated, using criteria formulated earlier.

We begin by asking whether means of the two predictor variables X_1 and X_2 can be recovered. From Figure 5, we can see that X_1 and R_{X_1} are d-connected via a direct effect, indicating that missingness on X_1 depends directly on X_1 itself. It is therefore impossible to d-separate these two variables using any other observed variables as dictated by Equation 6, NMAR holds, and we cannot recover an unbiased mean of X_1.

Considering estimates of X_2, we observe multiple paths that can be traced from X_2 to R_{X_2}; however, in this example, all of them traverse colliders and are already closed; for example, $X_2 \rightarrow A_2 \leftarrow \varepsilon_{A_2} \leftrightarrow \varepsilon_{R_{X_2}} \rightarrow R_{X_2}$. Application of d-separation reveals that X_2 and R_{X_2} are unconditionally independent of each other. That means that both listwise deletion and FIML or MI would yield consistent results of means of X_2, because an MCAR situation is present. Interestingly, A_2 would not be needed for consistent estimates, even though it is correlated with both X_2 and R_{X_2}. In fact, using A_2 is expected to induce biases in the estimate of X_2 because it induces a dependency between X_2 and R_{X_2}.

To determine the recoverability of the mean of Y we note that $Y \perp\!\!\!\perp R_Y | X_1, X_2, A_1$ holds. However, both X_1, and X_2 are only partially observed and therefore the simple recovery criterion is not fulfilled. The criterion that considers all variables to check recoverability of the mean of Y is also not fulfilled because the missingness term of X_1 has a direct effect from X_1 itself. We thus fail to meet sufficient conditions to recover the mean of Y and conjecture that the mean of Y remains biased. In summary, we would be unable to recover the mean of X_1 and Y, but would be able to get consistent mean estimates of X_2.

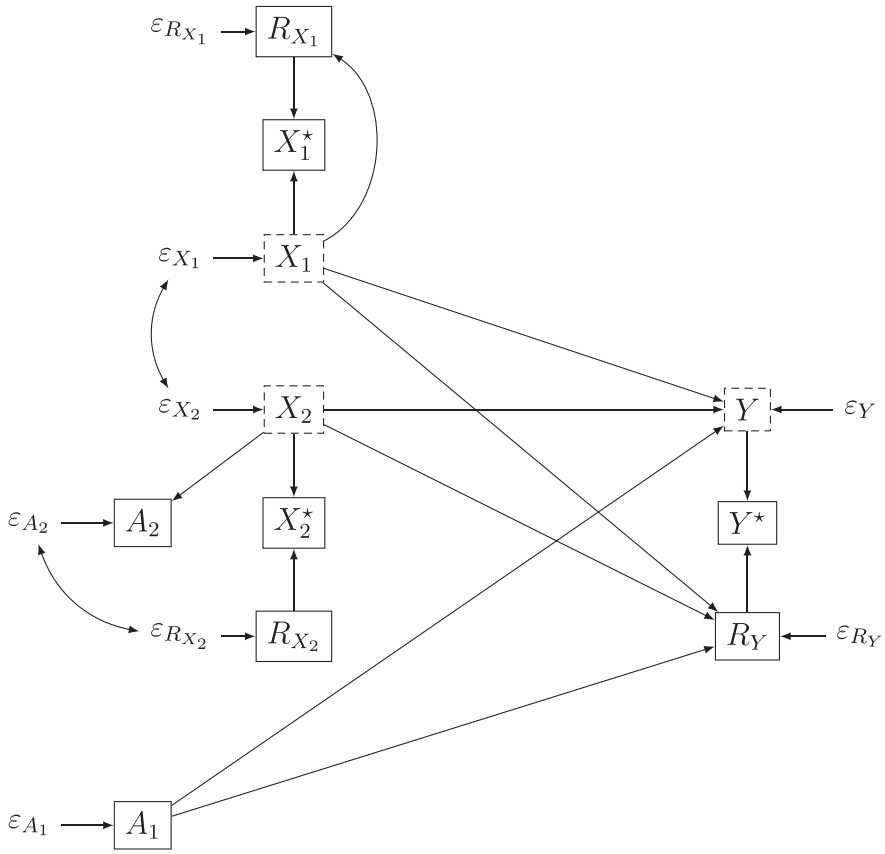

FIGURE 5 Illustrative example of regression problem with missing data.

Finally, we are interested in the recoverability of the regression coefficients from the Xs to the outcome Y. We first check whether $Y \perp\!\!\!\perp \{R_Y, R_{X1}, R_{X_2}\}|\{X_1, X_2\}$ holds. This is not the case, as A_1 still d-connects Y and R_Y. However, we also observe that $Y \perp\!\!\!\perp \{R_Y, R_{X_1}, R_{X_2}\}|\{X_1, X_2, A_1\}$ holds. Because we have introduced a new variable A_1 to our conditioning set, we also need to check whether the regression of A_1 on the Xs is also recoverable. We can do so by using Equations 7 and 8 again, this time testing whether $A_1 \perp\!\!\!\perp \{R_{A_1}, R_{X_1}, R_{X_2}\}|\{X_1, X_2\}$ holds. Here, we first observe that A_1 is fully observed (meaning that R_{A_1} will be a constant that is by definition unrelated to other variables), and that therefore this check can be reduced to $A_1 \perp\!\!\!\perp \{R_{X_1}, R_{X_2}\}|\{X_1, X_2\}$. This in turn holds, and we thus now know that the regression of Y on X_1, and X_2 can be recovered when using A_1 as an auxiliary variable. Interestingly, the regression coefficients can be recovered, even though the means of both X_1 and Y cannot be recovered due to NMAR situations. Finally, we note that an approach that does not consider an *m*-graph, but only checks correlations among partially observed variables, missingness indicators, and auxiliary variables, would come to the conclusion that both A_1 and A_2 should be used as auxiliary variables, because they are correlated with some variables and missingness indicators. The interested reader can find many more illustrative examples of *m*-graphs in Mohan et al. (2013).

Numerical Demonstration

We simulated data based on the model in Figure 5. For simplicity, we completely standardized all variables, and fixed every unidirected path to .3 on a standardized metric, and fixed the bidirected arrows (correlations) to .5. For effects on missingness indicators R, we modeled relationships between variables and an underlying continuous variable that represents the latent cause of missingness. All variables had a relatively large amount of 50% missing data and we likewise used a large sample size of 10, 000 to have relatively precise estimates of means and regression coefficients. The complete data were generated using TETRAD (Scheines et al., 1998). Missing data were imposed in R (R Development Core Team, 2011) and subsequent analyses were also performed in R, using the package mice (Van Buuren, Boshuizen, & Knook, 1999) to perform multiple imputation.

We estimated means of all variables and the simple regression model, regressing Y on both Xs, using either listwise deletion, an imputation approach based on the *m*-graph that only considers A_1 as an auxiliary variable (called *m*-graph imputation in Table 2), or a model-blind imputation approach that uses both A_1 and A_2 as auxiliary variables (called full imputation in Table 2). Table 2 presents results of these analyses.

In line with our expectations from applying graphical criteria, we observe bias in the means of X_1 and Y under listwise deletion, but no bias in the means of X_2. Regression coefficients are slightly biased under listwise deletion (note that bias in regression coefficients tends to be small if the logit of the probability to be missing is linearly related to causes of missingness, Collins, Schafer, & Kam, 2001).

TABLE 2
Parameter Estimates for Numerical Demonstration Under Different Missing Data Treatment

Parameter Estimate	*Complete*	*Listwise*	*m-graph Imputation*	*Full Imputation*
Mean X_1	.01	.25	.22	.21
Mean X_2	.00	.00	.02	−.11
Mean Y	.01	.30	.10	.10
Regression coefficient X_1	.30	.29	.30	.31
Regression coefficient X_2	.30	.28	.28	.26

Using MI based on a model-blind approach, we observe that the mean of X_1 is still biased, as we expected from the graphical criterion. The bias in the mean of Y has been substantially reduced, but Y is still not entirely free of bias. However, substantial bias in the mean of X_2 has been introduced, due to using A_2 as an auxiliary variable. We further observe small biases in regression coefficients.

Using MI based on the *m*-graph, we still see bias in the means of X_1 and (some residual bias) in Y, but A_2 remains bias-free, as was expected. Bias in regression coefficients remains very small for the coefficient of X_2.[4]

DISCUSSION

Our goal was to describe the classic missingness mechanisms using graphical models and argue that graphs can be a useful tool to think about missing data problems. We have demonstrated that the mechanisms can be expressed with graphs and importantly that the conditional independencies that are formally represented in formulaic expressions of equality of (conditional) probabilities can also be formally expressed in a graph using the d-separation criterion. Besides its use as a purely pedagogical tool, it might be useful for applied researchers to consider graphical models when they are interested in thinking about the structural relationships between variables and missingness indicators. As we have demonstrated, there are instances in which a variable could pose as a helpful auxiliary variable, but is in fact harmful. Although it is impossible to distinguish these variables using statistical criteria, the graphs allow for theoretical considerations to be expressed that might help to discover such a variable. Of course, at the same time, if assumptions that went into the graph are incorrect, then variables might be incorrectly classified as being harmful. The important message is that it is always better to use theoretical knowledge to think about the particular structure of an auxiliary variable, as opposed

[4]To rule out that these results were influenced by a particular random sampling, we replicated these results with an additional data set with a sample size of 100,000 and observed nearly identical results.

to always assuming that it never falls in the category of bias-inducing auxiliary variables.

In practice, an applied researcher would have to express his or her assumptions about structural relationships as precisely as possible. An often levied criticism is that such knowledge is not generally available, but in fact, as we have described earlier, it is not a prerequisite to have all available knowledge about all possible structural relationships between variables, missingness, and other auxiliary variables. Ignorance about specific relationships can be expressed by including additional directed or bidirected paths in the model. Correlated disturbance terms can always be included to express that unobserved causes of variables might exert influences on observed variables. Including additional paths implies that one is imposing fewer restrictions, at the expense of eventually having such a saturated model that any restrictions needed to fulfill MCAR or MAR disappear. This shows again that to claim MAR, certain independence assumptions between missingness and other observed and unobserved variables need to be made. A graphical display might help the applied researcher to judge whether these assumptions are plausible and whether it is plausible to claim MAR. Once the researcher has identified an assumed structural model, it is possible to determine the missingness mechanism and recoverability by examining the graphical criteria that we have described.

In summary, missingness mechanisms can be expressed in graphical models that are able to encode assumptions about conditional independence. These assumptions are crucial in the definitions of MCAR, MAR, and NMAR. We believe that graphs offer an alternative way to describe and explain the missingness mechanisms. We further believe that this approach might be accessible to a wide range of researchers and potentially easier to grasp for students and practitioners alike. We also tried to motivate use of such graphs in thinking about missing data using an illustrative example in which some background knowledge and assumptions about missingness, observed variables, and unobserved variables were available. We hope that these tools help students and applied researchers to think clearly about missing data mechanisms in general and in particular about the mechanisms that might be present in their own data.

ACKNOWLEDGMENTS

The authors would like to thank the participants of the colloquium of the Methodology Center at Pennsylvania State University and Judea Pearl for helpful feedback.

REFERENCES

Allison, P. D. (2001). *Missing data (quantitative applications in the social sciences)*. Thousand Oaks, CA: Sage.

Breitling, L. (2010). dagr: A suite of R functions for directed acyclic graphs. *Epidemiology*, *21*, 586–587.

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

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

Dawid, A. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B (Methodological)*, 1–31.

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

Enders, C. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*, 1.

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

Glymour, M. (2006). Using causal diagrams to understand common problems in social epidemiology. In J. M. Oakes & J. S. Kaufman (Eds.), *Methods in social epidemiology*, (pp. 370–392). San Francisco, CA: Jossey-Bass.

Graham, J., Taylor, B., Olchowski, A., & Cumsille, P. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*, 323–343.

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

Knüppel, S., & Stang, A. (2010). DAG program: Identifying minimal sufficient adjustment sets. *Epidemiology*, *21*, 159.

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

Martel García, F. (2013). Definition and diagnosis of problematic attrition in randomized controlled experiments. Retrieved from http://ssrn.com/abstract=2302735.

Mohan, K., & Pearl, J. (2014). On the testability of models with missing data. In S. Kaski, & J. Corander (Eds.), *JMLR Workshop and Conference Proceedings Volume 33: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 643–650). Retrieved from http://jmlr.csail.mit.edu/proceedings/papers/v33/

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

Muthén, B., Asparouhov, T., Hunter, A. M., & Leuchter, A. F. (2011). Growth modeling with nonignorable dropout: Alternative analyses of the star* d antidepressant trial. *Psychological Methods*, *16*, 17–33.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco: Morgan Kaufmann.

Pearl, J. (2009). *Causality: Models, reasoning, and inference, 2nd edition*. New York, NY: Cambridge University Press.

Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology*, *40*, 75–149.

Pearl, J., & Mohan, K. (2014). *Recoverability and testability of missing data: Introduction and summary of results*. (Technical Report R-417). Los Angeles, CA: UCLA Cognitive Systems Laboratory.

Potthoff, R. F., Tudor, G. E., Pieper, K. S., & Hasselblad, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research*, *15*, 213–234.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: Author. Retrieved from http://www.R-project.org

Raykov, T. (2011). On testability of missing data mechanisms in incomplete data sets. *Structural Equation Modeling*, *18*, 419–429.

Raykov, T., Lichtenberg, P. A., & Paulson, D. (2012). Examining the missing completely at random mechanism in incomplete data sets: A multiple testing approach. *Structural Equation Modeling*, *19*, 399–408.

Rosseel, Y. (2012). *lavaan: Latent variable analysis*. Retrieved from http://CRAN.R-project.org/package=lavaan

Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.
Schafer, J. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*, 3–15.
Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.
Scheines, R., Spirtes, P., Glymour, C., Meek, C., & Richardson, T. (1998). The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, *33*, 65–117.
Textor, J., Hardt, J., & Knüppel, S. (2011). DAGitty: A graphical tool for analyzing causal diagrams. *Epidemiology*, *22*, 745.
Thoemmes, F., & Rose, N. (2014). A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate Behavioral Research*, *49*(5), 443–459.
Van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*, 681–694.

APPENDIX A
A SMALL D-SEPARATION EXAMPLE

The DAG in Figure A.1 shows a small example in which it is of interest to determine whether X and Y are d-separated.

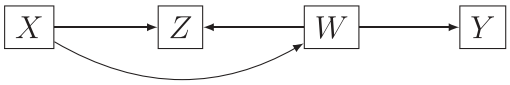

FIGURE A.1 A simple d-separation example. Disturbance terms are omitted.

There are two paths that connect X and Y. The first is $X \rightarrow Z \leftarrow W \leftarrow Y$. Because this path contains a collider, it is closed. The second path is $X \rightarrow W \rightarrow Y$. This path does not contain any collider and is therefore open. Thus, X and Y are d-connected. We can further explore whether X and Y are d-separated if we condition on any combinations of the remaining variables Z and W. If we condition on only Z we open a previously closed path, and thus X and Y are d-connected, given Z. If on the other hand, we only condition on W, all previously open paths are blocked, and therefore X and Y are d-separated, $(X \perp\!\!\!\perp Y)|W$. Finally, conditioning on both Z and W also blocks all open paths. The path that was opened by conditioning on the collider Z is again blocked, because W resides on this path as well. The concept of d-separation can also be applied to *m* graphs in which certain nodes represent missingness information.

APPENDIX B
COMPUTER CODE TO GENERATE THE *M*-GRAPH FOR THE ILLUSTRATIVE EXAMPLE

This computer code can be pasted directly in the "Model text data" in the DAGitty program (http://dagitty.net/) (Textor et al., 2011). The empty line in the following code is critical and must be retained when copying the text.

```
a1 1 @0.420,0.914
a2 1 @0.048,0.682
rx1 0 @0.115,0.364
rx2 0 @0.154,0.830
ry 0 @0.750,0.726
u1 1 @0.052,0.873
u2 1 @0.030,0.428
x1 1 @0.117,0.121
x1star 1 @0.172,0.241
x2 1 @0.153,0.541
x2star 1 @0.143,0.662
y E @0.744,0.508
ystar 1 @0.749,0.614
a1 ry y
rx1 x1star
rx2 x2star
ry yobs
u1 a2 rx2
u2 x1 x2
x1 y x1star ry rx1
x2 x2star y ry a2
y ystar
```

Testing Conditional Independence

DAGitty is a program that checks d-separation in graphs (along with other features, e.g., finding variable sets that block noncausal paths between two variables). *m* graphs can be easily created in DAGitty (see preceding code) by defining missingness indicators along with variables of substantive interest. DAGitty does not support drawing of bidirected arrows, however they can be easily represented as unobserved variables that have direct effects on both variables that should be connected with a bidirected arrow. We make use of these unobserved variables in the computer code and create U_1 to U_2, corresponding to each of the two bidirected arrows in Figure 5.

After a graph has been defined, DAGitty will automatically produce all (conditional) independence statements and report those on the right side of the graph. However, this list of d-separation statements can get very long. An alternative is to highlight two (or more) nodes in the graph and check their independence visually. To do this, we designate one node the "exposure" and the other node the "outcome." In the context of an *m* graph, this distinction is not meaningful, so it does not matter which node is labeled the exposure or outcome. For this example, we always denote the actual variable as the exposure and the missingness indicator as the outcome. Nodes are highlighted by hovering over them with the mouse cursor and then pressing "e" for exposure, and "o" for outcome. It is possible to designate more than one node as the exposure or outcome, which is helpful in checking joint independencies. Once two (or more) nodes have been selected, DAGitty presents all active paths that connect the two variables in either red or green (again, this distinction is not important in *m* graphs). These highlighted paths induce a dependency between two nodes. If there are no red or green paths, then the two nodes are d-separated from each other. If on the other hand there are highlighted paths, the applied researcher could check whether these paths can be blocked by some other variables that lie on highlighted sequences of paths. To use a variable as a conditioning variable (equivalent to an auxiliary variable in missing data estimation), the user has to hover the mouse cursor over this variable and press "a". The display (including highlighted paths) updates automatically and the user can check if after conditioning any highlighted paths remain, or whether two nodes are d-separated.

In our example, we can begin by labeling X_1 as the exposure and R_{X_1} as the outcome. We observe that the direct path between these two variables is highlighted, indicating that this path induces a dependency between the two nodes. Because we cannot condition on any variable in this direct path, we conclude that X_1 and R_{X_1} are d-connected and that the mean of X_1 cannot be recovered. We proceed by labeling X_2 as the exposure and R_{X_2} as the outcome and observe that no path is highlighted. We can conclude that X_2 and R_{X_2} are d-separated and thus the mean of X_2 is recoverable. Repeating this for Y reveals that the unobserved X_1, X_2, and the observed A_1 are needed to d-separate Y and R_Y. Because the conditioning set contains partially observed variables X_1 and X_2, the criterion is not fulfilled and we conjecture that the mean of Y will remain biased. Finally, to check recoverability of the regression of Y on both Xs, we highlight Y as the exposure, and highlight R_Y, R_{X_1}, and R_{X_2} as outcomes. In addition, we highlight X_1 and X_2 as adjustment variables. This allows us to check the compound d-separation $Y \perp\!\!\!\perp \{R_Y, R_{X_1}, R_{X_2}\}|\{X_1, X_2\}$, defined earlier. We see that conditioning on A_1 eliminates all highlighted paths. Because A_1 is fully observed in this example, we do not need to check any further d-separation criteria and can conclude that the regression coefficients are recoverable.