Graphical Causal Models for Survey Inference

Julian Schuessler*

Department of Political Science

Aarhus University

julians@ps.au.dk

Peter Selb

Department of Politics and Public Administration

University of Konstanz

peter.selb@uni.kn

April 19, 2021

*We are grateful to Eric Plutzer, Frauke Kreuter, Christopher Winship, Ulrich Kohler, Sebastian Lang, as well as the participants of the panel on Survey Methods at the EPOP Conference in Glasgow, September 2019, and the DGS Workshop on Graphical Causal Models in Potsdam, March 2020, for helpful comments. Financial support by the German Academic Scholarship Foundation, the Graduate School of Decision Sciences, and the Cluster of Excellence "The Politics of Ineuqality" is appreciated. We thank Fred Hockney for language editing and proofreading.

Abstract

Directed acyclic graphs (DAGs) are now a popular tool to inform causal inferences. Departing from their informal use in the survey research literature, we discuss how DAGs can also be used to encode theoretical assumptions about nonprobability samples and survey nonresponse and to determine whether population quantities, including conditional distributions and regressions, can be identified from a sample. We describe sources of bias and assumptions for eliminating it in selection scenarios familiar from the missing data literature. We then introduce and analyze graphical representations of multiple selection stages in the data collection process, and highlight the strong assumptions implicit in using only design weights. Furthermore, we show that the common practice of selecting adjustment variables based on empirical correlations is ill-justified and that nonresponse weighting when the interest is in causal inference may come at severe costs. Finally, we identify further areas for survey methodology research that can benefit from advances in causal graph theory.

1 Introduction

Surveys figure among the most prominent data collection tools for social research. Regardless of whether one is dealing with an election poll, health survey, or census, the general purpose of a survey is to provide information about the distribution of some individual attribute (e.g., an attitude, behavior, or social characteristic) in a population, usually based on self-reports in a sample of individuals selected from the population. The inferences involved – from self-reports to underlying attributes (i.e., measurement) and from samples to target populations (i.e., generalization) – are prone to various errors which may arise in the design, collection, and analysis of survey data. While the Total Survey Error (TSE) framework has emerged as the dominant conceptual foundation for studying the sources of error and ways of assessing and improving survey data quality (e.g., Groves et al., 2009), some influential work in survey methodology has used graphical models to illustrate errors (e.g., Groves,

2006; Groves and Peytcheva, 2008; Kreuter and Olson, 2011; Schafer and Graham, 2002; Olson, 2019). However, their usage has largely remained informal and has mostly focused on explaining problems with nonresponse, but not their solutions.

At the same time, research using directed acyclic graphs has greatly enhanced not only our understanding of causal inference (Pearl, 2009; Morgan and Winship, 2015) but also of sample selection bias and other missing data problems (Bareinboim, Tian and Pearl, 2014; Elwert and Winship, 2014; Thoemmes and Mohan, 2015; Mohan, Thoemmes and Pearl, 2018). In this article, we bring together these two disparate strands of research. We distill important insights from the graphical literature and apply them to prototypical problems that occur in research that uses surveys with non-response or based on non-probability samples. While prior research in sociology (e.g., Elwert and Winship, 2014) has mostly focused on uncovering problems with nonresponse when the interest is in causal inference, we here also discuss descriptive population inferences, and we show how to understand *solutions* to common problems graphically.

Furthermore, we use graphical and counterfactual models to analyze more specific practices in survey research. Based on our analysis, we make concrete suggestions to improve how researchers choose and use adjustment variables when data are missing. Thereby, we not only underline the value of causal graphs as an intuitive and powerful tool for survey researchers to communicate their assumptions, be it for inferences on descriptive or causal parameters, but also highlight their value in uncovering non-trivial insights for improving more complicated statistical practices.

After elaborating on the tight connection between graphs and statistical dependencies, we show how classic assumptions made in the adjustment for survey nonresponse ("missing at random", etc.) can be intuitively yet rigorously formalized using graphs. Inter alia, we emphasize the importance of using causal instead of correlational language and discuss how "collider bias" can explain findings of vote validation studies. Furthermore, we provide an explicit explanation and justification for common "inverse-probability-weighted" estimators,

going beyond the otherwise encompassing analysis in Wooldridge (2007).

Next, we use the graphical approach to contribute to three areas in survey research: First, we analyze multiple selection stages and the exclusive use of design weights. Here, we closely follow the TSE framework and go beyond existing analyses of sample selection using DAGs (e.g., Bareinboim, Tian and Pearl, 2014; Daniel et al., 2012; Moreno-Betancur et al., 2018; Thoemmes and Mohan, 2015). Specifically, we discuss the implicit strong assumptions behind the widespread practice of using "design weights" (and only those) when analyzing survey data (e.g., the General Social Survey).

Second, we critically investigate correlation-based strategies to choose adjustment variables. We show that requirements formulated in the literature (Kreuter et al., 2010; Peytchev, Presser and Zhang, 2018; Sakshaug and Antoni, 2019) for valid adjustment variables to be correlated with response indicators and variables of interest are ill-justified. Theoretical considerations—formulated in graphical language—trump correlational information.

Third, we discuss non-response weighting when the interest is in causal effects. Here we argue that for principled reasons and unlike in the case of estimation of means or associations, "imperfect" adjustment is generally ill-advised. We argue that the sample average causal effect in itself may be of scientific interest, while imperfectly weighted estimators of population average causal effects are consistent for neither the sample nor the population average effect.

In sum, we contribute to a development Groves and Lyberg (2010, 866) urged in their critical appraisal of the state of survey methodology a decade ago:

> Missing in the history of the total survey error formulation is the partnership between scientists who study the causes of the behavior producing the statistical error and the statistical models used to describe them.

We conclude our paper by outlining future areas of research where causal graphs can be used to better understand the assumptions of survey methods or to develop new methods:

regression adjustment, sample selection models based on instrumental variables, "transporting" causal effect estimates using observational data, sensitivity analysis, measurement error, and efficient covariate control.

2 Probability basics

In the following sections, we introduce some basic rules and principles from causal graph and probability theory which are necessary to follow the subsequent discussion. Our discussion is necessarily dense; we refer the interested reader to the expositions in Morgan and Winship (2015) and Pearl, Glymour and Jewell (2016).

We begin with *random variables*, Y and X, and their marginal population distributions $P(Y)$ and $P(X)$. In our running example, $P(Y)$ is the distribution of candidate preferences in the population, and $P(X)$ the distribution of educational degrees.

$P(Y|X)$ describes the conditional probability of Y given X. If this probability actually varies as a function of X, we say that X and Y are *dependent*, which we will often loosely refer to as correlated. If $P(Y|X) = P(Y)$, so that $P(Y)$ is constant across values of X, we say that Y and X are *independent* or uncorrelated. $E[Y]$ denotes the population mean of Y, and $E[Y|X]$ denotes the conditional mean of Y given X. The *law of total probability* asserts that

$$P(Y) = \sum_x P(Y|X = x)P(X = x). \tag{1}$$

That is, one can always divide the distribution of a variable Y into distributions of Y conditional on $X = x$ and then get back to the marginal distribution $P(Y)$ by summing over and weighting the conditional distributions with their relative size $P(X = x)$ provided that $P(X = x) > 0$. This works for means as well (*law of iterated expectations*):

$$E[Y] = \sum_x [E|X = x]P(X = x). \tag{2}$$

The law of total probability is the basis for post-stratification and other survey adjustment

strategies (e.g., Lohr, 2019). In such a case, one knows the group (x-specific) distribution of Y from the survey. Then one can estimate the overall population distribution of Y by weighting these group distributions by the size of the groups in the population known from external sources.

3 Graph basics

In general terms, causal graphs encode a researcher's assumptions about the causal process that generated the data in a population of interest. They consist of variables, or *nodes*, which are linked by *arcs*. We use arrows for arcs to depict the hypothesis that one node influences the other. The presence of an arc between two nodes merely indicates that a certain causal effect might be present. On the other hand, the absence of an arc between two nodes indicates a critical assumption, namely that one does not affect the other. A *path* is a sequence of arcs that links one node to another, regardless of the direction of arrows. Retracing arcs or going through the same node twice is not allowed. A *directed* or *causal path* is traced out along arrows tail-to-head. If there is a directed path from one node to another, the former is said to be an *ancestor* of the latter, the latter a *descendant* of the former. A Directed Acyclic Graph, or DAG, contains only directed arrows and no feedback loops (i.e., no variable is its own ancestor or descendant).

The causal assumptions encoded in a DAG constitute the theoretical model on which *identification analysis* rests. Identification analysis is concerned with estimating parameters of interest, regardless of random variability in the data due to small samples or measurement error. In this sense, it is closely related to the statistical notion of *consistency*. The main reason to focus on identification as a first step is that if a parameter cannot be identified using "infinite" and error-free data, we certainly cannot learn anything about it using finite data.[1] So, for the time being, we assume large samples and error-free measurements. We return to issues of estimation and measurement in the conclusion. Finally, the graphical

literature on sample selection and missing data often uses the term *recovery* or *recoverability* of some quantity (Mohan and Pearl, 2021). We will use the terms identification, recovery, and recoverability interchangeably.

4 D-separation

Having discussed the basic semantics of causal graphs, we now turn to the only graphical rule on which we rely: *d-separation*. It allows us to logically infer the absence of correlations from the structure of the graph. We give here a rather dense discussion of the formalities, which will later be illustrated through various examples.

Figure 1 depicts the three basic patterns one needs to understand when determining d-separation. In the left panel, Z is a common cause, or *confounder*, of X and Y. In this scenario, Z induces a statistical association between X and Y, although X and Y do not cause each other, which is indicated by the absence of an arrow between them. In the intermediate graph, the causal effect of X on Y is *mediated* through Z. This also has the consequence of producing a correlation between X and Y because the former influences the latter. In short, we say that these two paths are *open*. Finally, in the right panel, Z is a common effect, or *collider*, on the path between X and Y. In contrast to the two other cases, this does not produce a correlation between X and Y. Here, we say that the path is *blocked* or d-separated by the variable Z.

These patterns of association reverse once we look at the distribution of X and Y conditional on Z, i.e., $P(Y|X,Z)$. For example, this could be done by regressing Y on X while controlling for Z, or by stratification or matching on Z (e.g., Zhang, 2000). In the left graph, Z is the only common cause of X and Y. Accordingly, for units with the same value of Z, the value of X is not informative about Y, and vice versa. In the intermediate graph, conditioning on Z also blocks the information flow from X to Y. X and Y are said to be d-separated conditional on Z. However, in the collider graph on the right, an association

emerges. To understand why an example is helpful.

Consider two independent binary variables X and Y and a random variable Z that is the sum of X and Y. Therefore, Z can take on the values $\{0, 1, 2\}$. X and Y may be random coin flips, so knowing the value of X does not help in predicting Y. However, conditioning on Z means that we are told its value. Knowing that Z is 1, for example, and that X is 0, we know that Y has to be 1. Conditional on Z, X and Y are dependent or d-connected. Put differently, knowing the result of a process and the value of one of its independent inputs also lets us predict the value of the other input.[2]

In sum, in Figure 1, conditioning on the intermediate variable Z blocks the path between X and Y in the first two graphs but opens the path in the right graph. For deriving whether variables X and Y are (conditionally) independent in more complex DAGs, it turns out that one can just enumerate all paths between these variables. If all of these paths are blocked, perhaps conditional on other variables Z, then we say that X and Y are d-separated (conditional on Z). If all of the variables involved are measured, this statement is testable: The distribution of Y (for some value of Z) should not change for different values of X. For instance, if one commits to a specific regression model, the test involves regressing Y on X and Z; the coefficient on X should be zero (One could also use X as the dependent and Y as the independent variable). The only reason that the coefficients can be different from zero is that the DAG is incorrectly specified and there is at least one open path.[3]

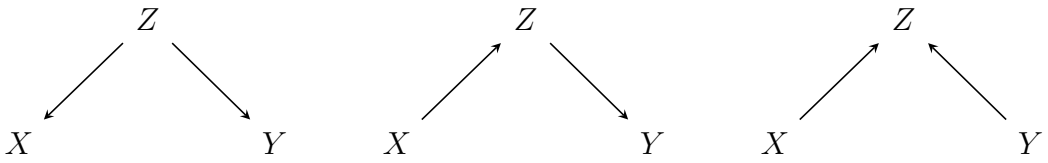

Figure 1: Basic causal patterns and d-separation: Z is a *confounder* (left), a *mediator* (center), or a *collider* (right) on the path between X and Y. In the latter scenario, the path is naturally blocked by Z, which entails that there is no statistical association between X and Y. Otherwise, paths are open and induce statistical association between X and Y.

5 DAGs in causal inference

In causal inference, a simple graphical criterion to determine the exogeneity of a variable X with respect to an outcome Y is the *back-door criterion* (Pearl, 2009, 79). It requires that we look for control, or "adjustment", variables Z such that (1) no such variable is a descendant of X, and (2) Z blocks all paths between X and Y that contain an arrow into X, that is, *back-door* paths. If such variables Z exist and can be measured, adjustment for them – be it through regression, matching, weighting, or another approach – allows estimating the causal effect of X on Y.

We will encounter similar formulations in the following. The DAG literature typically produces graphical criteria that are easy to check given a graph and cover many, perhaps even all, possible strategies to solve a certain problem. Such criteria fundamentally rely on the logic of blocking paths, often complemented with some additional requirements (e.g., that Z may not contain descendants of X). Finally, note that the back-door criterion asks us to consider the relationship (i.e., back-door paths) between X and Y in the graph. While this may seem like a trivial requirement, it implies that such statements as "X is exogenous" are not precise; X can only be exogenous with respect to an outcome Y, and in fact, with respect to a causal model.[4] A similar requirement also occurs for inferences from selective samples. We will return to issues surrounding causal inference and nonresponse weighting in the penultimate section of this paper.

6 Survey inference from a DAG perspective

In this section, we provide detailed explanations of the causal structures behind nonresponse biases and show how existing assumptions for adjustment can be better understood using graphs. While social science scholars interested in causal effects often use survey data, the prime interest of survey researchers typically lies in the population distribution of some variable Y, perhaps given another variable X. A highly visible example which we will be

referring to throughout the following exposition is election polling, where Y is a candidate or party preference, and X is a sociodemographic variable such as education. The context of elections provides ample opportunities for validating survey statistics (e.g., Ansolabehere and Hersh, 2012; Shirani-Mehr et al., 2018), and failures of polls to predict high-profile elections often trigger investigations which yield relevant insights into survey errors.

For instance, Kennedy et al. (2018) evaluate several theories as to why many polls preceding the 2016 U.S. presidential election, particularly in the Midwest, underestimated support for Trump. One theory maintains that preferences for the Democratic candidate (Hillary Clinton) were increasing in formal education in 2016, while they had been U-shaped historically. Since highly educated voters are often overrepresented in surveys, they used to "proxy" for citizens of lower education, but this ceased to be the case in 2016 when preferences differed between those two groups. Accordingly, surveys that did not adjust for education X overestimated Clinton's vote share Y.

Conceptually, selection into the survey sample can be formalized using a binary variable S that is 1 whenever an element of the population is included in a survey and 0 whenever it fails to be. The data one actually measures in the survey are conditional on $S = 1$; i.e., one does not observe $P(Y)$, but merely $P(Y|S = 1)$. Accordingly, sample selection is a problem because one is *forced* to condition on the variable S. From this perspective, unit nonresponse in a probability sample is equivalent to selection into a nonprobability sample, although the resulting biases are often larger and harder to adjust for in the latter case (Cornesse et al., 2020).

In graphs, such – inadvertent – conditioning can be indicated by putting boxes around the selection nodes in Figure 2. In the following, we use these three simple graphs to explain how to represent the prototypical scenarios discussed in the missing data literature (e.g., Little and Rubin, 2019; Rubin, 1976): Missingness Completely At Random (MCAR), Missingness (MAR), and Missingness Not At Random (MNAR) (also see Thoemmes and Mohan, 2015). In Appendix A1, we discuss how our formal setup differs from the one that is common in

the applied survey literature (e.g., Groves, 2006).

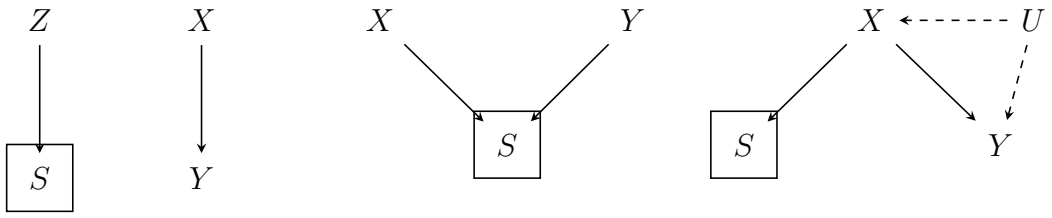

Figure 2: Prototypical selection scenarios. Left: Y is missing completely at random (MCAR), i.e., S and survey outcome Y have separate causes, Z and X. Center: Y influences missingness so that it is missing not at random (MNAR). Right: Y is missing at random (MAR) conditional on X, because X drives S and Y. U is an unobserved confounder.

Missingness Completely at Random (MCAR)

Suppose we are interested in estimating the population distribution of some variable Y from a sample of individuals with $S = 1$. In the left panel of Figure 2, selection node S and survey variable Y have separate sets of causes, Z and X. Since there is no path whatsoever between S and Y, the two variables are d-separated, and we have $Y \perp\!\!\!\perp S$ (read: Y is independent of S). Accordingly, we can write:

$$P(Y|S = 1) = P(Y), \tag{3}$$

so that the distribution we observe in the sample, $P(Y|S = 1)$, asymptotically equals the population distribution we are interested in, $P(Y)$. In the terminology of the missing data literature, the selection process is *ignorable*, and Y is *missing completely at random*.

In a prototypical survey without nonresponse, Z would be the output of some randomization device and the only cause of S. This, in expectation, warrants that the selection of respondents is independent of the survey variable of interest (and, in fact, any other variables including X). Just like randomized treatment assignment in experiments, random selection of respondents is a convenient *procedural* justification of independence. Yet, random selection is not necessary for $Y \perp\!\!\!\perp S$. For instance, assume that we are dealing with an election poll in which Y is party choice and X are policy preferences, where the selection process

is random sampling subject to nonresponse due to the lack of interest in politics Z among sampled individuals (see Groves, Presser and Dipko, 2004). In this situation, the population distribution of Y can be recovered from the survey sample despite nonresponse as long as there is no open path between political interest and policy preferences. On the other hand, even an intact probability sampling scheme will not lead to independent Y and S, if the design entails unequal inclusion probabilities (e.g., for regions) unaccounted for in the analysis, and mean values of Y differed across regions. Also, any probability sampling may produce associations between Y and S by chance (see the discussion in Valliant, Dorfman and Royall, 2000, 19-21).

Missingness Not at Random (MNAR)

In the center panel of Figure 2, both X and Y influence selection into the survey. It is well known that more highly educated persons (X) are overrepresented in surveys, and it was also suspected that there was a relationship between vote preferences (Y) and participation in 2016 pre-election polls (Kennedy et al., 2018). The graph would be consistent with both accounts. Recall that the data one actually observes are conditional on $S = 1$. If we are interested only in the marginal distribution of candidate preferences Y, we will have a problem. If, for example, supporters of a candidate trailing in the polls have a lower propensity to participate in the survey, then people that end up in the respondent pool will be more likely to have a preference for the leading candidate (e.g., Gelman et al., 2016). Accordingly, we cannot make consistent inference on $P(Y)$ since it differs from the distribution we measure, $P(Y|S = 1)$. Whenever the survey outcome directly affects sample inclusion, this problem cannot be solved, at least regarding the quantities of our interest.[5]

A more peculiar implication of this graph is that if the interest is in analytic statistics, in-sample correlations between independent and outcome variables can exist even though there is no relationship between these variables in the population. The existing literature on survey nonresponse does not discuss this phenomenon (Groves, 2006; Peytchev, Carley-Baxter and

Black, 2011; Peytchev, 2013; Mercer et al., 2017). For example, if we are interested in how candidate preference Y varies as a function of education X, we encounter a collider structure where we are forced to condition on selection S. This means that education and preference are correlated in the sample, although in this stylized example no such correlation exists in the overall population. Why is this the case? Assume that both education and preference have a positive effect on the probability of responding in a survey. Then, among those that actually respond ($S = 1$) and that have relatively low education (that is, $X = 0$), preference Y is more likely to be 1 (say, Democrat) because it has to "make up" for the low value of X in producing $S = 1$. Accordingly, we can derive that in such a case, education and preference are negatively correlated in the sample. If we allowed for an effect of X on Y or unobserved confounders to influence these variables, the result would not qualitatively change. For example, we could have a positive correlation in the population while the correlation among sampled subjects is zero or negative. Often, this is called "collider bias".[6] It represents a potential explanation for the result found in Lahtinen et al. (2019), who report that nonresponse bias leads to an underestimation of socioeconomic differences (X) in turnout (Y).[7]

This graphical structure also illustrates why explanations for nonresponse bias that are based only on correlations between observable variables (instead of graphs or structural equations) are insufficient. For example, Kennedy et al. (2018, 4) state that

> "if survey response was correlated with presidential vote and some factor not accounted for in the weighting, then a deficient weighting protocol could be one explanation for the polling errors."

Note that the statement does not refer to causal order. In the collider structure, we have the correlation structure Kennedy et al. describe: survey response S correlates with candidate preference Y, and "some factor" X also correlates with S. However, this itself is not a problem – the collider structure and the conditioning on $S = 1$, however, is – and

weighting or adjusting for X does not solve the problem. In fact, the bias introduced by weighting can be larger than the bias in the unadjusted estimate.

Missingness at Random (MAR)

Sometimes survey research adjusts for differences between sample and population to make consistent inferences. Such an approach is valid if the survey outcome of interest is *missing at random* conditional on these adjustment variables, regardless of the specific method (post-stratification, propensity score weighting, etc.) used (e.g., Valliant, 2020). We will now discuss a prototypical graph where such a strategy works before we return to the problem in more detail in Section 7. Figure 2 depicts a scenario where X are observed confounders—variables that affect both S and Y. In our running example, Kennedy et al. (2018) argue that a crucial element of X is education, which affected both survey participation and candidate preference.

In this case, Y is MAR because X d-separates the selection indicator from the survey outcome of interest. It is helpful to show graphically why this is the case. To check for d-separation, we have to enumerate all paths between S and Y. There are two such paths: $S \leftarrow X \rightarrow Y$ and $S \leftarrow X \leftarrow U \rightarrow Y$. We see that in the first path, X is a confounder, while in the second path it is a mediator (between U and S). Accordingly, both paths are open: There is a correlation between sample selection and survey outcome running through X. After the 2016 presidential election, Kennedy et al. (2018) argued that education positively correlated with both sample selection and preferences for Hillary Clinton. Our graph suggests that this correlation might not only be because of causal effects of education itself, but also because there are deeper unobserved variables (e.g., early-life experiences) that also affect formal education and candidate preferences, but, crucially, do not directly affect survey participation.

In sum, as in the MNAR case, there is a correlation between survey participation and candidate preferences such that $P(Y) \neq P(Y|S = 1)$. However, in this graph, the depen-

dence can be "broken" by conditioning on X. Doing so blocks the two paths connecting S and Y. Accordingly, we have $S \perp\!\!\!\perp Y|X$ and that $P(Y|X, S = 1) = P(Y|X)$. This means that the distribution of preferences as a function of education among sampled persons is the same as in the general population. Our analysis based on a substantively motivated graph and the use of d-separation shows why this works. In contrast, existing explanations rely on stating conditional independence assumptions coupled with informal explanations that are often imprecise, rather than deriving explanations from a deeper formal model. For example, Pfeffermann and Sverchkov (2009, 462) state that one needs to know "all the variables determining sample selection and response". In our graph, variables in U qualify for this criterion; however, we have shown that adjustment for them is not necessary.

If our interest is in the marginal distribution $P(Y)$, we can recover the population distribution of preferences if we know $P(X)$, the population distribution of education, perhaps from a census. We can then use the law of total probability and the assumption implied by the graph ($S \perp\!\!\!\perp Y|X$) to write:

$$\begin{aligned} P(Y) &= \sum_x P(Y|X = x)P(X = x) \\ &= \sum_x P(Y|X = x, S = 1)P(X = x). \end{aligned} \tag{4}$$

In the expression on the right-hand side, $P(Y|X = x, S = 1)$ can be estimated from our sample, and $P(X = x)$ is known from the census.

7 Adjustment for conditional distributions

The preceding discussion has covered all relevant cases for recovering marginal distributions. We have also seen that one can recover conditional distributions when the conditioning variable (regressor) X is not only of substantive interest but also suffices to break the dependence between S and Y. The more typical case is one where the substantive interest is on $P(Y|X)$,

but where it seems unlikely that X alone suffices for selection adjustment and additional variables Z are needed (see Figure 3 below). In this case, one cannot simply insert Z into the analysis (e.g., the regression model) because adjusting for it will usually change both the interpretation and the actual value of the regression coefficient of X. Curiously, we believe that this is a very common situation, yet it is never discussed in textbooks and has escaped the attention of the recent methodological literature (e.g., Elliott and Valliant, 2017). An exception is Pfeffermann and Sverchkov (2009), whose discussion is technical and not related to formal models of response behavior.

We saw earlier that if adjustment is possible, it involves adjusting for variables X that break the dependence between sample selection and outcome Y. If the interest is in a distribution already conditioned on X, but one thinks additional variables Z are needed, it is intuitive that identification is possible if X and Z together do the job to d-separate Y from S. In fact, a simple yet exhaustive graphical criterion for recoverability of $P(Y|X)$ is that (X, Z) d-separate S from Y, which implies $Y \perp\!\!\!\perp S|X, Z$.[8] Using this and the law of total probability, we can write (Bareinboim, Tian and Pearl, 2014, 2413):

$$P(Y|X) = \sum_{Z=z} P(Y|X = x, Z = z, S = 1)P(Z = z|X = x). \tag{5}$$

In short, this expression asks us to compute the distribution (or mean) of Y given X and Z in the sample, and then to average over the X-specific population distribution of Z to obtain an estimate of the population distribution of Y conditional on X. This is a post-stratification estimator not for means, but for conditional distributions and means (see also Gelman, 2007). Under similar assumptions, one can also justify the use of inverse-probability weighting estimators, including general maximum likelihood and two-stage estimators, where the regression of interest is on X, and the weights are computed using X and Z. We formally analyze such estimators in section A2 in the Appendix.

8 Multiple selection nodes

So far, we have followed the literature on selection graphs and collapsed the data collection process into a single selection node S in order to understand the classical scenarios from the missingness literature. However, selecting respondents for surveys usually involves multiple stages. The TSE framework distinguishes between at least three of them: the *coverage stage* (1), in which members of the target population are selected into the frame population from which individuals are sampled at the *sampling stage* (2), and the *response stage* (3), in which sampled individuals further self-select according to their contactability, ability and willingness to participate in the survey. In each of these stages, inclusion may hinge on its own set of factors and can be related to the survey variable of interest in different ways. For instance, landline sampling frames are known for their lack of coverage of young and low-status persons (e.g., Blumberg and Luke, 2007); age and gender further affect the contactability of sampled subjects, presumably via their effects on employment status and mobility (e.g., Stoop, 2005); and social involvement as well as interactions between survey and respondent features, such as topic interest or trust in the sponsorship of a survey, are considered to be crucial for the cooperativeness among those successfully contacted (e.g., Groves, Singer and Corning, 2000). In this section, we show that graphs are ideal for visualizing such multi-stage selection processes and that the common practice of using "design weights" only relies on very strong causal assumptions that are usually left implicit. For example, the General Social Survey only contains various design weights. Its codebook recommends to use these, but makes no similar recommendation (or provisions) for nonresponse adjustments (Smith et al., 2019, Appendix A).

Multiple selection stages are easily represented in a graph by separate selection nodes. Consider the graphs in Figure 3, each of which includes two selection nodes, S_1 and S_2, for the sampling and the response stage, respectively.[9] S_1 affects S_2 by design. Fundamentally, the obtained sample is conditional on both $S_1 = 1$ and $S_2 = 1$. The question, therefore, is whether the correlation between those two variables and the outcome of interest Y can be

broken using an adjustment strategy. The left graph encodes the assumption that Z affects the selection of a unit in the sampling stage (S_1), but does not directly affect whether a sampled unit actually responds (S_2). Furthermore, Z affects the outcome. In this case, the analysis is very similar to the MAR case discussed previously: Z d-separates both selection variables from Y because it acts as a confounder on the paths between these variables and the outcome. Accordingly, the post-stratification formula $P(Y) = \sum_z P(Y|S_1 = 1, S_2 = 1, Z = z)P(Z = z)$ applies. In fact, for purposes of adjustment, we gain nothing by differentiating between S_1 and S_2.

Even though this graph may strike one as highly simplified and loaded with unrealistic assumptions concerning Z and (S_1, S_2), it is one of the weakest sets of assumptions that can be used to justify the common practice of "base" or "design weighting". For example, in telephone surveys, a sample of households is often drawn using random-digit dialing. On the first call, one person is selected so that individual inclusion probabilities at the sampling stage S_1 are inversely proportional to the number of eligible persons per household Z. This yields design weights $P(S_1 = 1|Z) = \frac{1}{Z}$. These are then used for estimating marginal or conditional distributions. Such an analysis disregards the response stage S_2 completely. In this graph, this adjustment strategy works because Z and S_1 d-separate S_2 from Y, so

$$\begin{aligned} P(Y) &= \sum_z P(Y|S_1 = 1, S_2 = 1, Z = z)P(Z = z) \\ &= \sum_z P(Y|S_1 = 1, Z = z)P(Z = z). \end{aligned} \tag{6}$$

It should be clear by now that such an approach does not yield consistent estimators if Z also affects whether units respond to the survey request. This, unfortunately, seems quite likely (Gelman, 2007): Household size plausibly has a positive direct effect on contactibility and thereby S_2, consistent with the intermediate panel of Figure 3, while it has a negative effect (by design) on whether a unit is sampled in the first place. Formally, in this case, one would need to apply nonresponse weights based on $P(S_2|Z)$, as discussed previously.[10]

Furthermore, a simple simulation illustrates that in such a case, design-weighted estimates may be more biased than unweighted estimates. Suppose the (infinite) population is described by the following random variables:

$$Z \sim \text{Normal}(0, 1), \tag{7}$$

$$S_1 \sim \text{Bernoulli}(p = logit^{-1}(2 + Z)), \tag{8}$$

$$S_2 \sim \begin{cases} \text{Bernoulli}(p = logit^{-1}(-Z)) \text{ if } S_1 = 1, \\ 0 \text{ otherwise,} \end{cases} \tag{9}$$

$$Y \sim \text{Normal}(2 + 2 * Z, 5). \tag{10}$$

Under this model, which is consistent with the middle graph of Figure 3, Z has a positive effect on S_1, but a negative effect on S_2.

Our interest is in the population mean of Y. We simulate 1000 realizations of the above model, each with 1000 observations. Each realization is one random sample (of size 1000) for which Z is measured. However, Y is only measured for those units for which S_2 happens to be equal to 1 (depending on S_1, Z, and noise). The selection process means that Y is measured for around 40% of the sampled observations on average.

We compute design weights as well as nonresponse weights by running logistic regressions of S_1 and S_2, respectively, on Z in the realized sample. For each realized sample, we compute unweighted, design-weighted, as well as nonresponse-weighted means. The true population mean of Y is 2. On average, the unweighted estimate is around 1.47, the design-weighted estimate is around 1.17, and the nonresponse weighted estimate is around 1.88. Clearly, the nonresponse-weighted estimator has the smallest bias (which would approach 0 as the realized sample size increases), while the design-weighted estimator has the largest bias.

Finally, what are the consequences of unobserved confounders of the sampling and response stages, as in the right panel of Figure 3? Possession of a landline phone is an obvious

variable that comes to mind: It clearly affects whether a unit may be sampled and may also have independent effects on individual response behavior S_2. In contrast to the left graph, response is now informative about the outcome Y, conditional on being sampled. This is because S_1 is a collider on the path $S_2 \leftarrow U \rightarrow S_1 \leftarrow Z \rightarrow Y$. However, Z still suffices as an adjustment variable. In fact, merely using design weights works. This is because Z d-separates S_2 from Y, conditional on S_1, so S_2 can be ignored. Even though S_1 is a collider that is conditioned upon, Z is a confounder on the same path, and so blocks it again.

As a final note, multiple selection nodes can also be used to depict item-specific missingness, which makes sense with sensitive survey items subject to substantial nonresponse. However, unit nonresponse has become much more prevalent than item nonresponse in recent years (see Yan and Curtin, 2010). See Thoemmes and Mohan (2015) and Moreno-Betancur et al. (2018) for a graph theoretical account of item nonresponse.

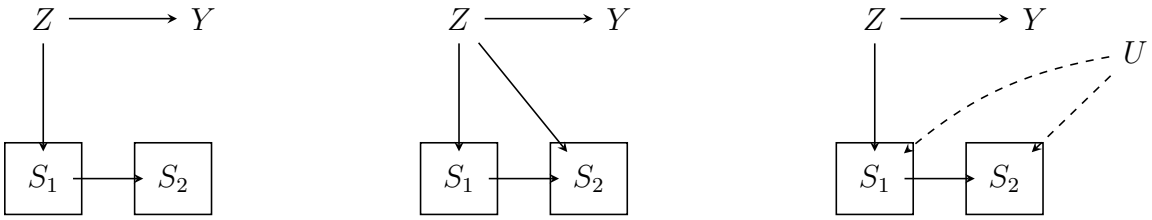

Figure 3: Recoverability with multi-stage selection: S_1 is selection through sampling, S_2 is selection at the response stage, Y is a survey variable of interest, Z is an auxiliary variable, and U is an unobserved variable.

9 Do adjustment variables need to correlate with sample selection and outcome?

In the preceding discussion, we have emphasized that the assumptions for non-response adjustment implicitly or explicitly invoked are (conditional) independence assumptions—and only those. This is at odds with a sizable literature that investigates new adjustment variables from the survey process or administrative sources (Kreuter et al., 2010; Kreuter

and Olson, 2011, 2013; Peytchev, Presser and Zhang, 2018; Sakshaug and Antoni, 2019). Here, one finds different, or at least additional, requirements for valid adjustment variables. A representative statement reads, "to reduce bias effectively without increasing variance, a covariate that is used for non-response weighting adjustment needs to be highly associated with both the response indicator and the survey outcome variable" (Kreuter et al., 2010, abstract).[11]

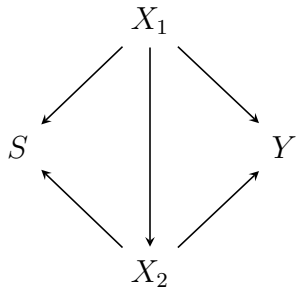

Figure 4: Graph where adjustment variables X_1 and X_2 may *not* correlate with Y due to offsetting paths, but the induced non-response bias may be substantial.

Using causal graphs, it is easy to see why this is not always true. Consider the graph in Figure 4. X_1 could be age, while X_2 could be income (Kreuter and Olson, 2011, 316), and Y political preferences.

Here, the population correlation between X_2 and Y consists of the paths $X_2 \rightarrow Y$ and $X_2 \leftarrow X_1 \rightarrow Y$. These two paths may produce a very small or even zero correlation, e.g. when the causal effect of X_2 on Y is positive, but the association due to the confounding path is negative. X_2's correlation with S may be zero for the same reasons. Accordingly, the heuristic used in the literature would suggest that X_2 is an unimportant adjustment variable. However, conditional on X_1—which is clearly needed for adjustment—one path between X_2 and Y (as well as S) would be blocked, and the association left over could be strong, indicating the significant consequences of additionally adjusting for X_2. Indeed, we later show via a simple simulation how X_2 may not correlate at all with Y and S, but would still be required for, and vastly improve, nonresponse adjustment.

At its core, this phenomenon occurs because conditioning on extra variables may decrease,

increase, or reverse existing correlations. However, research on similar phenomena in causal inference, mostly guided by graphs, suggests that this is not well known to applied researchers and that the associated issues can be subtle and complex (Pearl, 2011; Steiner and Kim, 2016). Inspired by experimental work on survey incentives, we also analyze a scenario in Appendix A3 where the opposite occurs: The randomized incentive seems like an ideal adjustment variable that is strongly correlated with both S and Y—but in fact adjustment for it can never help and may actually hurt.

Kreuter and Olson (2011) analyze a scenario very similar to the above graph using simulations, and illustrate it using a DAG. However, as far as we can see, they do not conclude that inspecting correlations between adjustment variables and selection/outcome has no natural value for whether the variables may be useful, and this may be one explanation for why it is still a wide-spread practice (Peytchev, Presser and Zhang, 2018; Sakshaug and Antoni, 2019).

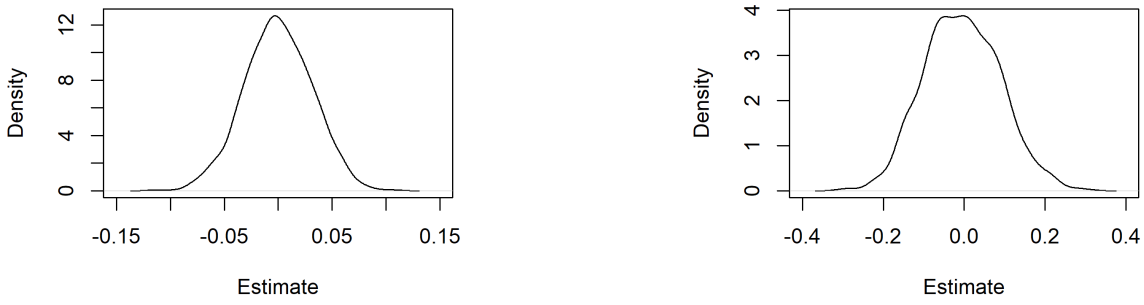

Figure 5: Sampling distributions of correlation between X_1 and S (left) and of coefficient from a linear regression of Y on X_1 among respondents (right) across 1000 replications.

To further illustrate the problem, we use a simple and straightforward simulation setup similar to the ones in Kreuter and Olson (2011), and consistent with Figure 4:

$$X_1 \sim N(0,1), \tag{11}$$

$$X_2 \sim N(X_1,1), \tag{12}$$

$$S \sim \text{ Bernoulli}(p = logit^{-1}(2 + X_1 - X_2)), \tag{13}$$

$$Y \sim N(2 + X_1 - X_2, 2). \tag{14}$$

The interest is in the population mean of Y. We simulate 1000 realizations of the above model, each with 1000 observations. As in our previous simulation, the whole model describes the random variables in the (infinitely large) population, and each realization is one random sample (of size 1000) for which X_1 and X_2 are measured. However, Y is only measured for those units for which S happens to be equal to 1 (depending on X_1, X_2, and noise). The selection process means that Y is measured for around 77.5% of the sampled observations on average across samples, so the response probability is still quite high.

We evaluate the following diagnostics and estimators, in line with existing practice: First, we look at the correlation of X_1 and X_2 with S, based on the sampling frame (i.e., all 1000 observations). We then look at the regression of Y on X_1 or X_2 in the realized sample (conditional on $S = 1$). Finally, we evaluate a weighting estimator for the mean of Y, based on inverse-probability weighting where we fit a logit model for S using both adjustment variables, or just the one that correlates with S or Y.

We start with the adjustment variable X_2. On average, it correlates strongly with sample selection ($\rho \approx -0.48$) and is also strongly associated with Y in the sample ($\beta \approx -0.75$). This would lead us to correctly infer that it is an important adjustment variable.

However, Figure 5 plots the sampling distributions for the same statistics for adjustment variable X_1. Clearly, these are centered around zero. Most of the time, we would ignore X_1 as an adjustment variable, because its correlations with both S and Y would be exceedingly small.

However, when we evaluate the weighting estimators, the one based on both variables clearly outperforms the one based only on X_2. The latter has a relative bias of almost 10%, while the former is approximately unbiased. Even more impressively, using X_1 reduces mean squared error by 75%.

Accordingly, looking at correlations of adjustment variables with S or Y (the latter in sample) can lead analysts astray—in the extreme example here, one adjustment variable is completely uncorrelated with both S and Y (in sample), but still vastly improves nonresponse adjustment. Based on our analysis, we recommend researchers regard correlations of adjustment variables with S and Y at most as tentative evidence, and not as a basis for deciding which variables to use.

We note a further problem with these empirical studies, in that they mistake the in-sample correlation of X and Y for their population correlation. The preceding argument clearly relies on (conditional) population correlations, but the latter are not directly available to researchers, as Y is measured only for respondents. The in-sample correlation analyzed in the literature (Kreuter et al., 2010; Peytchev, Presser and Zhang, 2018; Sakshaug and Antoni, 2019), on the other hand, may be badly biased for the population correlation. For example, in the graph above, conditioning on S opens the path $X_2 \rightarrow S \leftarrow X_1 \rightarrow Y$. But this is not a constitutive part of the population correlation of X_2 and Y, and it will introduce bias.

Finally, we may also have the opposite case: An observed variable correlates strongly with both S and Y (in-sample), but must be ignored for nonresponse adjustment. Survey incentivization experiments (e.g., Wing, 2019) may be one such case where plausible prior knowledge and graphical analysis tell analysts to disregard an observed variable completely.[12] We analyze this problem in Appendix A3. This is one more reason to disregard such sample-based estimates for diagnosis. Instead, researchers need to rely on theoretical, causal considerations (possibly depicted by a graph) first.

10 Nonresponse weighting when the interest is in causal effects

The preceding discussion was focused on inferring population means, associations, regression coefficients, and related quantities. Here, a graphical perspective (as well as the older "missingness" terminology) suggests a black-white picture: Either certain observed variables are sufficient to block connections between sample inclusion and survey variables of interest, or they are not. As shown by Kreuter and Olson (2011), in-between cases where only a subset of the necessary adjustment variables are used may lead to estimators that are more biased than unadjusted estimators. However, our impression is that when applied researchers use nonresponse weights, they generally hope for weighting to decrease (absolute) bias.

In this section, we show that when the interest is in estimating causal effects rather than marginal and conditional population quantities, "hopeful weighting" under the tacit acknowledgment of remaining biases is generally ill-advised. In essence, this is because inferences on means and regression coefficients are most often only interesting insofar as they apply to a larger population, but this doesn't hold true for causal effects.

Consider again our running example of education X and vote preference Y. Knowing the population correlation between these two variables informs us about societal cleavages and may also be useful for political parties to adjust their platforms or to target voters in other ways. An estimate of this correlation from a selective sample, on the other hand, is arguably only relevant insofar as it approximates the population quantity. However, the causal effect of education on vote choice from a selective sample may be of scientific value regardless of its relationship with the population causal effect. For example, a significant non-zero average effect in the sample of any size allows us to reject Fisher's "sharp null" hypothesis of no causal effect in the population. And a significant large average causal effect in the sample rejects the hypothesis that causal effects in the population are of at most limited (in some substantive sense) size.

There now is indeed a burgeoning literature that concentrates on statistical ("randomization") inference for such sample causal effects without any reference to a broader population of unsampled units (e.g., Abadie et al. 2020, Li and Ding 2017). And the implications of acknowledging the sample causal effect as a valuable quantity on its own are profound. For population means and regressions, the usual warning is that weighting may increase or decrease bias. But for sample causal effects, weighting always introduces (asymptotic) bias, unless average sample and population causal effects are the same, in which case weighting (asymptotically) does nothing.

We can make this point more formally. Consider Figure 6. Here, D is a (possibly randomized) treatment of interest and Y is the outcome of interest. There are observed adjustment variables X that influence both Y and S and for which population information ($P(X)$) is available. U stands for unobserved confounders that similarly impact on Y and S. X and U together create a connection between Y and S so that the sample is not representative of the general population. Furthermore, since both X and U influence Y, they are also possible effect modifiers. Across units, causal effects may vary due to differences in X or U or both.

Using potential outcomes notation (e.g., Morgan and Winship, 2015), the sample causal effect of switching D from d to d' is identified as

$$E[Y(d') - Y(d)|S = 1] = E[Y|D = d', S = 1] - E[Y|D = d, S = 1],$$

because S fulfills the back-door criterion with respect to this conditional effect of D on Y (e.g., Pearl, Glymour and Jewell 2016, 70).

At the same time, we can use our usual reasoning to quickly determine that the sample regression of Y on D that could, in turn, be used to estimate this causal effect is not representative of the population regression (and, therefore, also not useful for estimating the population causal effect). Using X for weighting does change this basic effect; unless we also measure and adjust for U, there is an open path between S and Y. At the same time,

weighting for X will generally change the resulting estimate.

In sum, we here have a case where the in-sample regression is unbiased for the (potentially) scientifically interesting sample causal effect, whereas imperfectly weighted estimates are generally inconsistent for both the sample and the population causal effect. The implications are clear: When researchers are interested in causal effects, believe that observed variables are not sufficient to recover from nonresponse, and are unwilling to make further parametric assumptions, there appears to be little reason to weight at all. This is in contrast to the situation where population means or regression coefficients are of genuine interest; in that case, one may at least hope that weighting decreases absolute bias.

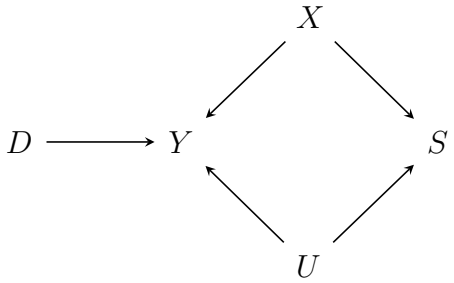

Figure 6: Graph for causal inference from survey data suffering from nonresponse: D is a treatment of interest, Y the outcome of interest. Observed variables X and unobserved variables U impact on Y and S and both are sources of effect modification (varying causal effects).

11 Conclusion

In this article, we have synthesized recent developments in the literature on causal graphs so that they are of great theoretical and practical value for survey researchers. By emphasizing the relationship between graphs and statistical dependencies, we have highlighted that graphs are tightly related to the statistical machinery that survey researchers are already using. Insofar as their use has been informal (e.g., Groves, 2006), we have argued that they should be seen as formal mathematical models of the survey response process.

Specifically, causal graphs can be used to understand and communicate complex assump-

tions for nonresponse adjustment such as MAR and to derive general and straightforward estimators of marginal and associational parameters if certain assumptions are met. Furthermore, multiple selection stages can easily be depicted using causal graphs, and we have shown the implicit assumptions behind using design weights. In this way, graphs allow for a unified view of random sampling and nonprobability samples, and design- and model-based survey inference (Elliott and Valliant, 2017; Kohler, Kreuter and Stuart, 2019). Finally, we have shown how graphical analysis can help us to understand why existing practices like checking correlations between potential adjustment variables and selection indicators or survey outcome, or the "hopeful" weighting of causal effect estimates, can lead us astray.

There are several interesting areas where future work can use causal graphs to visualize and better understand core assumptions in survey adjustment. One is the use of sample selection models when the assumptions for the adjustment strategies that we have discussed are not plausible. While early approaches (e.g., Heckman, 1979) relied on heavy parametric assumptions such as linearity in the structural models, more recent work relaxes these but emphasizes the importance of error independencies and exclusion restrictions (Das, Newey and Vella, 2003). Both of these assumptions sets can be assessed using causal graphs.

We have touched upon causal inference under sample selection and argued that concentrating on sample causal effects, without imperfect adjustments for nonresponse, may often be justified. Recent work on "transportability" (Bareinboim and Pearl, 2016) considers the distinct task of generalizing causal effect estimates obtained from one study population to different populations. Transportability requires causal assumptions as well as non-experimental, observational data from the target population. For the latter, survey data will be indispensable.

In causal inference, another active area of research is the analysis of sensitivity to unobserved confounders. Mercer et al. (2017) call for similar developments with regard to sample selection bias. Recently, Smith and VanderWeele (2019) have developed a fairly general approach to this problem, guided by graphical assumptions.

The inference from self-reports to underlying attributes – measurement – is a further central task of any survey researcher. Graphs are excellent tools to depict complex measurement models and are widely used in conjunction with parametric assumptions in the structural equation modeling tradition (e.g., van der Veld and Saris, 2004). In the causal graph literature, weak assumptions on the sign and heterogeneity of the effects of latent constructs on measured quantities have been used to infer the existence of causal effects (VanderWeele and Hernán, 2012). We suspect that there is ample room for the development of empirical strategies for survey researchers using these approaches, especially with regard to the interaction of errors induced by sample selection and measurement error (Groves, 2006; Tourangeau, 2019; Olson, 2019).

Finally, we have not discussed the finite sample behavior – and especially the mean squared error – of various adjustment strategies. This is because we have been concerned with identification, which is the first logical step in nonresponse adjustment strategies. However, it has long been known that nonresponse adjustments may lead to estimates with high variance (e.g., Little, 1986) This is usually investigated using Monte Carlo simulations, and research often relies on causal graphs to depict the structure behind the simulated data (Kreuter and Olson, 2011). There is some previous work in economics employing causal graphs to guide *efficient* covariate control for causal inference (White and Lu, 2011) that might have implications for nonresponse adjustment as well.

Notes

1. Often, however, parameters can be set-identified (bounded) under weak assumptions. Our interest in this article is on point identification. For bounding parameters under sample selection, see Manski (2009).

2. The same mechanics apply if we happen to know the realization of a descendant of Z. For example, let D be a variable that takes on the value 1 when Z equals 1, otherwise 0 (so that it is a binary proxy for Z). Knowing that D equals 1 and that X equals 0 also leads to the prediction that Y equals 1.

3. In finite samples, the estimate will never be exactly zero.

4. See Pearl (2009, 165–170) for a discussion of different concepts of exogeneity.

5. See Didelez, Kreiner and Keiding (2010) for a graphical account of what population quantities are recoverable under outcome-dependent sampling.

6. See Griffith et al. (2020) for phenomena in data on COVID-19 patients that might be explained by collider bias.

7. It is interesting to note that measurement error due to vote overreporting has the opposite effect since misreporters among nonvoters look like voters in terms of their SES (e.g., Ansolabehere and Hersh, 2012)

8. Although this technically is just a sufficient condition, and not necessary, it covers most scenarios where auxiliary variables are available (see Bareinboim, Tian and Pearl, 2014).

9. For ease of presentation, we assume that the frame population includes all members of the target population so that there is no selection through (under-)coverage.

10. This is because $S_2 = 1$ implies $S_1 = 1$ logically, so the sampled data $P(Y|S_1 = 1, S_2 = 1)$ can also be written as $P(Y|S_2 = 1)$. Therefore, weighting for S_2 is sufficient.

11. Similarly, Peytchev, Presser and Zhang (2018) state in their abstract that such an association is a "*sine qua non* for effective adjustment" (emphasis in original).

12. Instead, Wing (2019) shows that such an incentivization experiment can be used to perform an instrumental-variables-style analysis of the consequences of nonresponse.

References

Abadie, Alberto, Susan Athey, Guido W Imbens and Jeffrey M Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88(1):265–296.

Ansolabehere, Stephen and Eitan Hersh. 2012. "Validation: What big data reveal about survey misreporting and the real electorate." *Political Analysis* 20(4):437–459.

Bareinboim, Elias, Jin Tian and Judea Pearl. 2014. Recovering from selection bias in causal and statistical inference. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Bareinboim, Elias and Judea Pearl. 2016. "Causal inference and the data-fusion problem." *Proceedings of the National Academy of Sciences* 113(27):7345–7352.

Blumberg, Stephen J and Julian V Luke. 2007. "Coverage bias in traditional telephone surveys of low-income and young adults." *Public Opinion Quarterly* 71(5):734–749.

Cornesse, Carina, Annelies G Blom, David Dutwin, Jon A Krosnick, Edith D De Leeuw, Stéphane Legleye, Josh Pasek, Darren Pennay, Benjamin Phillips, Joseph W Sakshaug et al. 2020. "A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research." *Journal of Survey Statistics and Methodology* .

Daniel, Rhian M, Michael G Kenward, Simon N Cousens and Bianca L De Stavola. 2012. "Using causal diagrams to guide analysis in missing data problems." *Statistical Methods in Medical Research* 21(3):243–256.

Das, Mitali, Whitney K Newey and Francis Vella. 2003. "Nonparametric estimation of sample selection models." *The Review of Economic Studies* 70(1):33–58.

Didelez, Vanessa, Svend Kreiner and Niels Keiding. 2010. "Graphical models for inference under outcome-dependent sampling." *Statistical Science* 25(3):368–387.

Elliott, Michael R and Richard Valliant. 2017. "Inference for nonprobability samples." *Statistical Science* 32(2):249–264.

Elwert, Felix and Christopher Winship. 2014. "Endogenous selection bias: The problem of conditioning on a collider variable." *Annual Review of Sociology* 40:31–53.

Gelman, Andrew. 2007. "Struggles with survey weighting and regression modeling." *Statistical Science* 22(2):153–164.

Gelman, Andrew, Sharad Goel, Douglas Rivers, David Rothschild et al. 2016. "The mythical swing voter." *Quarterly Journal of Political Science* 11(1):103–130.

Griffith, Gareth, Tim T Morris, Matt Tudball, Annie Herbert, Giulia Mancano, Lindsey Pike, Gemma C Sharp, Tom M Palmer, George Davey Smith, Kate Tilling, Luisa Zuccolo, Neil M Davies and Gibran Hemani. 2020. "Collider bias undermines our understanding of COVID-19 disease risk and severity." *medRxiv* .
URL: *https://www.medrxiv.org/content/early/2020/05/08/2020.05.04.20090506*

Groves, Robert M. 2006. "Nonresponse rates and nonresponse bias in household surveys." *Public Opinion Quarterly* 70(5):646–675.

Groves, Robert M, Eleanor Singer and Amy Corning. 2000. "Leverage-saliency theory of survey participation: description and an illustration." *The Public Opinion Quarterly* 64(3):299–308.

Groves, Robert M and Emilia Peytcheva. 2008. "The impact of nonresponse rates on nonresponse bias: a meta-analysis." *Public Opinion Quarterly* 72(2):167–189.

Groves, Robert M, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer and Roger Tourangeau. 2009. *Survey methodology*. 2 ed. John Wiley & Sons.

Groves, Robert M and Lars Lyberg. 2010. "Total survey error: Past, present, and future." *Public Opinion Quarterly* 74(5):849–879.

Groves, Robert M, Stanley Presser and Sarah Dipko. 2004. "The role of topic interest in survey participation decisions." *Public Opinion Quarterly* 68(1):2–31.

Heckman, James J. 1979. "Sample selection bias as a specification error." *Econometrica* pp. 153–161.

Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua D Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen Olson, Douglas Rivers et al. 2018. "An evaluation of the 2016 election polls in the United States." *Public Opinion Quarterly* 82(1):1–33.

Kohler, Ulrich, Frauke Kreuter and Elizabeth A Stuart. 2019. "Nonprobability sampling and causal analysis." *Annual Review of Statistics and its Application* 6:149–172.

Kreuter, Frauke and Kristen Olson. 2011. "Multiple auxiliary variables in nonresponse adjustment." *Sociological Methods & Research* 40(2):311–332.

Kreuter, Frauke and Kristen Olson. 2013. "Paradata for nonresponse error investigation." *Improving surveys with paradata: Analytic uses of process information* 2:13–42.

Kreuter, Frauke, Kristen Olson, James Wagner, Ting Yan, Trena M Ezzati-Rice, Carolina Casas-Cordero, Michael Lemay, Andy Peytchev, Robert M Groves and Trivellore E Raghunathan. 2010. "Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173(2):389–407.

Lahtinen, Hannu, Pekka Martikainen, Mikko Mattila, Hanna Wass and Lauri Rapeli. 2019. "Do Surveys Overestimate or Underestimate Socioeconomic Differences in Voter Turnout? Evidence from Administrative Registers." *Public Opinion Quarterly* .

Li, Xinran and Peng Ding. 2017. "General forms of finite population central limit theorems

with applications to causal inference." *Journal of the American Statistical Association* 112(520):1759–1769.

Little, Roderick JA. 1986. "Survey nonresponse adjustments for estimates of means." *International Statistical Review/Revue Internationale de Statistique* pp. 139–157.

Little, Roderick JA and Donald B Rubin. 2019. *Statistical analysis with missing data*. John Wiley & Sons.

Lohr, Sharon L. 2019. *Sampling: Design and Analysis*. Chapman and Hall/CRC.

Manski, Charles F. 2009. *Identification for prediction and decision*. Harvard University Press.

Mercer, Andrew W, Frauke Kreuter, Scott Keeter and Elizabeth A Stuart. 2017. "Theory and practice in nonprobability surveys: parallels between causal inference and survey inference." *Public Opinion Quarterly* 81(S1):250–271.

Mohan, Karthika, Felix Thoemmes and Judea Pearl. 2018. Estimation with Incomplete Data: The Linear Case. In *IJCAI*. pp. 5082–5088.

Mohan, Karthika and Judea Pearl. 2021. "Graphical models for processing missing data." *Journal of the American Statistical Association* pp. 1–42.

Moreno-Betancur, Margarita, Katherine J Lee, Finbarr P Leacy, Ian R White, Julie A Simpson and John B Carlin. 2018. "Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies." *American journal of epidemiology* 187(12):2705–2715.

Morgan, Stephen L and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

Olson, Kristen. 2019. "Comments On "How Errors Cumulate: Two Examples" by Roger Tourangeau." *Journal of Survey Statistics and Methodology* .

Pearl, Judea. 2009. *Causality*. Cambridge University Press.

Pearl, Judea. 2011. "Invited commentary: understanding bias amplification." *American journal of epidemiology* 174(11):1223–1227.

Pearl, Judea, Madelyn Glymour and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Peytchev, Andy. 2013. "Consequences of survey nonresponse." *The ANNALS of the American Academy of Political and Social Science* 645(1):88–111.

Peytchev, Andy, Lisa R Carley-Baxter and Michele C Black. 2011. "Multiple sources of nonobservation error in telephone surveys: coverage and nonresponse." *Sociological Methods & Research* 40(1):138–168.

Peytchev, Andy, Stanley Presser and Mengmeng Zhang. 2018. "Improving traditional nonresponse bias adjustments: Combining statistical properties with social theory." *Journal of Survey Statistics and Methodology* 6(4):491–515.

Pfeffermann, Danny and Michail Sverchkov. 2009. Inference under informative sampling. In *Handbook of statistics*. Vol. 29 Elsevier pp. 455–487.

Rubin, Donald B. 1976. "Inference and missing data." *Biometrika* 63(3):581–592.

Sakshaug, Joseph W and Manfred Antoni. 2019. "Evaluating the utility of indirectly linked federal administrative records for nonresponse bias adjustment." *Journal of Survey Statistics and Methodology* 7(2):227–249.

Schafer, Joseph L and John W Graham. 2002. "Missing data: our view of the state of the art." *Psychological methods* 7(2):147.

Shirani-Mehr, Houshmand, David Rothschild, Sharad Goel and Andrew Gelman. 2018. "Disentangling bias and variance in election polls." *Journal of the American Statistical Association* 113(522):607–614.

Smith, Louisa H and Tyler J VanderWeele. 2019. "Bounding bias due to selection." *Epidemiology* 30(4):509–516.

Smith, Tom W., Michael Davern, Jeremy Freese and Stephen L. Morgan. 2019. *General Social Surveys, 1972-2018: cumulative codebook*. NORC.

Steiner, Peter M and Yongnam Kim. 2016. "The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases." *Journal of causal inference* 4(2).

Stoop, Ineke AL. 2005. *The hunt for the last respondent: Nonresponse in sample surveys*. Vol. 200508 Sociaal en Cultureel Planbu.

Thoemmes, Felix and Karthika Mohan. 2015. "Graphical representation of missing data problems." *Structural Equation Modeling: A Multidisciplinary Journal* 22(4):631–642.

Tourangeau, Roger. 2019. "How Errors Cumulate: Two Examples." *Journal of Survey Statistics and Methodology* .

Valliant, Richard. 2020. "Comparing alternatives for estimation from nonprobability samples." *Journal of Survey Statistics and Methodology* 8(2):231–263.

Valliant, Richard, Alan H. Dorfman and Richard M. Royall. 2000. *Finite population sampling and inference. A prediction approach*. John Wiley & Sons.

van der Veld, William and Willem E. Saris. 2004. Separation of Error, Method Effects, Instability, and Attitude Strength. In *Studies in Public Opinion: Attitudes, Nonattitudes, Measurement Error, and Change*, ed. Willem E. Saris and Paul M. Sniderman. Princeton, NJ: Princeton University Press pp. 37–59.

VanderWeele, Tyler J and Miguel A Hernán. 2012. "Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs." *American Journal of Epidemiology* 175(12):1303–1310.

White, Halbert and Xun Lu. 2011. "Causal diagrams for treatment effect estimation with application to efficient covariate selection." *Review of Economics and Statistics* 93(4):1453–1459.

Wing, Coady. 2019. "What Can Instrumental Variables Tell Us About Nonresponse In Household Surveys and Political Polls?" *Political Analysis* 27(3).

Wooldridge, Jeffrey M. 2007. "Inverse probability weighted estimation for general missing data problems." *Journal of econometrics* 141(2):1281–1301.

Yan, Ting and Richard Curtin. 2010. "The relation between unit nonresponse and item nonresponse: A response continuum perspective." *International Journal of Public Opinion Research* 22(4):535–551.

Zhang, Li-Chun. 2000. "Post-stratification and calibration—a synthesis." *The American Statistician* 54(3):178–184.

Supplementary Materials

A1 Comparison to the Analysis in Groves 2006

The three graphs in Figure A1 are adapted from Figure 1 in Groves (2006), and superficially similar to the three prototypical selection graphs in Figure 2 that we have discussed. However, there are important differences. Groves (2006), in line with some of the survey literature on non-response (Bethlehem, 2002), does not work with a binary selection indicator S, but instead with an unobserved response propensity P (normally a probability between 0 and 1). Nonresponse bias is introduced if P correlates with Y.

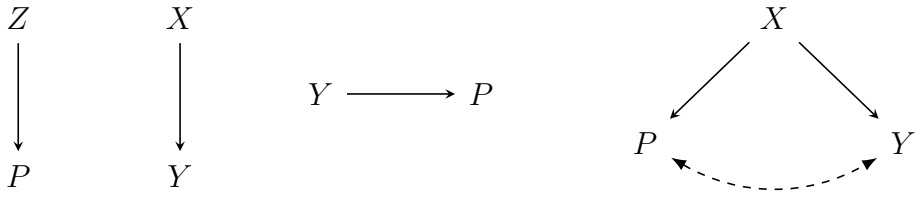

Figure A1: Three causal graphs, adapted from Figure 1 in Groves (2006). P stands for the latent response propensity. The bi-directed arrow in the right graph (presumably) indicates the association induced by X, not a separate unobserved confounder.

A fundamental advantage of using the binary S variable is that it consistent with the literature on partial identification (Manski, 2009) and that it allows for the calculation of assumption-free bounds. Modelling the problem using P does not allow for such calculations, if only because the distribution of P is not known, not even for respondents. Furthermore, this makes it hard to gauge how one could adjust for nonresponse. The adjustment strategies that we discuss rely on modelling empirically the relationship between S and X. With an unobserved P instead of S, this is not possible.

Furthermore, regarding the graphical models in Figure A1, the graph on the left is practically equivalent to the left graph in Figure 2. The center graphs, however, differ. It is only by introducing a second variable X also pointing into P or S that one is able to appreciate the collider bias phenomenon: Correlations may exist in-sample that are absent in the pop-

ulation. Again, it seems that the large literature on survey non-response has never explicitly discussed let alone explained this phenomenon. Finally, the graphs on the right also differ. In Figure A1, it seems (although it is not quite clear) that the additional bidirected path $P \leftrightarrow Y$ is meant to depict the consequence of having the common cause X, and not to visualize a separate unobserved confounder, as is common in the recent DAG literature (Pearl, 2009). This is not a trivial point. Using d-separation, we would conclude that successful non-response adjustment is not possible in this graph.

A2 Inverse Probability Weighting for M-Estimation

We here continue our analysis from section 7 in the main text and show how it relates to general inverse-probability weighted estimators. To reiterate, we assume that (X, Z) block all paths from S to Y and that we have information on $P(Y, X, Z|S = 1)$ from the sample and on $P(X, Z)$ from external data. Accordingly, we can assess the inclusion probabilities conditional on X and Z as

$$P(S = 1|X, Z) = \frac{P(X, Z|S = 1)P(S = 1)}{P(X, Z)}. \tag{A1}$$

The left-hand side is easy to estimate using some regression approach with outcome S if X, Z are available for each observation in the sampling frame. If not, a natural solution is to calculate the expression on the right-hand side for all X, Z.

Our interest is to find an estimator for the parameter(s) β defined as

$$\min_{\beta} E[g(X, Y, \beta)], \tag{A2}$$

where $g()$ is a specified function. For example, if X is a matrix of variables, y is a vector, and $X'X$ has full rank, then the population linear regression coefficients are $\beta = (X'X)^{-1}X'y$. Sample analogs of such minimization problems are called *M-estimators*.

Maximum likelihood estimators are another special case of this very broad class of estimators.

This problem is different from the one considered in Wooldridge (2007). In that article, Z alone is sufficient for adjustment, and both X and Y can be seen as "outcomes". Here, we are interested in a situation where X is both needed for adjustment and as a conditioning variable of substantive interest, whereas Z is merely an auxiliary variable needed for adjustment.

Our approach will be to consider weighted expressions of the function $g()$ for some β (which we suppress for notational clarity) applied to the sampled observations, now indexed by individual i:

$$\frac{S_i g(X_i, Y_i)}{P(S_i = 1|X_i, Z_i)}. \tag{A3}$$

To evaluate this expression, note that $P(S_i = 1|X_i, Z_i)$ is constant when X_i, Z_i are given. Also, $g(X_i, Y_i)$ is random only through Y_i when X_i, Z_i are given, so that under our causal assumptions $S_i \perp\!\!\!\perp g(X_i, Y_i)|X_i, Z_i$.[1] Using this, we have

$$\begin{aligned} E\left[\frac{S_i g(X_i, Y_i)}{P(S_i = 1|X_i, Z_i)}\right] &= E\left[E[\frac{S_i g(X_i, Y_i)}{P(S_i = 1|X_i, Z_i)}|X_i, Z_i]\right] \\ &= E\left[\frac{1}{P(S_i = 1|X_i, Z_i)} E[S_i g(X_i, Y_i)|X_i, Z_i]\right] \\ &= E\left[\frac{1}{P(S_i = 1|X_i, Z_i)} E[S_i|X_i, Z_i] E[g(X_i, Y_i)|X_i, Z_i]\right] \\ &= E\left[\frac{1}{P(S_i = 1|X_i, Z_i)} P(S_i = 1|X_i, Z_i) E[g(X_i, Y_i)|X_i, Z_i]\right] \\ &= E[g(X_i, Y_i)]. \end{aligned} \tag{A4}$$

The first equality uses the law of iterated expectations. The second equality uses the fact that $P(S = 1|X, Z)$ is constant with respect to the inner expectation as well as linearity of expectations. The third equality uses the independence assumption $S \perp\!\!\!\perp g(X, Y)|X, Z$. The fourth equality follows because S is binary. The fifth equality again uses the law of iterated expectations. Under weak regularity conditions, minimizing the sample equivalent

of equation A3 leads to an unbiased estimator of β (Wooldridge, 2007).

A3 Correctly ignoring adjustment variables even when they correlate with sample inclusion and survey outcome

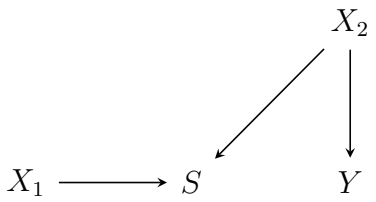

Figure A2: Graph where X_1 is a randomized survey incentive that only influences S, but not Y, and X_2 is another adjustment variable.

The analysis in Section 9 in the main text has uncovered a case where a variable that is needed for nonresponse adjustment does not correlate with S in the population or with Y in the sample. We now discuss the reverse case: A variable correlates with both S and Y, but adjustment for it is not necessary, and may in fact increase bias.

The situation is depicted in Figure A2. Here, X_1 is a randomized survey incentive (e.g., in monetary form) that is offered to some respondents. Due to randomization, there are no other variables pointing into it (that is, only the researchers or the randomization device influence X_1). Furthermore, due to substantive considerations, the incentive is thought to influence actual survey participation, but to have no effect on the survey variable of interest, Y. X_2 is another survey adjustment variable, e.g., demographics.

Clearly, X_2 is necessary and sufficient for nonresponse adjustment. Additional adjustment for the survey incentive X_1 is not needed. However, X_1 will possibly correlate strongly with survey participation to the extent that it actually affects the latter. Furthermore, while there is no open path and therefore no association between X_1 and Y in the population, there is one in sample. This is because S acts as a collider on the path $X_1 \rightarrow S \leftarrow X_2 \rightarrow Y$.

Accordingly, the heuristics used in the literature on adjustment variables will indicate that one should adjust for X_1, which actually introduces bias where none has been before.

Notes

1. The latter fact cannot be derived using d-separation.

References

Bethlehem, Jelke. 2002. Weighting Nonresponse Adjustments Based on Auxiliary Information. In *Survey Nonresponse*, ed. Robert M. Groves, Don A. Dillman, John L. Eltinge and Roderick J.A. Little. New York: Wiley p. 275–88.

Groves, Robert M. 2006. "Nonresponse rates and nonresponse bias in household surveys." *Public Opinion Quarterly* 70(5):646–675.

Manski, Charles F. 2009. *Identification for prediction and decision.* Harvard University Press.

Pearl, Judea. 2009. *Causality.* Cambridge University Press.

Wooldridge, Jeffrey M. 2007. "Inverse probability weighted estimation for general missing data problems." *Journal of econometrics* 141(2):1281–1301.