


Methodology and Research Practice

Prespecification of Structure for the Optimization of Data Collection and Analysis

Matthew J. Vowels¹  ^a

¹ Institute of Psychology, University of Lausanne, Switzerland

Keywords: Markovicity, data collection, conditional independence, causality, path modeling, structural equation modeling

<https://doi.org/10.1525/collabra.71300>

Collabra: Psychology

Vol. 9, Issue 1, 2023

Data collection and research methodology represents a critical part of the research pipeline. On the one hand, it is important that we collect data in a way that maximises the validity of what we are measuring, which may involve the use of long scales with many items. On the other hand, collecting a large number of items across multiple scales results in participant fatigue, and expensive and time consuming data collection. It is therefore important that we use the available resources optimally. In this work, we consider how the representation of a theory as a causal/structural model can help us to streamline data collection and analysis procedures by not wasting time collecting data for variables which are not causally critical for answering the research question. This not only saves time and enables us to redirect resources to attend to other variables which are more important, but also increases research transparency and the reliability of theory testing. To achieve this, we leverage structural models and the Markov conditional independency structures implicit in these models, to identify the substructures which are critical for a particular research question. To demonstrate the benefits of this streamlining we review the relevant concepts and present a number of didactic examples, including a real-world example.

Imagine you want to estimate the effect of a therapeutic treatment on depressive symptoms, and how this effect may be mediated via another variable, say, therapeutic alliance. One might suspect that these variables are linked through a complex causal web involving multiple other factors - but which of these other factors are necessary, in terms of data collection, for estimating the main effect of interest? Collecting too many variables increases the cost and time required to complete data collection, having an impact on participant fatigue (Lavrakas, 2008) as well as draining valuable project resources. Conversely, collecting too few may render the results of the statistical tests invalid. In this manuscript, we describe how to identify those variables which are strictly necessary to arrive at unbiased answers to pre-specified questions. Of course, other interests may influence data collection (such as subsequent applications and usage), but knowing what is strictly necessary allows one to make more informed decisions about what to include.

In this paper, we argue that the data collection and research project methodology can be optimized by specifying the causal structure underlying a theory in graphical form. Using rules from the structural modeling framework, one

can then use the graph to identify variables or scales which are either causally necessary or which can be omitted from the data collection process. This liberates resources to either improve the quality of the remaining scales (e.g., by using scales with a more comprehensive set of items), and/or to reduce participant fatigue by shortening the duration of a questionnaire and using these resources to increase the overall sample size. Indeed, concerns about inadequate statistical power are growing in response to the replication crisis (Aarts et al., 2015; Baker et al., 2020; Correll et al., 2020; Sassenberg & Ditrich, 2019), and researchers are thus encouraged to make sure they have sufficient data to estimate the effects of interest.

Furthermore, even if a researcher decides not to undertake any analyses (perhaps they are not able to collect data, for whatever reason) the process of reflecting a theory graphically nonetheless helps with transparency, reproducibility, and the meaningfulness of subsequent interpretation. Psychology has been accused of being 'not even wrong' (Scheel, 2022) on the basis that the theories are too vague to be adequately tested. By reflecting our theories in a graphical form, we thus improve the clarity and reduce the one-to-many relationship between our theories and our

^a Correspondence concerning this article should be addressed to Matthew J. Vowels, Institute of Psychology, University of Lausanne, Switzerland. E-mail: matthew.vowels@unil.ch

statistical models. Translating our theories to graphs also forces researchers to think carefully about the underlying process, and the concomitant implications for data collection. The specification can then be made explicit, preregistered (Nosek et al., 2018), and compared unambiguously against other work. This, in turn, facilitates more precise replication by subsequent researchers, as well as a clearer understanding of the relationships between the hypotheses being tested and the assumptions and theory which underpin the model specification and results (Grosz et al., 2020; Haslbeck et al., 2021; Navarro, 2021).

In this work we show how four related concepts - conditional independencies, Markov Blankets, projection, and causal identification - can be used to judiciously shrink the number of variables required to answer a research question, without impacting downstream analyses and without impacting the congruity of the model with the underlying theory. The process is not data-driven and is not the same as seeking model 'parsimony' - our approach does not fundamentally change the complexity of the underlying processes reflected by the 'full' model. Instead, using a set of rules which are consistent with the assumptions of the original graph being specified, our initial graphical representation can be reduced to focus in on the effects we really care about. Thus whilst the complexity of the statistical model reduces, it does so without introducing any additional simplifying assumptions beyond those which already existed in the original theory.

The techniques are relevant to a broad range of problems amenable to specification in graphical form. For example, the didactic examples given by Rohrer (2018) involve health problems and work satisfaction, genetics and child's depressiveness, or educational attainment and income. Additionally, social psychologists interested in complex, mediated processes and multiple baseline control variables could also benefit from the proposal presented here. To this end, as well as providing a set of experimental results to demonstrate the performance characteristics in a general and non-domain-specific way, we also provide an example application to a graph used in organizational behavior (Spurk & Abele, 2010). Our hope is that researchers can use the techniques presented in this work to optimize their data collection and analysis in a more transparent way which is tailored specifically to the particular relationships of interest.

We begin by motivating the specification of our theories in graphical form. Then, we introduce the relevant statistical/structural concepts needed to understand the process for reducing this model. We then walk through a number of didactic examples, comparing an assumed 'real-world' or Data Generating Process (DGP) against the minimal required model for estimating a set of causal effects of interest. We also provide the associated multiple linear regression models where a single regression model can be used to provide the same information, and present a real-world example. In supplementary, we also provide simulation results to demonstrate that the approach does not introduce bias, and in some cases can improve model fit and reduce standard error. Finally, in the supplementary we

also provide the code for an automatic tool for reducing the graph (along with a description of the associated algorithm). The code for reproducing the simulations as well as the automatic tool are provided here: <https://github.com/matthewvowels1/minSEM>.

Terminology and Conceptual Overview

In this work, we assume that psychologists/researchers are principally concerned with estimating a particular causal effect (e.g., the effect of treatment on an outcome). Indeed, this goal aligns with the causal nature of psychological theories (which, in general, describe causal processes), as well as the goal to design and implement effective interventions which improve peoples' lives. As such, we assume that a researcher wishes to test a particular hypothesis which concerns a (causal) effect size of interest.

We will refer to a number of objects which deserve to be defined up-front. In [Figure 1](#) we present examples of these objects for reference. Firstly, we assume that there exists some (potentially highly complex) real-world *Data Generating Process* (DGP). According to our existing theories, we wish to model this DGP in such a way that we are able to meaningfully represent it. One option for doing so involves the use of *Structural Equation Models* (SEMs). SEM provides us with a powerful and popular (Blanca et al., 2018) statistical framework to unambiguously reflect and test causal theories and relationships (Grosz et al., 2020; Pearl, 2009; Rohrer, 2018; Vowels, 2021; Wright, 1921, 1923). In particular, the SEM can be represented in an intuitive *graphical* (and therefore visual) way, thereby specifying our domain knowledge about the DGP.

The graphical representation of the theory, which we will refer to as the graphical or structural model, can be used early on in the research pipeline to inform the data collection methodology, by helping us specify which constructs we need to measure. Furthermore, early specification of a statistical model helps us with preregistration and research transparency (Wagenmakers et al., 2012). Such transparency is increasingly important in the fields of psychology and social science, where attention has been drawn to numerous problems with theory testing, research methodology, and analytical practice (Aarts et al., 2015; Flake & Fried, 2019; Gigerenzer, 2018; Marsman et al., 2017; McShane et al., 2019; Scheel et al., in press; Vowels, 2021).

As we will discuss, we will apply the rules of a type of graphical model known as a *Directed Acyclic Graph* (DAG) to the graphical representations of our SEM. These rules are actually more general than those specific to SEM, because whilst SEM assumes linear relationships between variables, the rules we use are applicable to problems with almost arbitrarily non-linear relationships. Using these rules, and in combination with a *Research Question* expressed as a set of target causal effects of interest, we can reduce its complexity (which we refer to as the *Reduced SEM*) without sacrificing our ability to estimate what we care about for a particular research question or hypothesis. This reduced model then determines which variables we are required to collect data for. In some cases, we may not need to use the typical

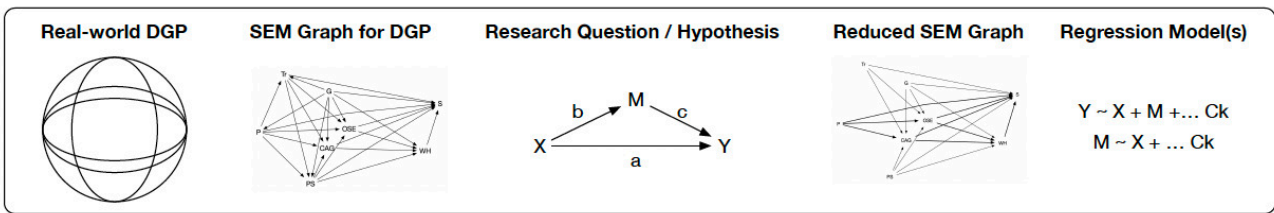


Figure 1. Top level terminology.

Note. We assume (left) there exists a real-world causal Data Generating Process (DGP), which we wish to model using a structural model. This structural model can be represented graphically (see SEM graph for the DGP in the figure). Using our proposed approach, this SEM can be simplified in such a way that does not jeopardise the estimation of a particular (causal) effect size which is of interest to our research. For example, we may be interested in estimating path coefficients/effects a , b , and c in a mediation model. Finally, the effect sizes may be estimated using straightforward regression models.

SEM estimation techniques to answer our research questions, and a simple multiple regression model may suffice. However, it is worth emphasising that this work is not concerned with the estimation of the coefficients themselves, but rather how we can use the graphical modeling rules to simplify the representation of a theory, and in turn streamline our data collection and study design.

Motivation

In this section we provide two principal motivations for our proposed approach: Statistical power, and model under- or mis-specification. In light of these motivations, we then provide a top-level overview of our proposal.

Statistical Power and Model Specification

Psychological research is frequently *underpowered* (Crutzen & Peters, 2017; Maxwell, 2004; Vankov et al., 2014), and the theory and analysis are often poorly specified (Grosz et al., 2020; Rohrer, 2018; Scheel, 2022; Scheel et al., in press; Vowels, 2021). The studies are underpowered to the extent that the sample sizes are insufficient to test a target hypothesis. For example, for a minimum assumed *true* effect size of interest, it is generally recommended that enough data are collected to yield a power of 80%, meaning that there is an 80% probability that we will find a statistically significant result (at a given threshold such as 0.05) (Gelman et al., 2021). Researchers are thus encouraged to ensure that their studies are adequately powered, and have been encouraged to do so for some time (Sedlmeier & Gigerenzer, 1989; Vankov et al., 2014). However, depending on the complexity of the theory under test, researchers may need to measure a large number of constructs, each with a large number of items. For example, depending on the format, the IPIP-NEO Big 5 inventory contains between 120-300 items (Goldberg, 1999; Goldberg et al., 2006) and therefore takes considerable time to complete. Besides the associated cost and time required to measure constructs using such comprehensive scales, the participants may also experience fatigue, lowering the quality of the responses (Lavrakas, 2008).

The second problem of under-specification has prompted meta-researchers to describe research in psychology as ‘not even wrong’ (Scheel, 2022). That is to say, if the theories are too vague to be specified unambiguously,

then it is not clear what it is that any particular statistical test is actually testing. If we are considered with understanding the real-time process of dyadic support, for instance, we might need to develop a statistical model which can capture the intricacies of back-and-forth, multi-modal (verbal, para-verbal, non-verbal) interactions between partners. Without unambiguously reflecting the complexity of the process in our statistical model, it is not clear what a typical model in psychology (*e.g.*, a multiple linear regression model) is really doing for us. The structural representation of this process can be a helpful aid to understand (a) what data we need to collect, and (b) whether the data can even be collected in principle (the acquisition of real-time, multi-modal data may in some cases be infeasible).

Furthermore, a single theory may admit multiple statistical models, each of which tests something slightly different but all of which are valid given the malleability of the underlying theory. Few psychological theories make it clear which variables are necessary to include as control variables, for instance. And yet, the inclusion of different control variables can have a large impact on the resulting parameter estimates, and it is not usually clear how these control variables are chosen or how they relate to the tested theory (Cinelli et al., 2020; Hullman et al., 2022; Vowels, 2021). As an example, in medical studies older patients may be more likely to choose medication over surgery, but also be less likely to recover. This makes age a key confounder that must be controlled/adjusted for to evaluate the treatment effects. However, perhaps there exist other, less obvious confounders which we have not collected and which we can therefore not adjust for. Some variables may need to be controlled for but be unattainable, some may be inconsequential (and can be omitted without consequence), and still others may actually be detrimentally biasing the model. In order to determine which control variables should or should not be included, and to therefore avoid what is known as structural misspecification (Vowels, 2021), researchers need to somehow formalise their theories.

The Proposed Solution

With respect to statistical power, there exists a need for compromise - maximising the quality of a survey such that it measures all that we need, at a sufficient level of quality,

for a sufficient number of participants. Of course, we acknowledge that there often exist multiple goals for studies in which new data will be collected - they may have either confirmatory or exploratory research questions, or both; they may wish to compare and contrast multiple competing hypothesized structures; they may want to 'future-proof' the study, such that additional variables are collected with a view that they may be necessary for answering research questions which are not yet specified.

At the same time, and in order to correctly specify a model with respect to a psychological theory, it is important that psychologists consider not only the structure between the primary constructs central to their theory, but also the full data-generating process (DGP) which leads to a set of observations. The theory can then be translated into a graphical/structural model which reflects this DGP, which we can use to make sure we are not missing variables which are key to answering a particular research question. The process of deriving a structural model from our theory has been previously discussed by Rohrer (2018) and others (Kline, 2005; Loehlin & Beaujean, 2017), and we do not describe the procedure in this work, but note that the graphical framework (more about this in later sections) makes the process quite intuitive.

The advantages of reflecting the theory unambiguously in a structural model include reproducibility (it is clear what exactly is being tested) and an increase in the interpretability and validity of the resulting effect sizes. Rather than the effect sizes being arbitrary consequences of *ad hoc* models loosely connected to theory, they reflect specific causal effects within a fully specified structural/causal process. Whilst the causal validity of effect sizes estimated using these models still depends on whether a number of strong assumptions hold (e.g., whether the hypothesized structure is correctly specified with respect to the actual, real-world structure), the transparent specification of the model makes subsequent criticisms and revisions more precise. The task of translating our theories may also highlight possible weaknesses in the theory, or call attention to possibly insurmountable difficulties for data collection. For instance, theories which involve dynamic processes that unfold at irregular intervals over time may require very specific, expensive, and challenging data collection procedures (Hilpert et al., 2019). Identifying the specifics of such challenges in advance could save a lot of wasted time and effort.

Unfortunately, the task of identifying all relevant variables will likely implicate a large number of secondary variables (such as demographics and other theoretically related constructs), and thus require longer questionnaires. The problems of statistical power, comprehensive scale inventories, and the need to collect a broad range of variables and constructs relevant to our theory puts a lot of pressure on researchers to find a suitable 'Goldilocks' design, and one or multiple methodological facets are likely to be compromised as a consequence. As such, after the specification of the full DGP, we should examine the resulting model to identify possible shortcuts in the data collection process. Indeed, and as we will show, even if a variable or construct

is relevant to a particular causal process, it may not be required for the actual analysis. To know this, however, the variable needs to be transparently situated in a causal model for us to understand whether it is essential for answering a target research question, or not.

Once the structure of the DGP is fully specified, and as we will describe in detail below, we are able to identify essential substructures which are sufficient for testing our intended hypotheses. The substructures, by definition, exclude certain variables. Thus, if we can identify these substructures in advance of data collection, we may be able to significantly reduce the number of constructs we need to measure. Indeed, in example 2i in [Figure 4](#) below, we show that it is possible to reduce the number of variables/constructs by two thirds, although this depends on how much of the causal process we are interested in testing. It goes without saying that any simplification must be done carefully. Indeed, the potential consequences of any resultant model misspecification can be severe, and includes heavily biased parameter estimates which are almost impossible to meaningfully interpret (Hullman et al., 2022; Vowels, 2021). However, there are no requirements for researchers to 'go all the way' with the simplification, and the proposal is flexible insofar as the degree of desired reduction can be determined by the researcher and their specific requirements.

We thus advocate that researchers consider the DGP upfront, before the data collection stage. Such prespecification in the form of a structural (or, as we will present, graphical) model represents a beneficial step in terms of preregistration and transparency, helps researchers distill their theories into testable models, thereby increasing the validity and meaningfulness of downstream statistical inference and results interpretations, and provides us with an opportunity to 'prune' the structure to optimize for statistical power during data collection.

Background

In this section, we introduce a number of relevant technical concepts for reducing our structural models. In general, we assume that the model is being specified in graphical form as a path model, or a Structural Equation Model (SEM), where directed paths/arrows correspond with causal links. As we mention above, the techniques we use are more general than the SEM framework, and come from the graphical models literature. A number of existing resources discuss the implications of changes in causal structures on statistical estimation. For example, Matthew J. Vowels (2021) discusses the problems that arise due to misspecification of causal models, and notes the potential to focus on specific effects within a causal process; and Cinelli, Forney, and Pearl (2020) provide a laconic summary of how to choose control variables such that the choice does not induce bias in our parameter estimation. Unfortunately, these resources do not discuss the possibility of reducing our SEMs to the most simple model which can still yield unbiased estimates of (possibly multiple) causal effects.

To best communicate our approach, we begin with a brief review of the relevant background. We aim to review four

related concepts in particular: causal identification, conditional independence, Markov Blankets, and projection. Briefly, identification is the goal of isolating causal from non-causal statistical dependencies, and, when possible, facilitates the estimation of causal effects. It relies on conditional independencies, which describe how statistical dependencies arise due to the underlying causal process, and how conditioning on these variables enables us to isolate or disentangle different sources of dependence. Markov blankets show that, through the use of conditional independencies, we can completely isolate an entire substructure in a graph, thereby making it clear that not all variables are necessarily required for a particular research question. Finally, projection enables us to combine/reduce the number of paths. This is particularly true in the case of mediation, where a mediator can be excluded entirely if the researcher is not interested in estimating the mediation *per se*.

Interested readers are encouraged to consult useful resources by Hünermund and Bareinboim (2021; Cinelli et al., 2020; Kline, 2005; Koller & Friedman, 2009; Loehlin & Beaujean, 2017; Pearl, 2009; Pearl et al., 2016; Peters et al., 2017; Vowels et al., 2022). In terms of notation, we use X (or, e.g. A, B, C etc.) to denote a random variable, and bold font \mathbf{X} (or, e.g. $\mathbf{A}, \mathbf{B}, \mathbf{C}$ etc.) to denote a set of random variables. We use the symbols $\perp\!\!\!\perp$ and $\not\perp\!\!\!\perp$ to denote statistical independence and statistical dependence, respectively. For linear systems, such statistical dependence may be identified using correlation, but the majority of our discussions are general and non-parametric. We use directed arrows to denote a directional structural/causal dependence, and U (or \mathbf{U}) for a single (or set of) unobserved variable(s).¹

For example, in SCM terminology $A := f(B, C, U_A)$ indicates that A is some general function f of B and C . Here, U_A tells us that A is also a function of exogenous random process U_A . Indeed, it is this U_A which prevents the relationship between A and B and C from being deterministic. Structural Equation Models (SEMs), on the other hand, assume that all endogenous variables are the result of a linear weighted sum of others, such that $A := \beta_{BA}B + \beta_{CA}C + U_A$. Here, the β s are structural parameters (also called path coefficients or effect sizes) which we wish to estimate. The walrus-shaped assignment operator $:=$ tells us that the left hand side is a structural outcome of the right hand side; the equations are not intended to be rearranged and there is very much a directional relationship involved.

As we construct system of equations representing our SEM (or, indeed, our SCM) it is often convenient to represent these relationships graphically/visually. For example, consider the following set of (linear) structural equations:

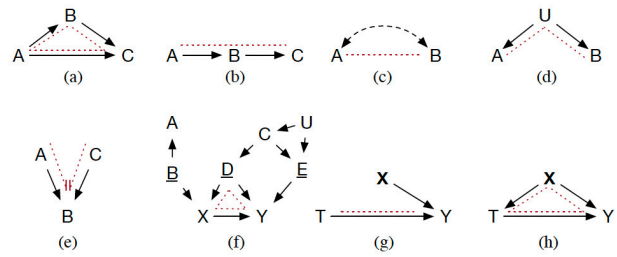


Figure 2. A set of demonstrative graphs.

Note. This figure provide a number of example graphical models. Solid black lines indicate causal dependencies, dashed red lines indicate statistical dependence, parallel red bars indicate a 'break' in statistical dependence (example (e)), **boldfont** indicates a set of variables, and the letter U is reserved to denote unobserved variables.

$$\begin{aligned} A &:= U_A \\ B &:= \beta_{AB}A + U_B \\ C &:= \beta_{AC}A + \beta_{BC}B + U_C. \end{aligned} \tag{1}$$

These can be represented simply as the mediation model depicted in black, solid arrows in Figure 2(a). The variables U are generally not included unless they are statistically dependent. Of course, they frequently *are* dependent in psychology, and this may be denoted using a curved, bidirected edge, as between variables A and B in Figure 2(c), or by explicitly including the relationship as in Figure 2(d). Such relationships can, of course, also be included in the system of equations comprising the SEM. Note that, as a result of the causal structures present in the DGP, there are induced a number of statistical dependencies indicated in Figure 2 by the red dashed lines. By induced statistical dependency, we mean that the variables are correlated, or, more generally, statistically dependent, by consequence of the causal relationships between the variables in the underlying causal process.

The Data Generating Process

It is worth maintaining conceptual separation between: (1) the process occurring in the real world, which we consider to be the true Data Generating Process (DGP), (2) Our SEM, which we generally want to sufficiently capture the process in the real world, and (3) the specification of a multiple linear regression. Note that (1) and (2) do not have to match precisely. Indeed, when we create our SEM we expect it to be a significant simplification of the real-world process, but it needs to be somewhat *consistent* with the true process (and the degree to which this is achieved is one of the primary aims of our research). If it is not sufficiently consistent, we might deem it to be *misspecified*, and it will not yield meaningful statistical estimates.

¹ Note that the theory we discuss is applicable to models with latent constructs (such as factor or measurement models), as well as those without (such as path and structural models), and generalises beyond linear models. The theory we discuss is part of the general Structural Causal Modeling (SCM) and Directed Acyclic Graph (DAG) frameworks (Pearl, 2009). Path models and SEMs both represent a subset of the family of SCM and DAG models, where the *functional* relationships between variables are assumed to be linear. In other words SCMs and DAGs make no assumptions about whether one variable is an arbitrarily complex function of another (strictly, there are exceptions to this, as discussed by Maclaren & Nicholson, 2020).

For example, if we have a strong theory that the true DGP can be adequately represented by a fully mediated process $A \rightarrow B \rightarrow C$, then we would be advised to employ an SEM which is consistent with this structure. By consistent we mean that the model we use facilitates the unbiased estimation of the parameters of interest, and that these estimated parameters correspond with something meaningful in the real-world (e.g., causal effects sizes).² One option we have is to specify everything about our theory explicitly using an SEM, and this can be done in graphical form to aid formalisation. However, what we aim to show is that if we are primarily concerned with a subset of parameters (*vis a vis* all path coefficients in the model), then in some cases we can significantly reduce the complexity of our model without affecting the consistency of our resulting model. In the case of the *full* mediation, it is interesting to note, for example, that including a direct path in the SEM (in addition to the indirect effect) does not bias our estimates of the indirect path parameters. This is because the direct path will have an estimated effect of zero if it does not exist in the real-world, and its inclusion does not influence the value of the coefficient estimated for the indirect path. This is an example of how *increasing* the complexity of the SEM does not necessarily result in ‘disagreement’ or misspecification with respect to the SEM and the real-world DGP. In contrast, failing to include a direct path which *does* exist in the real-world DGP, can affect the resulting path estimates. As such, in some cases assumptions which simplify the graph can be more ‘dangerous’ than those which increase the complexity of the graph, and it is especially important any simplification be done with care to avoid biasing the estimates of the remaining path coefficients.

Finally, note that the effect sizes of interest in the final SEM can be estimated using multiple regression. Indeed, the specification of an SEM using the popular *lavaan* R library (Rosseel, 2012) follows a very similar syntax to that used to estimate each path using the *lm* regression library. Note that this may not always be possible, particularly if one needs to estimate latent factors. However, we provide the equivalent regression syntax to highlight the equivalence between the techniques, and to show that even if a structural model is used to specify the DGP, it may be possible to use a straightforward linear regression model for the actual estimation.

Identification and Disentangling Statistical Influence

Identification concerns whether or not, for a given graph, the causal effect we are interested in is actually estimable from the observed data, even in the absence of an experiment (Huang & Valtorta, 2006; Shpitser & Pearl, 2008). In the case where the full graph is given and there are no unobserved confounders, all causal effects are tech-

nically identifiable from the data. This means that there exists a mathematical expression which expresses the causal effect(s) of interest as a function of the observed statistical associations. If a causal effect is identifiable, it may be possible to estimate it with only a fraction of all the observed variables. Furthermore, if researchers are only interested in estimating a single path coefficient in a structural model, it may not be necessary to run the full SEM estimation process, and instead researchers can run a multiple regression (possibly employing machine learning techniques) to directly estimate the effect of interest (van der Laan & Rose, 2011; Vowels et al., 2021).

In the case where researchers *are* interested in the estimation of multiple paths (for example, in a mediation model), one can choose either to undertake a series of multiple regression analyses (and we provide examples of this below), or to estimate them simultaneously using the SEM estimation framework. In both cases, however, all effects of interests must fulfil the requirements for identification. In other words, the estimation multiple causal effects (e.g., from treatment to mediator and from mediator to outcome) requires that all effects can be identified from the data, which is obviously entails more stringent requirements than does the estimation of only one of these paths.

A detailed description of how to use identification is beyond the scope of this paper, but we describe below how to isolate/disentangle statistical influence using the conditional independency properties below. For now, let us consider the case where we are interested in estimating only one path coefficient / causal effect - the rules generalize to multiple coefficients. Consider the graphs in [Figure 2\(g\) and \(h\)](#). Graph (g) represents the canonical Randomized Control Trial setup, where T represents some treatment, Y some outcome, and \mathbf{X} some set of covariates which help to explain the outcome Y . In this graph, the covariates \mathbf{X} are independent of treatment T because of the random assignment of treatment. Such a structure means the only statistical dependence that exists between the treatment and the outcome is a result of the treatment itself. This statistical dependence is thus equivalent to the causal dependence we are interested in. As such, the effect can be directly estimated by comparing the outcome under different treatments. Note that one may still wish to consider \mathbf{X} too - it can be used to explain additional variance in Y in order to tighten the estimate of the treatment effect. In other words, the inclusion of these variables may reduce the standard errors associated with a particular causal effect size estimate.

In contrast, in observational studies patients may select their own treatment, and graph [Figure 2\(h\)](#) is more appropriate. For instance, if age is one of the covariates, older patients may prefer medication and have a lower chance of recovery, whilst younger patients may prefer surgery and have a higher chance of recovery. Thus, if we wish to esti-

² For the estimation task itself, we can either use the SEM estimation framework (and estimate all the included paths), or alternatively, we can derive a set of equivalent regression equations.

mate the *causal* influence of treatment T on the outcome Y , we cannot simply compare the outcomes of the two treatment groups, but now also need to somehow adjust or ‘control’ for the additional statistical dependence that exists between Y and T which results from the ‘backdoor’ non-causal path $T \leftarrow \mathbf{X} \rightarrow Y$. This is non-causal because there is no directed path between T and Y via X (the arrow points from X to T , not the other way around). Knowing the rules of conditional independencies described below, we will be able to isolate the causal effect of interest such that the remaining statistical dependence between T and Y corresponds with the causal dependence we actually wish to estimate.

Note that we will use the term control variables to mean variables which we wish to adjust for to identify causal effects of interest, and which would otherwise leave an opening for non-causal, statistical association. For example, the set of variables \mathbf{X} in Figure 2(h) could be considered to be a set of relevant control variables which enables us to get unbiased estimation of the effect of treatment T on the outcome Y . However, it is worth considering that a set of control variables itself may comprise a complicated structure in its own right, and we consider two cases in the examples section below.

Conditional Independencies

The visual graphs provide us with a way to directly read off the conditional independency structure of the model. Conditional independencies tell us whether the inclusion of additional information changes anything about our knowledge. For instance, consider the (illustrative) fully mediated model Testosterone \rightarrow Bone Length \rightarrow Height. This model tells us that, in the absence of a direct path from Testosterone to Height, if we already know someone’s Bone Length, knowing their Testosterone in addition changes nothing about their likely height. In other words, no more of the statistical dependency between Testosterone and Height is left to explain once Bone Length is known. Equivalently, if we condition our knowledge on Bone Length, Testosterone is rendered *conditionally independent* of Height. Indeed, if a linear regression is used to estimate the effect of Testosterone on Height, but we include Bone Length as a control variable, the coefficient on Testosterone will tend towards zero. This is a useful example which highlights the importance of a consideration for structure and the associated conditional independencies - if we do not already know that the process is fully mediated, we might incorrectly arrive at the conclusion that Testosterone is unrelated to Height.

If our graph Testosterone \rightarrow Bone Length \rightarrow Height is a sufficient representation of the process in reality, and if the statistical relationships hold in the data we observe, then the graph is also said to be *Markovian* (i.e., the ‘Markov con-

dition’ holds). In fact a Markovian graph is simply a graph for which its implied conditional independencies hold in the data it is being used to model. Conversely, if there exists one or more unobserved variables which we have failed to include in our model, and which influence the statistical dependencies in our data such that the Markov condition no longer holds, the graph is said to be *semi-Markovian*. If we suspect a graph is semi-Markovian because of the presence of some unobserved confounder(s), we should do our best to update our graph and include this unobserved factor, so that the rules apply to our (now Markovian) model. If we find this unobserved variable is necessary for identification, but we simply cannot collect data for it (it might not be an easily measurable factor), then it may not be possible to estimate the causal effects of interest.³ Whether or not a causal effect of interest is identifiable is important to understand early on, because it may determine the feasibility of the study. This is another reason why a graphical specification of a theory can be useful.

We can use conditional independencies to isolate causal from non-causal statistical dependence (the task of identification described above), as well as to identify which variables we need to include or exclude in our SEM. Starting with the example in the full mediation model of Figure 2(b), we see that variable C cannot contain information about A which does not already ‘pass’ through B . Therefore, if we already know B , knowing A tells us nothing more about C than we already knew. This renders A statistically independent of C given B , which can be expressed as: $A \perp\!\!\!\perp C \mid B$. This is known as a conditional independence statement, because it tells us which sets of variables are independent of each other given a set of conditioning variables. It is worth noting that when we run a regression (logistic or otherwise) we are estimating some expected outcome *conditioned on* some set of predictors. Running the regression to estimate $\mathbb{E}[C|B, A]$ (i.e., the expected value of C , controlling for B and A) from data generated according to a fully mediated DGP will result in the same consequences as above: the fact we have included B means that the importance given to A will be zero (notwithstanding finite sample deviations). Clearly, therefore, an understanding of the structure is absolutely crucial for constructing the regression models (Vowels, 2022). For instance, if A is a treatment variable and we do not recognise B as a mediator, the inclusion of B in the model will result in a negligible coefficient estimate for A which may well mislead us to think the treatment is ineffective.

To generalise this result to other graph structures, it is worth committing some rules to memory. If a graph contains these two substructures:

$$\begin{aligned} A \rightarrow B \rightarrow C, \\ A \leftarrow B \rightarrow C, \end{aligned} \quad (2)$$

³ One might consider sensitivity analysis as a means to quantify the extent to which a causal effect can be explained by unobserved third variables (Diaz & van der Laan, 2013).

then knowing/conditioning on B renders A and C statistically independent. Of course, without this conditioning, A , B , and C are all statistically dependent. These two graphs are known, respectively, as a chain and a fork. One can start to write the complete list of conditional independencies which are implied by *both* of these two graphs is:

$$A \perp\!\!\!\perp B, A \perp\!\!\!\perp C, B \perp\!\!\!\perp C, C \perp\!\!\!\perp A|B, C \perp\!\!\!\perp B|A, B \perp\!\!\!\perp C|A. \quad (3)$$

The first, $A \perp\!\!\!\perp B$, means that A is not statistically independent of B (because A causes B), the second means that A is not statistically independent of C (because A causes C through B), and so on. Importantly, both of the graphs in Eq. 2 imply the same set of conditional independencies, and therefore there is no way to tell them apart using statistical dependencies alone.⁴ Alternatively, if a graph is structured as follows:

$$A \rightarrow B \leftarrow C, \quad (4)$$

we have what is known as a *collider*. Unlike the examples in Eq. 2, variables A and C are actually already independent such that $A \perp\!\!\!\perp C$. A collider is also depicted in Figure 2(e), and the parallel vertical red lines depict the ‘break’ in statistical dependence between A and C . Furthermore, conditioning on B in this structure actually *induces* statistical dependence between A and C - a phenomenon known as explaining away (Pearl, 2009; Pearl et al., 2016). A corresponding list of conditional independency statements for this collider is therefore:

$$A \perp\!\!\!\perp B, B \perp\!\!\!\perp C, A \parallel C, A \perp\!\!\!\perp C|B, \quad (5)$$

Variables are known as *ancestors* of downstream *descendants* if there exists a directed path between the variables. A direct descendent is also called a child, and the direct ascendant is called a parent. Note that conditioning on *descendants* of the variable B in the two graphs depicted in Eq. 2 can *partially* render A and C independent (because it essentially contains critical information from A via B). Similarly, conditioning on a descendent of the collider variable B in Eq. 4 can also render variables A and C *partially* dependent. Of course, two variables are either dependent or not, and the partial terminology is used here to communicate that the effect of conditioning is not as strong as would be the case using B itself, as opposed to one of its descendants. We can actually test for these conditional independencies using conditional independence tests (which, in the linear Gaussian setting are essentially partial correlations). These tests can then be used to *discover* the underlying structure in the data - a task known as causal discovery, for which many methods exist (Vowels et al., 2022).

Finally, returning to Figure 2(h), which was discussed above in relation to estimating the effect of treatment T on outcome Y given some confounders \mathbf{X} , we know that for the substructure $T \leftarrow \mathbf{X} \rightarrow Y$, we can achieve $Y \perp\!\!\!\perp T | \mathbf{X}$ in

order to essentially simulate the structure of the graph for the RCT in Figure 2(g). In other words, by conditioning on \mathbf{X} we ‘block the backdoor’ path of *confounding* statistical dependence which ‘flows’ from treatment to outcome by conditioning on \mathbf{X} . This leaves only the one statistical path, which is also the causal path we care about. In this case, the statistical dependence is equivalent to the causal dependence we wish to estimate. Thus, we have used conditional independency rules to isolate the causal statistical dependencies, and disentangle them from the non-causal statistical dependencies.

Markov Blanket

The conditional independency rules introduced above can be used to define a Markov Blanket. Essentially, the blanket constitutes a set of variables which yield conditional independence between variables ‘within’ the blanket, and those outside it. The notion of a Markov Blanket confirms the idea that not all variables are necessarily needed to estimate or identify a particular causal effect. The implication of this is that if we have knowledge of a set of conditioning variables, other variables which are causally ‘downstream’ of these conditioning variables become effectively ‘disconnected’ from those which are upstream.⁵

Consider Figure 2(f) which depicts a Markov blanket around variables X and Y . The underlined variables B , D , and E constitute the Markov blanket - knowing or conditioning on these variables renders X and Y independent of variables A and C , which are outside of the blanket.

An SEM model can be reduced in size to comprise only the variables and paths necessary to estimate set of paths of interest. Considering, again, Figure 2(f), if we are only interested in the path coefficients proximal to the variables X and Y , we do not need variables A or C , thus reducing the number of estimated paths from ten (if we include the paths from unobserved U) to five. We discuss more opportunities below.

Projection

A cause-effect relationship can often be broken down into smaller and smaller subdivisions, until one starts talking about the effect of one molecule on the next to explain a simple game of billiards. As per Figure 3, each subdivision of the cause-effect relationships between X and Y could be represented as a mediating path with an infinite number of intermediate mediating paths. By consequence of the Markov assumption (described above) it is thankfully not necessary to model all these intermediate mediators, and it suffices to abstract to the key ‘beginning and end points’.

4 Given that the chain and the fork are yield statistically equivalent data, it is worth considering the implications for testing for mediation structures.

5 It is possible to have variables which fall into the set of defining Markov blanket variables but which do not need to be explicitly conditioned on. This can occur, for example, in the presence of a collider structure which may already render upstream variables (which are outside of the blanket) as statistically independent of those within the blanket, without conditioning (recall that conditioning on a collider can open up an otherwise ‘closed’ path).

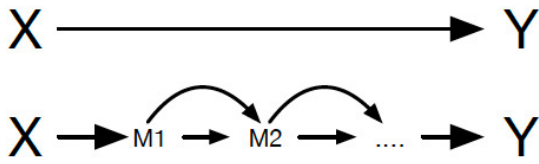


Figure 3. An illustration of ‘infinite mediation’.

Note. This figure illustrates that between any two cause-effect pairs, there exists an almost infinitely decomposable chain of intermediate mediators.

For instance, it is not necessary to know the intermediate position and velocity of a billiard ball (assuming these are well known), but it may be important to know when/if it changes course following a collision. One can, for example, reduce $X \rightarrow M \rightarrow Y$ simply to $X \rightarrow Y$ (Glymour, 2001, p. 40). Of course, if one is specifically interested in a mediating variable then one can collect the relevant data and explore the process (such examples are provided below). Some reductions may yield an intractably blunt abstraction, or, in the extreme, a form of infinite causal regress (e.g. regressing all first causes to our birth or the beginning of time), and one might instead consider more modest examples, such as whether a treatment is mediated by some psychological mechanism(s). In this case, one can nonetheless reduce the problem (via projection) to an estimation of the total effect of treatment on the outcome, thus aggregating the intermediate direct and indirect effects and thereby reducing the complexity of the graphical representation.

Reducing SEMs - Worked Examples

In the previous section we reviewed four concepts which we will use for simplifying our SEMs without introducing bias into our effect estimates: (1) causal identification, (2) conditional independencies, (3) Markov Blankets, and (4) projection. In order to demonstrate these various techniques, we will walk through a number of examples which are presented in Figure 4. For each example, we specify (a) a full DGP as our starting point which we assume to be true and complete (‘Full DGP’ in Figure 4), (b) a set of causal effects of interest, that must be identifiable for subsequent estimation (‘Research Question’ in Figure 4), (c) a minimal SEM (denoted Reduced in Figure 4), and (d) syntax for the R *lm()* function for a multiple regression. Five example DGPs are shown in Figure 4. Again, whilst we are not concerned with the estimation itself, note that one can choose to either use the SEM framework to estimate all the path coefficients in the resulting model, or one can undertake (possibly multiple) regressions to arrive at the same goal. In both cases, the graphical representation of the theory is what enables us to reduce the model in a way which does

not invalidate the subsequent analysis (as well as increasing transparency, helping us to think more deeply and concretely about the causal process, etc.).

In practice, the graphical representation of our DGP will be developed using domain knowledge and/or causal discovery techniques (Glymour et al., 2019; Vowels, 2021; Vowels et al., 2022). For now, we provide general examples with a view to demonstrating the ways in which the concepts reviewed above can be used to reduce our SEM. Similarly, in practice the set of paths of interest will be determined by our research questions and our hypotheses. Note that it may be possible to simplify SEMs bearing in mind other techniques which are applicable to linear models (such as instrumental variables) (Bollen, 2018), but we focus on those techniques reviewed above because they are generally applicable to a much broader family of problems. Finally, it is worth remembering that if a set of variables and paths are not needed for the SEM, then we also do not need to collect these variables to begin with, thus saving additional time and expense which could be used to, for example, collect more samples of the variables that really matter. Note that some variables may not strictly be necessary for the estimation of the effect but may nonetheless be worthwhile including. For example, proximal causes of an outcome which do not interfere with our estimation of other desired causes can be used to increase the precision/tightness of our estimates (Cinelli et al., 2020).

Unobserved variables and/or latent constructs may also be integrated into the specification of the graph. In terms of the planning, these objects can be considered in the same way as other observed variables, at least insofar as they relate to the estimation of the causal dependence we are interested in. One may find, for example, that the existence of certain unobserved variables fundamentally preclude identification (i.e., the estimation of the target effect), perhaps because they induce a backdoor/confounding path between the ‘treatment’ and the outcome. Conversely, one may find that either certain unobserved variables, or particular latent constructs are not necessary for the identification of the target effect. We later consider a number of worked examples involving unobserved variables (Examples 3 and 4).

To motivate the examples, we will attempt to describe semi-plausible DGPs for psychological processes, but note that these examples are likely to be overly simplistic, and are only intended to illustrate the process. We will discuss each of the examples in Figure 4 in turn. Finally, in the supplementary material we also provide simulation results for DGPs 2-5 in Figures A1-A3.⁶

Example 1: Mediated Treatment

Starting with the first example depicted in Figure 4, let us begin by considering what this graph could possibly represent. Variable Y could be an outcome (e.g. depressive

⁶ We omit simulations for DGP 1 because it represents a reduction of the other examples, and so including it is somewhat redundant.

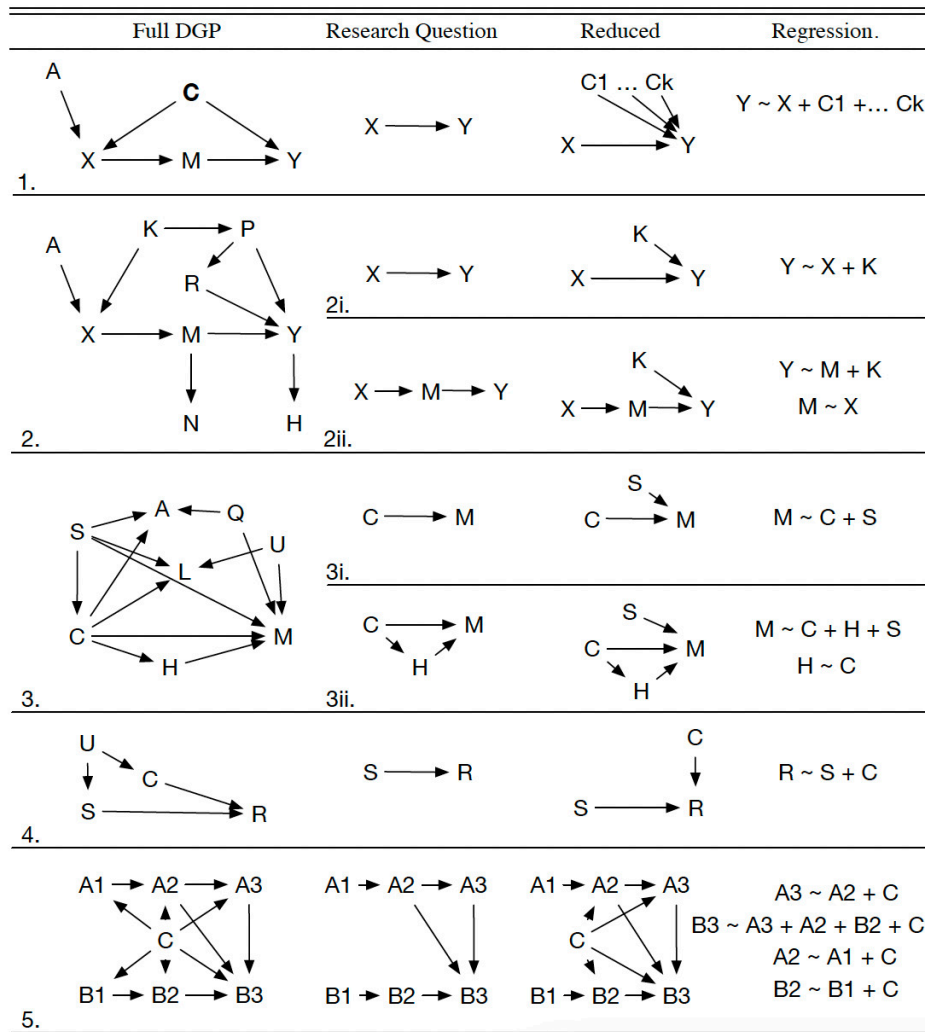


Figure 4. Finding the reduced model.

Note. This figure presents a number of examples for taking the full 'true' Data Generating Process (DGP) and finding the reduced graph and minimal linear/logistic regression required to answer a given research question.

symptoms) for a therapy X , the effect of which is mediated by therapeutic alliance M . The set C represents covariates that influence the choice of therapy modality as well as the likelihood of recovery, and includes factors such as age, gender, history of mental health problems, and so on. Finally, variable A could represent a personal attitude which influences the choice of treatment but which does not influence whether the person recovers.

For this example, let us assume that our research question concerns estimation of the efficacy of treatment on the outcome, *i.e.*, $X \rightarrow Y$. The reduced model (denoted in Figure 4 as Reduced) requires three fewer paths to estimate this effect. Firstly, if we are not interested in the particulars of the mediated path $X \rightarrow M \rightarrow Y$ then we do not need to include $X \rightarrow M \rightarrow Y$, or to therefore collected data for M (afforded by the projection concept reviewed above). Secondly, even though there exists a spurious/confounding/backdoor path $X \leftarrow C \rightarrow Y$, we do not need to estimate the actual path $X \leftarrow C$ so long as we include the path $C \rightarrow Y$. The inclusion of C facilitates identification of the principal effect of interest $X \rightarrow Y$. Note that in this case we do not

have to use SEM for the estimation procedure. Indeed, in this example we are not interested in the path coefficients linking C to Y either, even though these paths must be included to acknowledge the dependence that Y has on C and to block the backdoor path. Given we are only interested in the path from X to Y , we can simply run a multiple regression, using C as control variables and restricting interpretation to the coefficient on X . Note that the resulting *lm()* syntax contains only the two necessary components as predictors - X and the set of control variables C .

Finally, we do not need to include A in the model (neither do we need to collect data for A) because it is not necessary for the causal identification of the target causal effect of interest. Adding the path $A \rightarrow X$ into the model is superfluous to the effect we are interested in.⁷

Example 2: Structured Controls

The first graph with structured controls is given as example 2 in Figure 4. We can consider the meaning of variables A , X , M , and Y to be the same as in Example 1, that is

attitude, treatment, treatment-outcome mediator, and outcome, respectively. The difference now is that we also have a mediation child N , an outcome child H , and a structured set of control variables K, P , and R . If, as indicated in example 2i, we are only interested in estimating the effect of X on Y then, as in the first example, we can ignore A and M . Similarly, we can also exclude N and H for our reduced model, as their existence in the DGP does not change the principal relationship we are interested in.

There still exists a backdoor path through the control variables K, P, R , and Y , and so we need to understand which of the associated variables and paths to include in our reduced model to adjust for this spurious path. There exist the following options which block this path: $K \rightarrow Y$, $K \rightarrow P \rightarrow Y$, and $R \rightarrow Y \leftarrow P$. Note that $R \rightarrow Y$ is not an option by itself because this would leave the path through $P \rightarrow Y$ open. Note also that we do not need to estimate the path $K \rightarrow X$ because we are not interested in this effect. Thus, overall, our initial/complete model reduces to the estimation of only two paths (reduced from ten), as in the previous example. The linear regression also remains equivalent.

If our research question involved the estimation of the mediation, as in example 2ii in Figure 4, then the only change to the model needs to be the inclusion of the mediation $X \rightarrow M \rightarrow Y$. The linear regression now involves two stages to decompose the problem into two sets of paths (one from $X \rightarrow M$, and the other comprising the paths $M \rightarrow Y$ and $K \rightarrow Y$).

Example 3: Colliding Controls

One might be forgiven for thinking that the safest thing to do with a set of control variables is to always include them in the model to make sure we are blocking the backdoor paths. In the previous example, for instance, we could just play it safe by including $\{K, P, R\}$. However, example 3 in Figure 4 shows that some putative control variables may include collider structures. Let us consider that variables C, M , and L are class-size, math exam score, and language exam score, respectively. H represents a mediator such as whether a student does their homework, S represents Social Economic Status (SES) - perhaps children with higher SES attend schools with smaller class sizes and have better grades overall - U represents an unobserved attribute of intelligence Q a measured attribute of intelligence, and A musical ability.

Based on example 3i we are interested in the effect of class size on math exam score. It might be tempting to include the paths concerning the other related scores (such as language score, or musical ability). In the case of musical ability, we *could* include the paths $C \rightarrow A \leftarrow Q \rightarrow M$ without causing any problems, but it doesn't actually help us estimate the effect we are interested in. Indeed, the collider structure $C \rightarrow A \leftarrow Q$ prevents any backdoor information

affecting our estimation of $C \rightarrow M$, so we do not need these paths for causal identification. Another collider exists between $C \rightarrow L \leftarrow U \rightarrow M$, and even though the structure is the same, the fact that U is unobserved means we cannot and should not include L in the model. Indeed, if L were to be included (without U as U is unobserved) we would induce a spurious path linking C to M through L and U . Thus, whilst these might appear to be tempting control variables which we might think would, at best increase precision and at worst do nothing, in fact they should not be included owing to the collider structure with an unobserved variable.

We have no need to include paths relevant to A or L in our model. Including the path $Q \rightarrow M$ may improve the precision of our desired estimate, but it is not necessary. The partial mediation through H , if not part of our research question, does also not need to be included. The only path we have to be concerned about is $C \leftarrow S \rightarrow M$, and we can deal with the induced statistical path by simply including the path $S \rightarrow M$. In this case, the reduced model contains two paths, whilst the full model (including the unobserved paths) involves thirteen. The corresponding linear regression is equally simple, and only includes C and S as predictors.

If we are interested in the partial mediation of class size, homework, and math exam score, then we can simply augment the reduced model from example 3i to include this additional structure. The linear regression also changes to accommodate the estimation of the additional paths, as with example 2ii.

Example 4: Simple Unobserved Confounding

The fourth example is relatively straightforward. Here, R, S , and C could represent relationship satisfaction, partner support, and communication style, respectively, where the unobserved confounder U between support and communication. The unobserved confounder induces a non-causal statistical dependence between S and R through C , and the reduced model therefore needs to include the path $C \rightarrow R$. The linear regression, similarly, needs only S and C as predictors.

Example 5: Longitudinal Dyadic Effects

The final example concerns a longitudinal dyadic process, whereby variables for *e.g.* relationship satisfaction for two individuals A and B are collected at three time-points, but there exist intermediate opportunities where confounding could occur. This confounding could represent, for example, shared stressful events. The target causal effects involve all of the 'actor effects' (that is, autocorrelation in each individual's variables which results in similar values across consecutive timepoints), as well as two partner effects from $A_2 \rightarrow B_3$ and a 'concurrent' effect

⁷ Indeed, its inclusion can even increase the standard errors on the effect of $X \rightarrow Y$ because it makes it 'harder' to disentangle the variance in Y that stems from X and the residual variation of A which is also contained in Y .

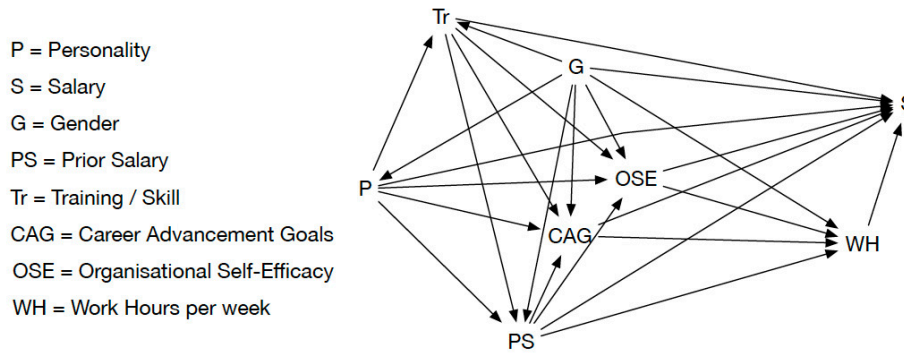


Figure 5. Real-world example graph.

Note. Real-world example graph adapted from Spurk and Abele (2010).

$A_3 \rightarrow B_3$.⁸ This example demonstrates when the use of SEM may be less complicated than undertaking a series of multiple regression tasks; our research question concerns the estimation of six separate causal effects, all of which have to be identified.

We do not need to estimate the paths $C \rightarrow A_1$, so long as we include the path $C \rightarrow A_2$, which enables us to block the backdoor path from A_1 to A_2 via C and thereby identify the effect $A_1 \rightarrow A_2$. For the same reasons, we do not need to estimate the path $C \rightarrow B_1$. In this example, we are not able to make any data collection savings (*i.e.*, we need to collect all variables), even though some of the path coefficients are not needed for estimation of the principal causal effects of interest.

Real World Example

To motivate the application of the techniques to non-synthetic examples, we have chosen a graph adapted from a paper published in the domain of business psychology and organizational behaviour. The graph is shown in Figure 5, and was presented to test the relationship between personality ('P' in the graph), and salary ('S'). First, let us consider the model required in the case where our research question solely concerns $P \rightarrow S$. The only non-causal path from personality to salary, assuming the graph shown in Figure 5, is via gender: $P \leftarrow G \rightarrow S$. The reduced graph is shown in Figure 6i. In this case, the simple regression $S \sim P + G$ would suffice, and the graphical representation of the SEM would be $P \rightarrow S \leftarrow G$. Once again, it is only possible to confirm this if we already have a representation of our model which enables us to identify the required control variables.

In the original work (Spurk & Abele, 2010), the researchers were specifically interested in a double-media-

tion by occupational self-efficacy ('OSE') and career advancement goals ('CAG'), which represent the first set of mediating variables, and working hours ('WH') which represents a second mediation of the effect of personality on salary. In this case, all variables are required for the analysis, and no savings can be made at the data collection stage, but we can nonetheless reduce the number of paths to be estimated. The reduced graph is shown in Figure 6ii. Identifying this reduced solution by eye is already becoming challenging, and automated tools (such as the one provided in supplementary material) are helpful in ensuring the reduction is correct. In addition, identifying the set of multiple regressions which can yield unbiased estimates of each of the target paths is also quite involved, and this example demonstrates how the SEM estimation framework might provide a more convenient alternative. In any case, it can be seen that six out of a total of 24 paths were not required.

Discussion

We have provided a number of didactic examples showing that if we are presented with a specific question regarding a relatively complex process, we can simplify our SEMs considerably. The simplification process takes advantage of a number of graphical rules, and does not introduce any additional assumptions to those which already apply to the full model. Furthermore, researchers are also free to choose whether they actually wish to estimate all the path coefficients using SEM framework itself, or whether a multiple regression would be more straightforward. Indeed, in cases where only a single causal effect needs to be estimated, one might consider using the graphical representation first, and then estimating it using a multiple regression instead. In this work we provided both the graphical representation of the SEM that one needs to estimate in order to answer a re-

⁸ Even though the causal framework does not strictly admit simultaneity (there must be some time delay between the cause and the effect), we assume that this concurrence is permitted according to the data collection procedure (*i.e.*, within wave three, partner A can influence partner B with some arbitrary time delay which is not distorted by the otherwise cross-sectional nature of the data collection methodology).

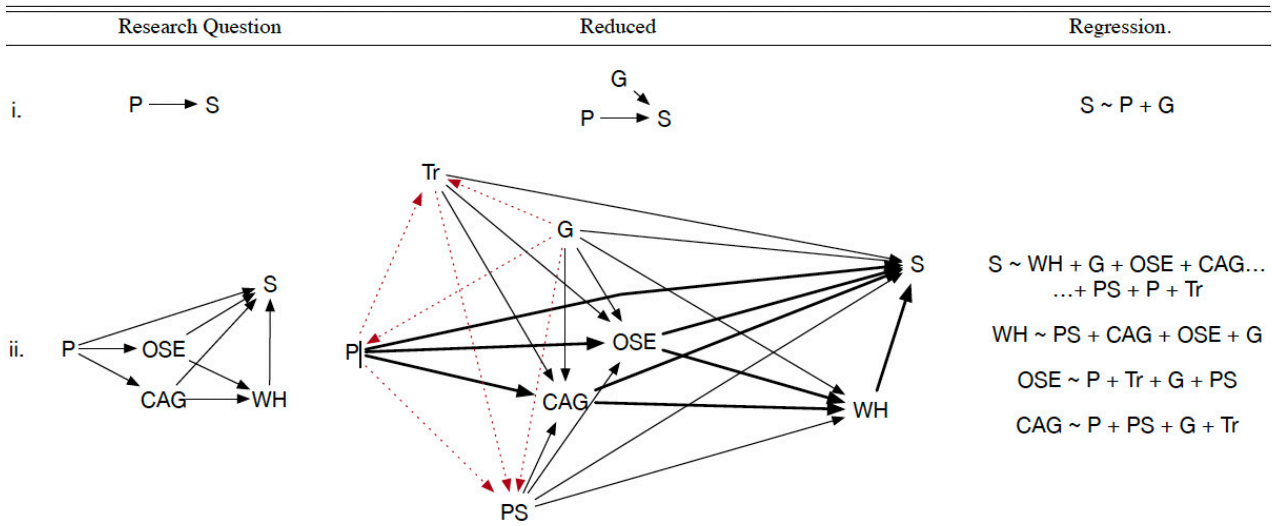


Figure 6. Reduced real-world example graph.

Note. Reduced real-world example graphs for the real-world DGP assumed in Figure 5. Bold black lines are those key to a multiple-mediation research question, whereas red dashed lines are those that may be excluded from a graphically specified SEM without affecting the estimation of the target paths.

search question relating to one or more causal effects, as well as the equivalent multiple regression equation(s).

In one of the demonstrative examples, an SEM with upwards of thirteen paths was reduced to only two. The simulation results provided and discussed in the supplementary material highlight unsurprising improvements in adjusted model fit metrics (unsurprising because simpler models are penalised less than complex models according to such metrics). Importantly, note that the simplification process does not bias the effect size estimates.

Even without the simplification process, translating a psychological theory into a graph is a worthy exercise, particularly when undertaken *before* the data collection stage. It helps us be transparent and unambiguous about our model and assumptions, increases specificity for preregistration, and can highlight potential methodological challenges and difficulties before any resources have been expended. It may even highlight cases where estimation is not possible, and this relates to the problem of causal identification. For example, if there exists an unobserved confounder between X and Y in the graph $X \rightarrow Y$, *i.e.* $X \leftarrow U \rightarrow Y$, the causal effect cannot be estimated because the non-causal statistical association induced by the confounder cannot be adjusted for without access to U . These problems can, again, be seen by an inspection of the graph, and it is worthwhile identifying these problems sooner rather than later. In practice, such problems may be common, and either a researcher must do all they can to account for the possible unobserved confounders, or they must assume that a sufficient number have been already collected to assume that the problem is ‘ignorable’ (Pearl, 2009). In general, it is important to remember that the goal of estimating causal effects rests on a number of strong (and often untestable) assumptions. However, it is only by taking causality seriously that we can understand what these assumptions are and whether they are reasonable.

Limitations

We have used SEM throughout the text because researchers in psychology may be familiar with this framework (Blanca et al., 2018). Furthermore, if they wish/need to estimate latent variables, the SEM framework readily facilitates this. Note, however, that SEM is generally considered to be an estimation framework, rather than a means to graphically represent one’s causal theory. Furthermore, SEM usually assumes linear (or at least pre-specified) functional relationships between variables. Fortunately, and as we briefly discussed earlier, all the rules and techniques discussed in this work belong to a broader class of graphical model known as Directed Acyclic Graphs (DAGs). DAGs do not make assumptions about the parametric (*e.g.*, normally distributed vs. non-parametric) form of the variables, nor about the functional (linear vs. non-linear) form relating variables. This means that when one uses our proposed method to construct and subsequently simplify a graphical structure, they can also consider themselves to be working directly with a DAG. If the researcher then wishes to avoid making assumptions about the functions and distributions, they do not have to use the SEM framework to do the estimation, but can instead use non-parametric regression or machine learning techniques (a discussion about which is beyond the scope of this paper). Indeed, another reason that we provide the multiple regression syntax is because its specification can be generalized relatively straightforwardly to non-parametric settings. For example, the specification of the regression $Y \sim X + C$ relates to the estimation of $\mathbb{E}[Y|X, C]$, which is the conditional expectation of Y given X and C . The conditioning set given on the right hand side of the tilde in the regression syntax, or the right hand side of the conditional expectation, are the variables/predictors in the regression which are being used to identify the causal effect(s) of interest, and this can be done in

both linear parametric as well as non-linear, non-parametric settings.

The reduction which is achievable depends on the research questions being asked, as well as the requirements of the researcher. We foresee that some researchers may wish to collect more variables than are strictly required for identification to future-proof their datasets, thereby facilitating the testing of currently unspecified hypotheses. The collection of extra variables can not only provide the opportunity for researchers to answer potentially unforeseen research questions, but it also enables researchers to include 'hedge' variables, in cases where the theory specification is uncertain and researchers do not want to risk variable omission. Indeed, if the researcher is contending with multiple hypothesized graph structures, they may wish to avoid putting all their eggs in one basket by collecting only the smallest set of variables relevant for one particular graph and one particular research question. By 'over-collecting' variables, they may also open up opportunities to undertake causal discovery - a data driven approach to the validation of putative causal structures. Without the extra variables, researchers would be somewhat stuck with what they have.

Finally, researchers should be mindful that the success of the approach rests on the degree of correct specification achieved when the DGP model is constructed. However, this limitation applies to *all* statistical approaches which concern the estimation of interpretable / causal effects, and this approach does not alleviate the consequences of model misspecification. Furthermore, reducing model complexity may reduce the precision of the estimation because less explanatory power may be available to estimate an effect. This is evidenced by a review of the simulation results for the p -values in the supplementary material. This downside is somewhat offset by the possibility that, with a simpler model, a larger sample size may be acquired for equivalent cost. For example, if the simplification process indicates that a number of constructs with large inventories are no longer required, we may gain back significant data collection time which can be put towards the recruitment of more participants. Such possibilities therefore enable us to increase statistical power for estimating the effects we really care about. Furthermore, the specification of larger models increases the chances of misspecification (simply put, in the specification of larger graphical models, there is more opportunity for error). Reducing the model and being specific and less ambitious about the number of primary effect sizes of interest (as opposed to wishing to estimate as many effects as possible) increases the likelihood that, at the end of the project, we have estimated something meaningful.

Related Options

It is worth noting that other approaches for streamlining data collection and reducing study cost, such as the tools for the development of short-form scale design (Greer & Liu, 2016; Smith et al., 2012) and planned missingness design (Wood et al., 2018). In the case of the former, researchers can use statistical techniques to identify reduced scale designs which provide similar performance in terms of certain scale quality measures, such as validity. In the case of the latter, there are a number of planned missingness techniques which enable researchers to amortize data collection cost over the course of a longitudinal design, or to leverage statistical associations to compensate for foreseen missing data. These methods differ significantly from our proposal, and can even be used in combination with ours. Specifically, the short-form scale design approaches are motivated by the fact that there may exist redundant information in a scale which is already represented by other items (or combinations, thereof). In contrast, our approach is concerned with the assumptions about and formal specification of the causal structure of data generating process itself, and does not concern redundancies in the scales used to measure the constructs/variables within this structure. The data generating process can therefore be considered independently of scale-item redundancy. Similarly, planned missingness techniques include split form designs (Raghuathan & Grizzle, 1995) which split large questionnaires into multiple smaller blocks, each of which is completed by participants at different stages of a longitudinal design. Alternatively, multiple imputation provides researchers with a way to leverage statistical associations to compensate for instances of missing data. Again, in contrast with our proposal, this approach does not consider the opportunities already implicit in the specification of our theory.

Conclusion

In summary, graphical representations of our theories provide us with an opportunity to encode our domain knowledge about a particular phenomenon of interest. In this paper we showed that, by using graphical modeling rules (in particular, the concept of conditional independencies), we can significantly shrink the required causal structural model without affecting the validity of the associated estimates, thereby reducing the required sample size and enabling us to redirect resources and funds towards the collection of variables which are critical to answering the questions we care about.

Submitted: September 23, 2022 PST, Accepted: January 20, 2023 PST



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

References

- Aarts, A. A. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943–950. <https://doi.org/10.1126/science.aac4716>
- Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2020). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*.
- Blanca, M. J., Alarcón, R., & Bono, R. (2018). Current practices in data analysis procedures in psychology: What has changed? *Frontiers in Psychology*, *9*. <https://doi.org/10.3389/fpsyg.2018.02558>
- Bollen, K. A. (2018). Model implied instrumental variables (MIIVs): An alternative orientation to structural equation modeling. *Multivariate Behavioral Research*, *54*(1), 31–46. <https://doi.org/10.1080/00273171.2018.1483224>
- Cinelli, C., Forney, A., & Pearl, J. (2020). A crash course in good and bad controls. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3689437>
- Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'Small', 'Medium', and 'Large' for Power Analysis. *Trends in Cognitive Sciences*, *24*(3), 200–207. <https://doi.org/10.1016/j.tics.2019.12.009>
- Crutzen, R., & Peters, G.-J. Y. (2017). Targeting next generations to change the common practice of underpowered research. *Frontiers in Psychology*, *8*. <https://doi.org/10.3389/fpsyg.2017.01184>
- Díaz, I., & van der Laan, M. J. (2013). Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *The International Journal of Biostatistics*, *9*(2), 149–160. <https://doi.org/10.1515/ijb-2013-0004>
- Flake, J. K., & Fried, E. I. (2019). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*. <https://doi.org/10.31234/osf.io/hs7wm>
- Gelman, A., Hill, J., & Vehtari, A. (2021). *Regression and other stories*. Cambridge University Press.
- Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218. <https://doi.org/10.1177/2515245918771329>
- Glymour, C. (2001). *The Mind's Arrows*. The MIT Press. <https://doi.org/10.7551/mitpress/4638.001.0001>
- Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, *10*. <https://doi.org/10.3389/fgen.2019.00524>
- Goldberg, L. R. (1999). *Personality psychology in europe* (I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf, Eds.; pp. 7–28). Tilburg University Press.
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. <https://doi.org/10.1016/j.jrp.2005.08.007>
- Greer, F., & Liu, J. (2016). *Principles and methods of test construction: Standards and recent advances* (K. Schweizer & C. DiStefano, Eds.; pp. 272–287). Hogrefe Publishing.
- Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, *15*(5), 1243–1255. <https://doi.org/10.1177/1745691620921521>
- Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2021). Modeling psychopathology: From data models to formal theories. *Psychological Methods*. <https://doi.org/10.1037/met0000303>
- Hilpert, P., Brick, T. R., Flueckiger, C., Vowels, M. J., Ceuleman, E., Kuppens, P., & Sels, L. (2019). What can be learned from couple research: Examining emotional co-regulation processes in face-to-face interactions. *Journal of Counseling Psychology*.
- Hoyle, R. H., & Panter, A. T. (1995). *Structural equation modelling: Concepts, issues, and applications* (R. H. Hoyle, Ed.; pp. 158–176). SAGE Publications.
- Huang, Y., & Valtorta, M. (2006). Pearl's calculus of intervention is complete. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, *arXiv:1206.6831*, 217–224. <https://doi.org/10.5555/3020419.3020446>
- Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. *arXiv Preprint*, *arXiv:2203.06498*.
- Hünernund, P., & Bareinboim, E. (2021). Causal inference and data fusion in econometrics. *arXiv Preprint*, *arXiv:1912.09104v3*.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling*. Guilford Press.
- Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.
- Lavrakas, P. (2008). *Encyclopedia of survey research methods*. 1. <https://doi.org/10.4135/9781412963947>
- Loehlin, J. C., & Beaujean, A. A. (2017). *Latent variable models: An introduction to factor, path, and structural equation analysis*. Routledge Taylor and Francis.
- Maclaren, O. J., & Nicholson, R. (2020). What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. *arXiv Preprint*, *arXiv:1904.02826v4*.

- Marsman, M., Schönbrodt, F. D., Morey, R. D., Yao, Y., Gelman, A., & Wagenmakers, E.-J. (2017). A Bayesian bird's eye view of 'Replications of important results in social psychology.' *Royal Society Open Science*, 4(1), 160426. <https://doi.org/10.1098/rsos.160426>
- Maruyama, G. (1998). *Basics of structural equation modeling*. SAGE Publications, Inc. <https://doi.org/10.4135/9781483345109>
- Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*, 9(2), 147–163. <https://doi.org/10.1037/1082-989x.9.2.147>
- McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(sup1), 235–245. <https://doi.org/10.1080/00031305.2018.1527253>
- Navarro, D. J. (2021). If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology. *Perspectives on Psychological Science*, 16(4), 707–716. <https://doi.org/10.1177/1745691620974769>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Pearl, J. (2009). *Causality*. Cambridge University Press. <https://doi.org/10.1017/cbo9780511803161>
- Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. Wiley.
- Peters, J., Janzing, D., & Scholkopf, B. (2017). *Elements of causal inference*. MIT Press.
- Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90(429), 54–63. <https://doi.org/10.1080/01621459.1995.10476488>
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27–42. <https://doi.org/10.1177/2515245917745629>
- Rosseel, Y. (2012). An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, 2(2), 107–114. <https://doi.org/10.1177/2515245919838781>
- Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, 31(1). <https://doi.org/10.1002/icd.2295>
- Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (in press). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*.
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105(2), 309–316. <https://doi.org/10.1037/0033-2909.105.2.309>
- Shpitser, I., & Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9, 1941–1979. <https://doi.org/10.5555/1390681.1442797>
- Smith, G. T., Combs, J. L., & Pearson, C. M. (2012). Brief instruments and short forms. *APA Handbook of Research Methods in Psychology, Vol 1: Foundations, Planning, Measures, and Psychometrics.*, 395–409. <https://doi.org/10.1037/13619-021>
- Spurk, D., & Abele, A. E. (2010). Who earns more and why? A multiple mediation model from personality to salary. *Journal of Business and Psychology*, 26(1), 87–103. <https://doi.org/10.1007/s10869-010-9184-3>
- van der Laan, M. J., & Rose, S. (2011). *Targeted learning - causal inference for observational and experimental data*. Springer International.
- Vankov, I., Bowers, J., & Munafò, M. R. (2014). Article Commentary: On the Persistence of Low Power in Psychological Science. *Quarterly Journal of Experimental Psychology*, 67(5), 1037–1040. <https://doi.org/10.1080/17470218.2014.885986>
- Vowels, M. J. (2021). Misspecification and unreliable interpretations in psychology and social science. *Psychological Methods*. <https://doi.org/10.1037/met0000429>
- Vowels, M. J. (2022). Trying to outrun causality with machine learning: Limitations of model explainability techniques for identifying predictive variables. *arXiv Preprint, arXiv:2202.09875*.
- Vowels, M. J., Camgoz, N. C., & Bowden, R. (2021). Targeted VAE: Variational and Targeted Learning for Causal Inference. *2021 IEEE International Conference on Smart Data Services (SMDS)*. <https://doi.org/10.1109/smds53860.2021.00027>
- Vowels, M. J., Camgoz, N. C., & Bowden, R. (2022). D'ya like DAGs? A survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4), 1–36. <https://doi.org/10.1145/3527154>
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6), 632–638. <https://doi.org/10.1177/1745691612463078>
- Wood, J., Matthews, G. J., Pellowski, J., & Harel, O. (2018). Comparing different planned missingness designs in longitudinal studies. *Sankhya B*, 81(2), 226–250. <https://doi.org/10.1007/s13571-018-0170-5>
- Wright, S. (1921). Correlation and causation. *Journal of Agriculture Research*, 20, 557–585.
- Wright, S. (1923). The theory of path coefficients: A reply to Niles' criticism. *Genetics*, 8(3), 239–255. <https://doi.org/10.1093/genetics/8.3.239>

Appendix: Simulation Results

The purpose of the simulation is to illustrate the differences in χ^2 , Root Mean Squared Error of Approximation (RMSEA), Comparative Fit Index (CFI), Mean Absolute Error (MAE) and p -values, between two models which differ in complexity but which are otherwise correctly specified (with respect to the true, underlying DGP. It is worth noting that χ^2 is known as an ‘absolute’ fit index, and is not adjusted for model complexity. A lower χ^2 value indicates better fit and provides a measure of how much our sample covariance matrix differs from our fitted covariance matrix. In contrast, RMSEA adjusts for the model complexity (favouring model parsimony), and here a lower value is preferred. Finally, CFI is not adjusted for model complexity, and higher values are preferred. For more information on these metrics, readers are pointed towards works by Maruyama (1998; Hoyle & Panter, 1995).

It is important to note that under these conditions (and when researchers use the process/tools presented in this work), the causal effect size estimates are unbiased regardless of whether the full model or the reduced models are used. As such, even though the use of these tools can have an effect on the standard errors (and therefore also the p -values and null-hypothesis significance testing), it does not affect the large-sample performance of the model. Indeed, this is evidence in the lower four plots of [Figure A1](#), which confirm that the choice of model does not affect the effect size estimates (all are unbiased). Nonetheless, it is important to understand the possible impact on the various model metrics to understand that two different correctly specified models can yield different finite-sample behaviours. These differences are discussed in more detail in this section.

Simulation results for DGP examples 2-5 in [Figure 4](#) are shown in [Figures A1-A3](#). We use the *sem* function in the *lavaan* library (Rosseel, 2012) to estimate a single target effect for each variant. For the MAE and the p -values, we provide results for a single effect of interest. For example, for the DGP research question 2ii in [Figure 4](#), we specify the SEM models given in the ‘Full DGP’ and ‘Reduced’ columns and generate MAEs and p -values for the total effect of X on Y . Similarly, for DGP research question 3ii, we specify the SEM models given in the ‘Full DGP’ and ‘Reduced’ columns, and generate MAEs and p -values for the total effect of C on M . Finally, for example 5, we specify the SEM models given in the ‘Full DGP’ and ‘Reduced’ columns, and generate MAEs and p -values for the total effect of $A1$ on $B3$.

For each of the example DGPs, we generate data across a range of sample sizes (10-200), and for each sample size we undertake 100 simulations. The results of these 100 simulations are used to derive means and standard deviations for each of the metrics, thus allow us to compare the results when specifying the full DGP model compared with the reduced models.

Starting with the results for the model fit metrics χ^2 in [Figure A1](#), we see that for DGPs 2-4 the *reduced* models have better fit (lower χ^2 indicates better fit). This comes as no surprise because here the complexity of the model im-

pacts our ability to reduce error for the path coefficients we are estimating (reducing the degrees of freedom). For similar reasons, it is also not surprising that the differences for the full and reduced models for DGP 5 were not different - the reduced model did not differ greatly in its reduction of complexity. In this sense, reducing the complexity of the model can have an effect on the resulting χ^2 , in such a way that yields a value which is considered desirable (of course, in practice we should specify theories based on more than just the resulting fit-statistics).

In [Figure A1](#) we provide estimates for the target effect size ‘Coefs’, on top of the true effect size ‘True Coef’. Importantly, the results confirm that the simplification process does not bias the estimates - all model variants correct estimate the effect size.

Results for CFI (higher is better) and RMSEA (lower is better) are shown in [Figure A2](#). Once again, the smaller models are preferred and yield higher CFI values. This again comes as a consequence of the complexity of the larger models and the concomitant impact on estimation. This notwithstanding, as the sample size increases, the results converge fairly quickly. The RMSEA results indicate a great improvement with the use of the reduced models, particularly for smaller sample sizes. This is not surprising because RMSEA is an adjusted metric, and so the results are consistent with the expectation that lower RMSEA values are associated with smaller models.

Finally, the p -values and MAEs for the target effect size estimates are shown in [Figure A3](#). For DGP 2 (top left plot), the p -values are higher for the reduced model than the complete model. This is consistent with the expectation that the inclusion of more variables can help increase the precision of our estimates. Indeed, in general we expect that the inclusion of variables into a structural equation model will reduce the standard error and, by the mathematical expressions relating these quantities, also reduce the p -values. However, this is only reliably the case if the model is correctly specified, and the reason it happens is because we are able to partial out the variance more completely. For example, consider the graph $X_1 \rightarrow Y \leftarrow X_2$. Here, Y has two causes, but let’s say that we actually only care about the link $X_1 \rightarrow Y$. In this case we have two options: create an SEM which includes $X_2 \rightarrow Y$ (in addition to the $X_1 \rightarrow Y$ link), or create an SEM which does not. Note, however, that the inclusion of $X_2 \rightarrow Y$ can help us estimate $X_1 \rightarrow Y$ because it partials out variance in Y which, in a finite sample, might otherwise be attributable to X_1 . Unfortunately, in practice it may not be as simple as this, because every time we include a new variable and a new path, we also increase the chances that we incorrectly specify the graph. Thus, whilst the option to reduce standard error by the inclusion of more paths is perhaps still a good thing to consider/understand in general, doing so requires us to be more and more confident that our specification is correct as we include more and more paths in our model.

Returning to the examples in the figure, the reduced model in DGP 2i only includes two effects of the outcome Y , which is X and K . However, other more proximal variables P and R exist, and their inclusion would improve the

quality of the estimate. In this case, R and P would be doubling as both control variables (adjusting for the backdoor path from X to Y , as well as variables which aid in precision (Cinelli et al., 2020). Note also that the standard deviation of these p -values is higher, indicating greater variation across simulations. This increased variance also results in a higher MAE, which is also evidence in the DGP2 - MAE plot in [Figure A3](#) (third row, first column). Thus, even though the effect size estimates will be unbiased (owing to correct specification of the reduced model with respect to the full DGP), the removal of explanatory variables can impact the precision of the estimates. In order to compensate for this, one can choose to retain variables which have explanatory power so long as their inclusion does not contradict the full, underlying model. DGP 2 represents a useful example insofar as variables R and P can be included (optionally in addition to K), to help explain the effect of X on Y .

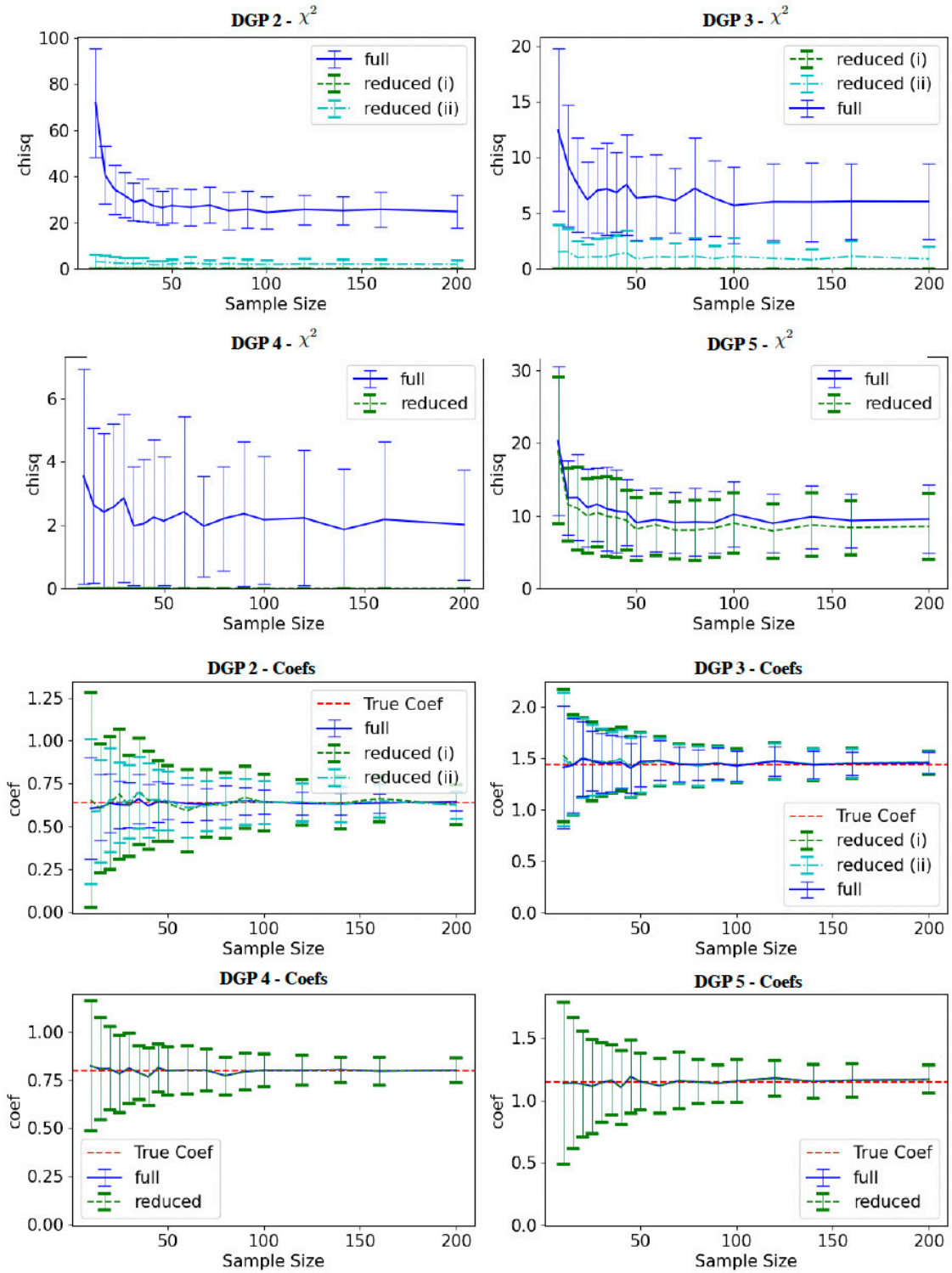


Figure A1. Simulation X^2 and Coefficient Estimation Results.

Note. Averages and standard errors over 100 simulations with varying sample sizes for χ^2 and estimated coefficient values for data generated from Data Generating Processes (DGPs) 2-5 in [Figure 4](#).

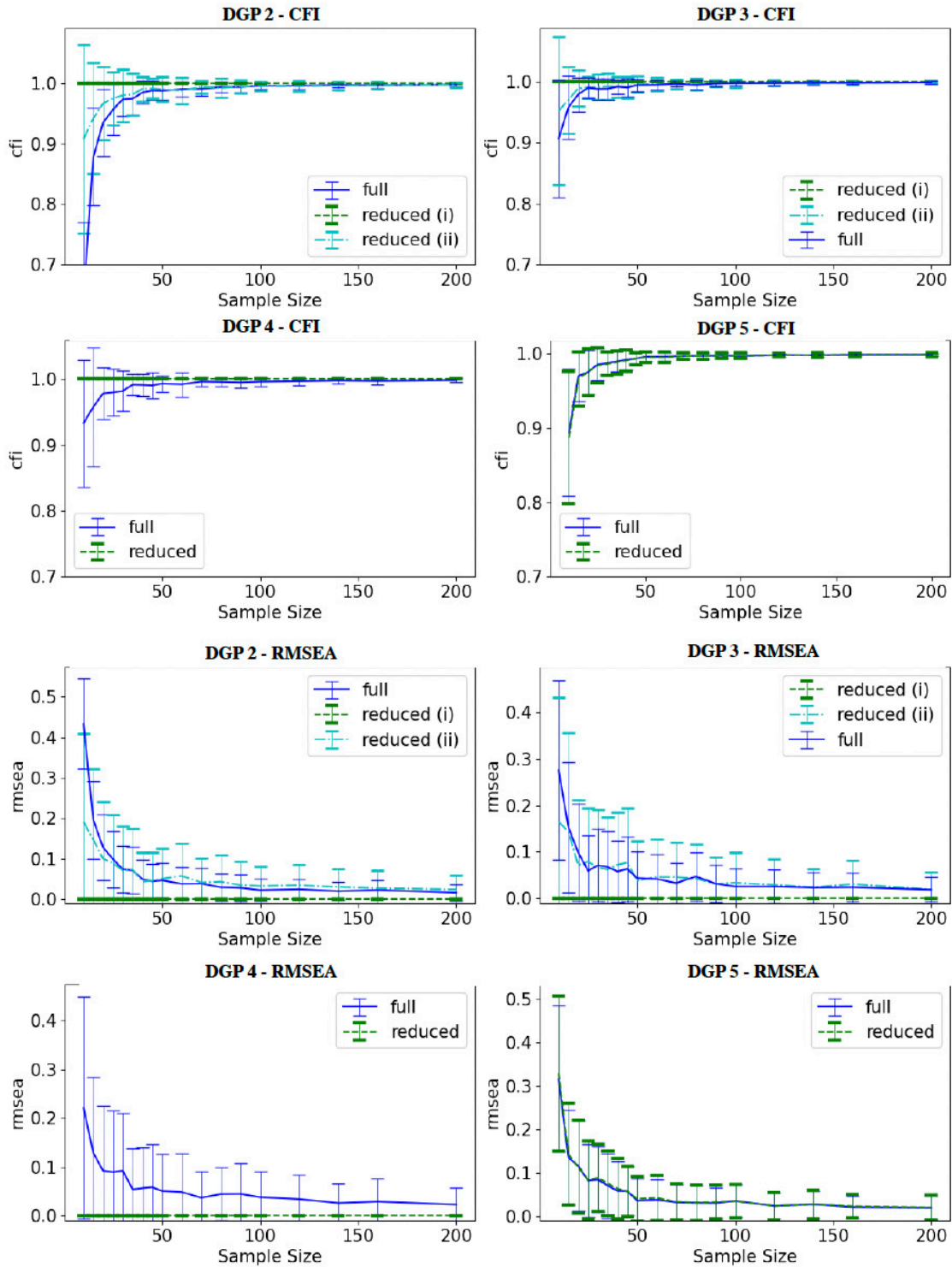


Figure A2. Simulation CFI and RMSEA Results.

Note. Averages and standard errors over 100 simulations with varying sample sizes for Comparative Fit Index (CFI) and Root Mean Squared Error of Approximation (RMSEA) for data generated from Data Generating Processes (DGPs) 2-5 in [Figure 4](#).

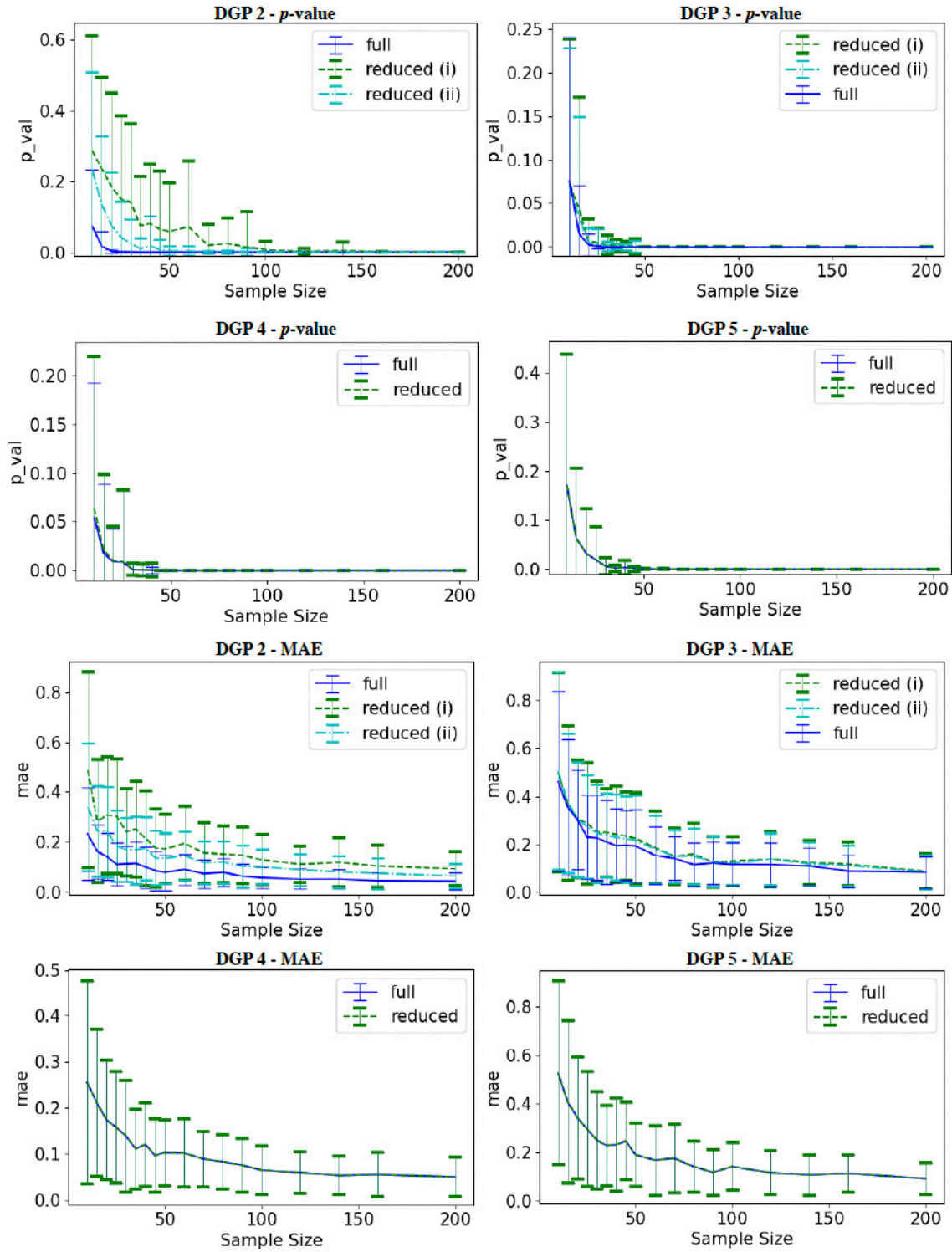


Figure A3. Simulation p-value and MAE Results.

Note. Averages and standard errors over 100 simulations with varying sample sizes for p-values and Mean Absolute Error (MAE) for data generated from Data Generating Processes (DGPs) 2-5 in Figure 4.

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/71300-prespecification-of-structure-for-the-optimization-of-data-collection-and-analysis/attachment/148422.docx?auth_token=xkgpNrDLLX2LywifUK-P

COI_DAS

Download: https://collabra.scholasticahq.com/article/71300-prespecification-of-structure-for-the-optimization-of-data-collection-and-analysis/attachment/148423.docx?auth_token=xkgpNrDLLX2LywifUK-P
