Aligning Methodology with Research Questions

By

ANNA WYSOCKI

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Mijke Rhemtulla

Emilio Ferrer

Philippe Rast

Committee in Charge

2023

Acknowledgements

First, I want to thank my advisor Mijke Rhemtulla. Even after 6 years of knowing and working with you, I barely believe that I got an advisor as intelligent, kind, understanding, and inspiring as you are. I generally have good luck, but with you I had wildly good luck. Thank you for all the hours you've spent mentoring me. Thank you for never laughing at my not-smart questions. Thank you for making me publish papers when I wanted to just add more stuff in perpetuity (how will I know a project is done without you?). Thank you for teaching me so much. Thank you for tolerating my chaos. Thank you for letting me teach your toddlers to trust fall. Thank you for letting me end Zoom meetings immediately (no awkward silences!!). Thank you for the knowledge, for the dinners, for the care, for the constant (and so appreciated) support. Thank you, thank you, thank you. Advisors can make or break a graduate student's experience, and you made the last 6 years the best and most intellectually expansive years of my life.

Thank you also to Philippe Rast and Emilio Ferrer who are not only members of my dissertation committee (and every other committee I've had in graduate school) but also mentors. Thank you for the time you have spent advising and counseling me. I have learned so much from each of you, and I am deeply appreciative. Thank you also to Lisa Welling for mentoring me on my path to a PhD program.

Thank you to Valerie Starratt—my first academic mentor. I think often about how you believed in and pushed college-aged me, and it was (with no hyperbole) life changing. You showed me the qualities I should look for and expect in a mentor. Your standards and expectations were always high, but so was your support and generosity. From the bottom of my heart, thank you.

To all my friends (Kristi, Jen, Ted, Nate, Madeline, Marcela, Chloe, Lindsey, Phil, Clara, Tanya, Kurt, Kailey, Arianne, Donnie, Lindsay, Cody, Sydney, Nora, Kelli, Kait, Çağrı, Eva, Tyler, Justin, and Michael), thanks for listening to me talk about my feelings (rarely), statistics (sometimes), and the wilderness (all the time). You have all been the weight on the life side of my work-life balance. Without all of you I would have published more but I would have been a lot less happy.

To my siblings, I'm sorry I'm Mom's favorite. That must be really hard for you. But can you blame her?

To Michelle—my sister and second mom—I am constantly humbled and empowered by how much you believe in me. The fact that you exist makes me braver and the world feel safer.

And finally, Mom. It's difficult to express how much I appreciate you. Being a parent seems really exhausting, and you were both parents to a band of 7 gremlins with cute faces. I don't know how you made sure we all survived, but I am so glad we got you as a mom and as a math teacher. You taught me algebra, independence, determination, kindness, and curiosity, and I love you with my whole heart.

Abstract

Psychological researchers investigate complex phenomena with imperfect measures and limited resources. Quantitative methods can ameliorate these challenges, but only when the methods applied match the research question and goal. Because of this, methodologists explore the consequences of mismatched methodology, and develop guidelines for methodology selection. Regardless of the research question, it is important to (1) clearly state the theoretical target, (2) select an estimate that can inform this theoretical target, and (3) select a valid estimator for the estimate. The following chapters in this dissertation focus on proposing a guide for variable selection, developing a model that allows for the integration of information across models (both important for estimate selection), and evaluating different estimation approaches (important for selection of the estimator).

Table of Contents

Chapter 1. Introduction

The Importance and Challenges of Complex Systems

Psychologists study complex systems with many interacting parts. One reason psychologists are interested in these systems is because understanding their structures and mechanisms has real-world implications. For example, knowing how race and gender biases impact decision making can make college admissions more equitable. Or understanding the principles of childhood learning can guide the development of a fourth-grade reading curriculum.

The complexity of these systems, however, means that learning about them is difficult. These systems are often centered on multi-dimensional constructs with definitions that are still debated among experts and that are difficult to measure. These systems also have a multitude of important influencing factors making it difficult to know or measure the full set of important variables. Further, the relations between these sets of factors are challenging to model due to the number and complexity (e.g., non-linear, multi-way interactions). As such, most psychological research studies have an important research question but only a fraction of the resources or information necessary to easily glean a valid answer.

These challenges are why psychologists rely heavily on statistical models. Statistical models can be used to "make up the distance" between our research goals and our resource, design, measurement, and data limitations (Smaldino, 2016; 2017) For example, latent variable models can remove measurement error from imperfectly measured constructs (making up the distance between how we measure variables and how we wish we could measure variables; Wansbeek, & Meijer, 2001) and random-effects models can remove unmeasured confounding from causal estimates (making up the distance between the set of variables we measured and the set of variables we wish we could measure; Kim & Steiner, 2021). But these methods will only improve the validity of inferences if the modeling choices—which variables are included in a

model, what model is selected, which estimator is used—are matched with the research question and goal (Lundberg, Johnson, & Stewart, 2021).

The Impact of Misaligned Research Questions and Statistical Models

Often in research articles, pages are devoted to describing a complex system and underscoring the importance and novelty of the research question. But then there is a jump to describing the methods and models used with little discussion of the link between the research question and the methodology (Lundberg, et al., 2021; Grosz, Rohrer, & Thoemmes, 2020). This jump is a problem because a mismatch between the research goal and the methodology can result in incorrect inferences about the systems of interest.

The negative impact of this misalignment is a topic that psychometricians have pointed out across various areas. For example, Rhemtulla and colleagues (2021) explained how latent variable models—used to remove measurement error from a construct of interest—will not remove measurement error or provide unbiased estimates if the construct being modeled is not a common factor underlying its indicators. Rohrer (2018) described how failing to control for an appropriate set of confounders can lead to biased estimates. Hamaker and colleagues (2015) wrote about how failing to model random intercepts when there are stable between-person differences leads to a conflation of between and within-person effects in estimated parameters.

Improving the Link between Research Questions and Statistical Models

Given the wide-spread impact of a misaligned research goal and methodology, multiple frameworks have been proposed to guide researchers through the selection of the research methods and models (e.g., Borsboom et al., 2021; McElreath & Smaldino, 2015; Lundberg et al., 2021). Many of these guides focus on the importance of discussing the rationale behind design

and modeling decisions and underscore the value of acknowledging the uncertainty in those decisions.

One such framework, proposed by Lundberg and colleagues (2021) suggests that all research studies should have 3 steps when developing a research question and matching it with a methodological approach. First, specify a theoretical target—the object of inquiry—which is the unit-specific quantity of interest as well as the target population. Here, researchers should argue why this theoretical target is important for understanding the larger system. Second, select the parameter that will be estimated to learn about this theoretical quantity—the estimate. Here, it is important to link the estimate to the theoretical quantity—how does this estimate help us learn about the theoretical quantity? Third, select an estimator—what approach will be taken to estimate the previously selected parameter? The selection of the estimator is informed by the target estimate as well as the data availability and constraints.

Making these three steps distinct tasks that build on each other allows researchers to make clear what question they are interested in, justify their interest in that question, and defend how their specific research question is informative for the larger theory. And, most importantly to this dissertation, it allows researchers to then center all following decisions—what variables to measure, what model to select, and how to estimate that model—on the clearly specified theoretical target. The following chapters discuss specific frameworks and models for selecting an estimate once the theoretical target has been specified (Chapter 3 & 4) and selecting an estimator once the estimate has been specified (Chapter 2). In the following subsections we outline chapters 2, 3, and 4 and describe how they add to the literature on the alignment of research goals and methodology.

Estimating Network Models (Chapter 2)

Selecting the correct estimate for a theoretical target is a challenging but important endeavor, and it will often depend on the underlying structure of the construct(s) of interest. Common factor models, where measured variables are hypothesized to reflect a latent variable, are widespread in psychology for modeling constructs such as personality and psychopathology wherein the latent factor is thought to cause the measured variables (e.g., behaviors, symptoms; Borsboom, 2008; Borsboom, Mellenbergh, & van Heerden, 2003). More recently, however, psychometric network models, which depict psychological constructs as a system of behaviors, symptoms, or attitudes with direct causal influences on each other have been proposed as an alternative model (Borsboom & Cramer, 2013; Epskamp, Maris, Waldorp & Borsboom, 2018). If the construct of interest is a system of interacting components, then a network model can likely capture the construct better than the latent variable model.

Network models were introduced to psychology a decade ago, and since then, have become popular particularly for modeling psychopathologies (Borsboom, 2017; McNally et al., 2015). But psychometric network models are new and as such how to estimate them (the final step in the 3-step framework outlined above) is an ongoing topic of research. Initially, the default approach to estimating network models was a regularization technique which induces a sparse network by deleting many of the paths connecting pairs of variables in a network (Friedman, Hastie, & Tibshirani, 2008). However, regularization was developed for high-dimensional settings—when the number of variables exceeds the sample size—which are uncommon for psychological data (Foygel & Drton, 2010; Friedman et al., 2010). Regularization decreases variance across estimates which in high-dimensional settings can make a complex model estimate-able. But it is likely not necessary for estimation in low-dimensional settings. As such, it is important to investigate how regularization performs in low-dimensional settings. In chapter

2 we explore whether regularization techniques are appropriate for psychological settings by comparing the performance of a set of regularization techniques in low-dimensional settings. Our goal here is to evaluate whether regularization is an appropriate estimation approach for the estimates of interest—the direct relations between variables.

Causal Inference (Chapters 3 & 4)

Psychology is having a causal reckoning where some researchers are starting to realize that it is impossible to explain and understand processes without invoking causality (Grosz, et al., 2020; Foster, 2010). Many of the research questions that psychologists are interested in are centered on explaining systems and understanding how to change or intervene on these systems (Foster, 2010). This means that many psychological theoretical quantities are causal.

However, the methodologies and research practices used in psychology are often not, by themselves, suited for causal inference (Rohrer, 2018). Fortunately, tools for causal inference have been extensively developed and discussed in fields like epidemiology and economics (e.g., Greenland, Pearl, & Robins, 1999; Hernán, Hernández-Díaz S & Robins, 2004; Hernán et al., 2008). But for these tools to be useful for psychological research, they need to be adapted to the research designs and variables that psychologists deal with. For example, variables in epidemiology tend to be simpler to measure (e.g., are you a smoker?) or discrete events (e.g., what medication were you prescribed?). Whereas in psychology, many variables are complicated to measure and describe states (e.g., how satisfied are you with your life?) rather than events.

As such, psychometricians have begun to adapt existing causal inference tools from other fields and develop new tools with the goal of making causal inference available to psychological researchers (e.g., Grosz et al., 2023; Kim & Steiner, 2021; Rohrer, 2018). These recent endeavors are centered on developing both estimates and estimation approaches that are valid for

learning about causal theoretical quantities in psychology. Chapter 3 and Chapter 4 in this dissertation develop two such contributions.

When selecting an estimate that can inform on a causal theoretical quantity, a key step is controlling for the correct set of confounding (or biasing) variables. Chapter 3 outlines why the current approach to selecting control variables in psychology is not well-suited to the goal of causal inference and proposes a better suited framework for control variable selection. When the theoretical quantity is causal, it is inherently longitudinal (causes precede effects; change occurs over time), but psychologists regularly have cross-sectional data. Likely any estimate or estimation approach based on cross-sectional data will not be aligned with the causal theoretical quantity. Chapter 4 develops a model that allows researchers to integrate existing longitudinal information into cross-sectional estimates to create more robust and less biased estimates. Overall, the chapters in this dissertation develop methods and frameworks with the goal of helping psychologists better align their methodologies to their research questions.

Chapter 2. On Penalty Parameter Selection for Estimating Network Models[1]

Network models are becoming popular in psychology (Borsboom, 2017; McNally et al., 2015) largely because they provide a theoretical alternative to latent variable models and the common cause framework (Borsboom & Cramer, 2013). For example, psychopathologies such as depression are often conceptualized as arising from a common or underlying cause. As this cause is unobservable (i.e., the latent variable), the symptoms are considered passive indicators that allow for inquiry and the ability to diagnose the disorder. On the other hand, networks conceptualize constructs as systems arising due to interactions between variables rather than due to an underlying cause (Epskamp, Maris, Waldorp & Borsboom, 2018). In practice, network models are used to estimate relations between nodes (e.g., symptoms), identify hubs (i.e., highly connected nodes), and visualize the overall structure of a construct (McNally, 2016); for a theoretical discussion and comparison of latent versus network models see Borsboom and Cramer (2013).

One commonly estimated network is the Gaussian Graphical Model (GGM) wherein nodes represent random variables, and edges represent conditional independencies, estimated as partial correlations, between variables (Lauritzen, 1996). The estimation of partial correlations can produce rich inferences as variables that directly activate each other will be connected assuming all important variables are included in the model. As applied researchers can never be certain that this criterion is met, partial correlations provide a possible causal skeleton for a construct (Edwards, 2012). In psychology, GGMs are used to estimate, for example, symptom,

[1] This chapter was slightly adapted from published paper: Wysocki, A. C., & Rhemtulla, M. (2021). On penalty parameter selection for estimating network models. *Multivariate Behavioral Research, 56*(2), 288-302.

personality, and health behavior networks (Costantini et al., 2015; Fried et al., 2017; Kossakowski et al., 2016).

An important aspect of GGMs is not only the identification of important conditional relations, but also the identification of truly zero edges, which is necessary to achieve a sparse network. Setting edges to zero is a key feature of network estimation as having a fully connected model is less helpful than having a few potentially meaningful connections to focus on in future experiments or interventions. However, it is important to achieve this sparsity in a justified manner.

In psychology, inducing sparsity is typically done through a form of penalized maximum likelihood, the graphical lasso or "glasso" (Friedman, Hastie, & Tibshirani, 2008), using a penalty selection method called EBIC (Foygel & Drton, 2010) for selecting the degree to which the likelihood is penalized. Minimizing the EBIC is used to select the penalty parameter λ, that in turn achieves the goal of edge selection. Although there are multiple methods to select λ for the glasso equation (Kuismin & Sillanpää, 2017), EBIC has emerged as the default in psychology with no published work establishing if its performance is superior for psychological data. We understand that 'psychological data' is a broad term and given the variety of research that is done in psychology (e.g., neural, psychopathology, social research) no single dataset or template could characterize all psychological data. We use the term to refer to psychopathology symptoms and personality scales, which have been the target of most glasso estimated networks to date (Beard et al., 2016; Briganti, Kempenaers, Braun, Fried, & Linkowski, 2018; Bryant et al., 2017; Pereira-Morales, Adan, & Forero, 2019)

The selection of the penalty parameter term in glasso, λ, is critical as different values applied to the same data can result in different networks (Epskamp & Fried, 2018; Kuismin &

Sillanpää, 2017). For example, when $\lambda = 0$ the network is no longer penalized, and, assuming no sample partial correlations are precisely zero, the resulting network is fully connected (i.e., all edges are non-zero). As λ increases, so does the penalization of the network, resulting in an increasingly sparse network (i.e., fewer edges are estimated) eventually resulting in an empty network (Liu, 2013). Within the glasso framework, penalty selection is done through an automated data mining process whereby a sequence of λ are tested, and one is selected based on whether the corresponding network optimizes some criterion. This criterion depends on which penalty selection method is being used. As different methods have divergent priorities (e.g., stability, sparsity, predictive ability), they often select different λ and can return vastly different networks (Kuismin & Sillanpää, 2017). The aim of the present work is to characterize the performance of penalty selection methods. We seek to fill this gap in the literature by comparing four penalty selection methods for the glasso in conditions representative of psychological data.

The rest of this chapter is outlined as follows. In the next section, we describe network estimation. Then, we outline four penalty selection methods: CV, StARS, RIC, and EBIC, and discuss the advantages and limitations of each approach. Then, with a motivating example, we use each method to estimate the network structure of PTSD symptoms, where we highlight how each method can (sometimes) estimate drastically different networks. Importantly, we also provide an overview of specific network characteristics (e.g., network density and partial correlation size) based on a review of published psychological networks. We next present two simulation studies and their results ending with a discussion of these findings in relation to psychological networks, the practical implications of this work, and future directions for methodological inquiry.

Network Estimation

A key feature of estimating networks is imposing a sparsity pattern on the precision matrix (**Θ**), the inverse of the covariance matrix (**Σ**). The sparsity pattern of **Θ** provides the structure of the network model where a non-zero value in the off-diagonal represents an estimated edge between nodes. Networks discussed and estimated in the current paper are undirected, as are almost all networks estimated from cross-sectional data (Borsboom et al., 2021b). This is because of the large number of equivalent directed models that result in the same fit (Hitchcock, 2001). Without experimental manipulation or strong theory, it is challenging to decrease the space of possible directional graphs. The precision matrix can be used to obtain partial correlations following:

(1)
$$cor(y_i, y_j \mid y_{-(i,\,j)}) = \frac{-\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}},$$

$$\text{where } \Theta = \begin{bmatrix} \theta_{ii} & \cdots & \theta_{ji} \\ \vdots & \ddots & \vdots \\ \theta_{ij} & \cdots & \theta_{jj} \end{bmatrix}$$

where the partial correlations represent the correlation between each pair of variables, controlling for all other variables in the model. When variables can be assumed to follow a multivariate normal distribution, the precision matrix can be estimated using normal-theory maximum likelihood estimation. In psychological applications, the most common estimation method used when the data approximate multivariate normality is the graphical lasso (glasso), a form of penalized maximum likelihood that uses an ℓ_1-penalty (Friedman et al., 2008), minimizing the likelihood:

(2)
$$\iota(\widehat{\boldsymbol{\Theta}}) = \log\det\widehat{\boldsymbol{\Theta}} - tr(\boldsymbol{S}\widehat{\boldsymbol{\Theta}}) - \boldsymbol{\lambda}\sum_{i\neq j}(|\widehat{\boldsymbol{\Theta}}_{i,j}|)$$

where **S** is the sample covariance matrix, $\widehat{\boldsymbol{\Theta}}$ is the estimated precision matrix, and $\lambda \epsilon \{0, 1\}$ is the penalty parameter that is applied to the sum of the absolute edge weights. This penalty results in all edge weights being shrunk toward zero, and many of them set to exactly zero. In other words, glasso performs both edge selection and weight shrinkage, with the choice of λ affecting both which edges are estimated and the size of the estimated edge weights. When λ is set to 0, the penalty term drops out and the equation returns to normal-theory (i.e., non-penalized) maximum likelihood such that is typically used in regression:

$$\iota\left(\widehat{\boldsymbol{\Theta}}\right) = \log\det\widehat{\boldsymbol{\Theta}} - tr\left(S\widehat{\boldsymbol{\Theta}}\right) \tag{3}$$

Again, there are different methods to select λ with divergent criteria. It is important to note, because glasso estimates the entire network in one step, only one penalty parameter is used in estimation. This is in contrast to methods that use a series of univariate regression models to estimate edges for each node individually (Meinshausen & Bühlmann, 2006; Ravikumar, Wainwright, & Lafferty, 2010). In the next section, we outline four penalty selection methods.

Methods for Parameter Selection

Cross-Validation (CV)

In psychology, there has been a surge of interest in predictive modeling, which stands in contrasts to more traditional explanation-centric frameworks (Yarkoni & Westfall, 2017). The goal of cross-validation (CV) is not exclusively to make inferences about individual parameters, but to select a model that is able to predict out-of-sample data. Although inferences are still possible, cross-validation is prone to over-selection which results in a higher false positive rate (Chetverikov, Liao, & Chernozhukov, 2021; Yu & Feng, 2014). These limitations hold for CV in both regression and network settings. Although there are different ways to implement CV, the

general procedure is to partition the data into a training set and a test set. Different forms of CV (e.g., leave-one-out, *K*-fold) have been applied to network estimation (Efron, Hastie, Johnstone, & Tibshirani, 2004; Friedman et al., 2008; Friedman, Hastie, & Tibshirani, 2010; Zhang, 1993). For our purposes, we will focus on *K*-fold CV, which is computationally more efficient and has been found to have greater stability than other forms of CV (Homrighausen & McDonald, 2014; 2017).

K-fold CV partitions the data into *K* non-overlapping subgroups. Using Equation 2, a sparse precision matrix is estimated with the pooled data from *K* - 1 of the subgroups, inverted into a sparse covariance matrix ($\widehat{\boldsymbol{\Sigma}}_{\text{train}}$) and then tested on predicting the covariance matrix from the remaining group (i.e., the testing group; ($\widehat{\boldsymbol{\Sigma}}_{\text{test}}$). The prediction error is computed across folds (i.e., until each subgroup has been the test group) and averaged. The prediction error is estimated as follows:

$$\ell_{CV}\left(\hat{\Sigma}\right) = \frac{1}{K}\sum_{i=1}^{K} -\log\det\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{train}} - \text{tr}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{test}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{train}}^{-\mathbf{1}}) \quad (4)$$

This procedure is repeated across the range of λs resulting in a mean prediction error (prediction error is averaged across folds) for each λ. The λ that minimizes the mean cross-validation error is selected (for more details see Bien and Tibshirani, 2011).

Stability Approach to Regularization Selection (StARS)

StARS uses a re-sampling method to select a λ that provides maximal network stability (Liu, Roeder & Wasserman, 2010). There are parallels between this method and bootstrapping (although bootstrapping re-samples with replacement and StARS without) wherein both methods use re-sampling to assess the variance or stability of a model across samples (Efron & Tibshirani,

1993). StARS does this by drawing *K* random, overlapping subsamples and fitting the range of λs to each of the subsamples. StARS begins with a large λ that results in an empty network providing stability with no variation between groups and λ is gradually reduced until there is a small but acceptable amount of variability between the subsample networks. Total instability for a given λ is defined as

$$D = \frac{\sum_{i \neq j} \left(2 \left(\xi_{ij}(\lambda) \right) \left(1 - \xi_{ij}(\lambda) \right) \right)}{\frac{p(p-1)}{2}} \quad (5)$$

where *p* is the number of variables in the dataset and $\xi_{ij}(\lambda)$ is the probability of a network having a specific edge calculated as

$$\xi_{ij} = \frac{\sum_{i \neq j} \hat{a}_{ij}}{K} \quad (6)$$

Where $\hat{a}_{ij} = 1$ when $\theta_{ij} \neq 0$ and $\hat{a}_{ij} = 0$ when $\theta_{ij} = 0$. The smallest λ with a total instability between .01 < D(λ) < .08 is selected (Liu, Roeder & Wasserman, 2010).

Extended Bayesian Information Criterion (EBIC)

Information criteria, for the purpose of model section, are commonly used in psychology, and most can be justified in several ways. For example, minimizing the BIC approximates selecting the most probable model, assuming the true model is in the candidate set (Raftery, 1995). When the ratio of sample size to number of variables is small, it has been noted that the BIC does not necessarily select a parsimonious model (Chen & Chen, 2008). As such, EBIC was developed by introducing an additional manually-set penalty, γ, to the BIC equation (Foygel &

Drton, 2010).[2] γ controls the prior probability of sparse models resulting in the return of sparser networks as γ increases (Chen & Chen, 2008; 2012). EBIC is calculated as

$$\text{EBIC} = -2\,\iota\left(\widehat{\boldsymbol{\Theta}}\right) + \text{E}\log(\text{n}) + 4\gamma\text{E}\log(\text{p}) \tag{7}$$

where $\iota\left(\widehat{\boldsymbol{\Theta}}\right)$ is defined in Equation 2 and E is the size of the edge set (i.e., the number of non-zero elements of $\widehat{\boldsymbol{\Theta}}$. When γ = 0 the added penalty is dropped, and the equation is reduced to the BIC. The selected network minimizes the EBIC with respect to λ. Per recommendations by Foygel & Drton (2010), the default setting for γ in popular R packages for estimating network models, such as **qgraph**, **glasso**, and **huge**, is .5 (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012; Friedman, Hastie, & Tibshirani, 2019; Zhao, Liu, Roeder, Lafferty, & Wasserman, 2012.

Epskamp (2016) investigated EBIC's performance across conditions ($p = 25$, $50 \leq n \leq 2500$) that are typical to psychology and showed that the sensitivity (defined in section *Performance Meausures*) of EBIC increased with sample size. Specifically, when sample size was less than 500, sensitivity was generally below 75% (i.e., the method only estimates 75% of the true edges). Additionally, EBIC is sensitive to the magnitude of edges in the population network wherein larger partial correlations resulted in worse performance (Williams, Rhemtulla, Wysocki, & Rast, 2019)

[2] The method originally proposed by Foygel & Drton (2010) is slightly different than the implementation of EBIC in the R package **huge** and, until recently, **qgraph**. Specifically, the original suggestion was to use EBIC to select a penalized model structure, and to use non-penalized maximum likelihood to estimate the model parameters once the structure was chosen. The **huge** and original **qgraph** implementations of EBIC instead use it to select both a penalized model structure and to obtain penalized parameter estimates.

Rotation Information Criterion (RIC)

The RIC uses a rotation procedure, similar to a non-parametric permutation test, to select λ (Zhao et al., 2012; Zhu & Cribben, 2018). The data are reshuffled by randomly rotating the rows within each column of data. This procedure produces a rotated dataset in which all relations among variables are spurious. The reshuffled dataset is used to construct the expected sampling distribution under the null hypothesis (i.e., the null distribution). The RIC then finds the smallest value of λ that accurately regularizes all (spurious) edges to 0. This rotation procedure is repeated a number of times (the default in its R package, **huge**, is 20) and the smallest calculated λ across all rotations is returned as the selected penalty for the network.

Method Performance

The glasso depends on a number of assumptions and necessary conditions to perform well. One assumption is that the underlying population matrix is sparse. In the statistics literature, sparsity has been defined as having fewer true edges than sample size (Meinshausen & Bühlmann, 2006; Tibshirani & Wasserman, 2015). This is the minimum level of sparsity required for a network to be estimated. However, in psychology, sample sizes generally greatly exceed the number of variables, so the number of possible edges rarely meets this limit (see Section *Psychological Network Review*). As such, it is unclear what the impact of sparsity will be in this case. Second, there are two conditions that must be met to ensure consistent estimation. The irrepresentable condition is satisfied when unimportant variables are not highly correlated with important variables. More specifically, this condition is satisfied when the sum of irrelevant covariance is less than 1 (Zhao & Yu, 2006). The beta-min condition is satisfied when non-zero coefficients are sufficiently large; see Gauraha and Swapan (2018) for a full discussion of these necessary conditions.

Further, as these methods were developed for data where the ratio of sample size (*n*) to the number of variables (*p*) is small, most simulations have explored glasso's performance in such high-dimensional settings (Foygel & Drton, 2010; Friedman et al., 2010). Of the simulations that have looked at glasso's performance in low-dimensional settings, where *p* is smaller than *n*, many have only used EBIC to select the penalty parameter (Epskamp, 2016; Epskamp & Fried, 2018; Williams et al., 2019). However, there is a small amount of work comparing penalty parameter selection methods. For example, StARS was found to be competitive against BIC and AIC in conditions with a small *n* to *p* ratio (performance was measured using F_1-scores, a measure of both recall and precision; Liu et al., 2010). Overall, StARS consistently estimated sparser graphs than other methods and as a result had a lower false positive rate but was also less sensitive (see Section *Performance Measures* for definition). *K*-fold CV was also compared to StARS and was found to return a denser network with more false positive errors but fewer false negative errors (Liu et al., 2010). Mohammadi and Wit (2015) compared EBIC, RIC, and StARS across four low-dimensional conditions. They found that RIC and StARS were competitive with EBIC even outperforming (performance again measured by F_1-scores) EBIC in multiple conditions particularly those most comparable to psychological data.

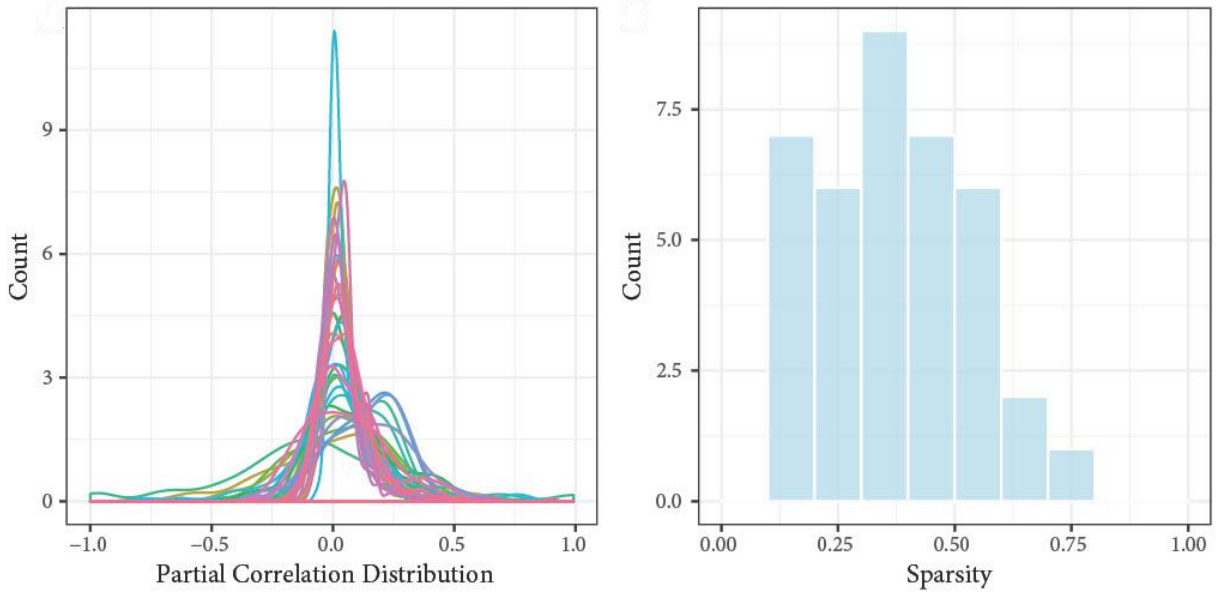
In simulations comparing penalty selection methods, there have been few low-dimensional conditions. This means that there has not been a full characterization of the impact of sample size and variable number on performance across methods. It also may be important to characterize these methods specifically for psychology-typical data. GGMs have typically been used to estimate gene or neural networks (Hecker, Lambeck, Toepfer, van Someren, & Guthke, 2009). However, in psychology they are being used to estimate constructs such as symptom and personality networks. As these are vastly different areas, there is the possibility that

psychological networks contain different sized partial correlations than gene and neural networks. Given the impact partial correlation size seems to have on these methods' performance, it is important to characterize method performance with partial correlations that approximate the size found in psychology. Finally, to our knowledge, no simulation comparing penalty selection methods has directly manipulated sparsity, the percentage of truly zero edges, across conditions. Therefore, we do not know whether sparsity or the lack thereof impacts penalty selection methods differently. As such, it is important to assess how these methods perform in conditions and data that are typical to psychology.

Psychological Network Review

To better understand the characteristics of psychological networks, we reviewed 37 recently published psychological networks assessing psychopathology (n = 33) and personality (n = 4) constructs. As the data for these two constructs were similar, their results will be presented together. Our review focused on the sample size, number of variables, sample sparsity and estimated partial correlations for each network. Figure 1 (left panel) depicts the distribution of sample partial correlations. Note these are non-regularized partial correlations (i.e., unbiased estimates). From this figure, we can see most networks have many partial correlations near 0, and large partial correlations are less frequent across networks. It is likely that many of these small partial correlations would be set to 0 through regularization. Although small partial correlations are more frequent, most networks have moderate to large partial correlations as well (see Table 1 for median partial correlation ranges and Table 2 for percentage of partial correlations within different ranges).

Figure 1. *Sparsity and Partial Correlation Distributions from Network Review*



Note. Left Panel: The distribution of estimated partial correlations across 37 networks. Each line represents a different network. Partial correlation size is depicted on the x-axis. Right Panel: The distribution of sample sparsity across 37 networks. Sparsity is depicted on the x-axis.

To obtain an estimate of sparsity without the use of a regularization method, we set any partial correlations less than .05 to 0. Figure 1 (right panel) depicts the sparsity distribution across networks. Network sparsity varied from 25% to 75%. 76% of the networks had between 50% and 75% sparsity.

Table 1. *Descriptive Statistics for Study Characteristics from Network Review*

	Median	**Mean**	**SD**
Sample Size	404.5	1,044.74	2420.04
Variables	18.5	19.84	7.25
Max PC	.52	.56	.18
Min PC	.000	.001	.003

Table 2. *Percentage across Ranges for Study Characteristics from Network Review*

Sparsity		**Partial Correlation[3]**		**Sample Size**		**Number of Variables**	
Range	Percentage	Range	Percentage	Range	Percentage	Range	Percentage
0 - .25	0%	0 - .15	74.1%	100 - 250	42%	0 - 10	18%
.25 - .50	24%	.15 - .25	13.5%	250 - 500	13%	10 - 20	47%
.50 - .75	76%	.25 - .50	9.7%	500 - 1,000	26%	20 - 30	24%
.75 - 1	0%	.50 - 1	2.7%	>1,000	18%	>30	11%

Table 2 displays the percentage of networks that fall between specific ranges for the features sparsity, sample size, and variable number (see Figure A1 in the supplementary material for the distribution of sample size and variables across the networks). 55% of the networks had a sample size less than 500, and most networks had between 10 and 30 variables. From this review, we surmise that psychological networks are often not extremely sparse and contain many small and a few larger partial correlations. Further, sample sizes are typically under 500, and the number of variables ranges from 10 to 30.

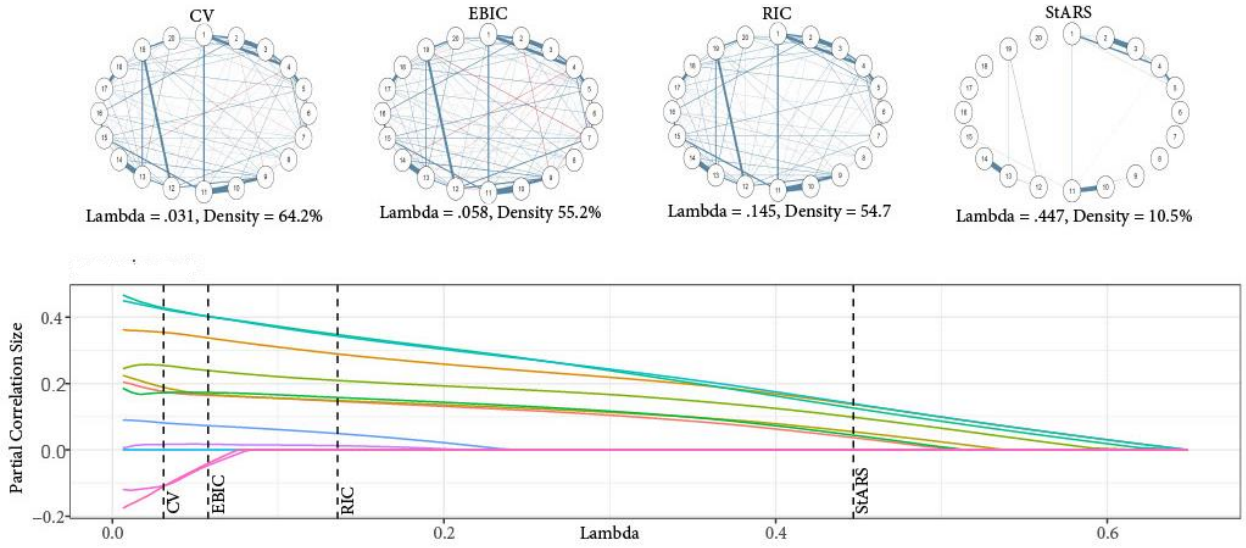
Motivating Example

To highlight the differences between methods, we estimated four network models, each using a different penalty selection method, from a post-traumatic stress disorder (PTSD) data set (McNally, 2016). Each method selected a different value of λ and estimated networks with different sparsities, although EBIC and RIC had near identical sparsities (< 1% difference; See

[3] Based on the absolute values of the partial correlations

Figure 2 (top panel) for visualization of each network). Even if a similar sparsity level is estimated across methods, the choice of λ also affects the edge weights (see Figure 2 lower panel). Comparing EBIC and RIC, even though each network has a similar number of edges, the edge weights within the RIC-selected network are smaller. Further, CV not only selects the densest network but also estimates the largest absolute edge weights while StARS selects the sparsest network with the smallest absolute edge weights. This example underscores how the choice of λ affects which edges get estimated and the edge weights, and, given that each penalty selection method selects a different lambda it bolsters the need for guidance on which method to use.

Figure 2. *Estimated Networks and Partial Correlation Size from Applied Example*



Note. Top Panel: Returned networks estimated from a psychological dataset for each method. Bottom Panel: The solution path for each method. Each line depicts an edge (12 were selected out of the edge set) and the change in partial correlation size across lambdas. The dashed line represents the selected lambda for different methods.

Simulation Study

In psychology, no recommendations exist for the use of one penalty selection method over another. Rather the performance of one method, EBIC, has been examined across various

conditions (Epskamp, 2016; Epskamp & Fried, 2018). In an effort to bridge this gap, we conducted two simulation studies. The first used partial correlations estimated from a psychological dataset (Section *Simulation 1: Empirical Partial Correlations*), whereas the second simulated data to assess the effect of partial correlation size on performance (*Section Simulation 2: Simulated Partial Correlations*) By using empirical partial correlations from a psychological dataset we are establishing how these methods may reasonably perform when used to estimate a psychological network. Then, by varying partial correlation size, still within a range that is representative of psychological data, we are establishing how performance may vary with respect to different partial correlation ranges. The code for both simulations can be found at https://github.com/AnnaWysocki/Network-Penalty-Selection .

Simulation 1: Empirical Partial Correlations

In Simulation 1, we used a 20 variable PTSD symptom dataset (McNally et al., 2015). We selected this dataset for two reasons. First, the data have previously been used to assess the performance of EBIC allowing for greater comparability between our results and previous simulations, and second, simulating data based on a psychopathology dataset provides partial correlations that are comparable in size to those that are likely to appear in psychological network applications (see Figure A2 in the supplementary material for the distribution of estimated partial correlations).

Our goal was to outline how these methods compare to each other along with how data characteristics such as sparsity, number of variables, and sample size influenced their performance. In addition, preliminary simulations suggested that the performance of these methods varied not only across samples and conditions but also across populations within a condition (e.g., two networks with the same level of sparsity, number of variables, and sample

size with a different population matrix may return different false positive rates, on average, across randomly sampled data). To better assess this within-condition variability, we performed a nested model simulation to estimate both sampling variability within populations and between-population variability within conditions (see Figures A3 and A4 in the supplementary materials for the between-population result figures). In other words, we simulated multiple datasets from the same population matrix as well as multiple populations within each simulation condition (described in greater detail below). We varied the number of variables (p; 10 and 20), sparsity level (50% and 80%), and sample size (n; 100, 200, 250, 500, 1,000, 2,000, and 3,000) across conditions.

The simulation procedure to create each population matrix was as follows:

1. We used the PTSD dataset to form a bank of partial correlations
2. All partial correlations within the range of ± 0.05 were removed (following Epskmap's (2016) simulation procedure)
3. X partial correlations were randomly sampled, without replacement, from the bank to create the population partial correlation matrix where

$$X = \left(\frac{p(p-1)}{2}\right)(1 - sparsity), \tag{8}$$

$p \in \{10, 20\}$, sparsity $\in \{.50, .80\}$, and X is rounded to the nearest integer.

Thus, we had 2 (sparsity levels) by 2 (p) by 7 (n) conditions (i.e., 28 conditions). For each condition, 100 population matrices were created. For each population matrix we carried out the following procedure:

1. Simulate 1,000 multivariate normally distributed datasets of size n

2. With each dataset, estimate four networks using the four previously outlined penalty selection methods (i.e., CV, StARS, RIC, EBIC)
3. Compute performance measures (see Section *Performance Measures*)

Simulation 2: Simulated Partial Correlations

Simulation 2 assessed the effect of partial correlation size on method performance and more fully characterized the effect of sparsity. The edge weights were randomly generated from a G-Wishart distribution which is frequently used to simulate multivariate data as it can be defined with only two parameters $\mathbf{\Theta} \backsim W_G(df,\ I_p)$ where I_p represents a p by p identity matrix and df represents the degrees of freedom (Mohammadi & Wit, 2017). The degree of freedom parameter determines the degree of shrinkage towards the identity matrix. As the parameter increases the distribution of θ_{ij} narrows. In other words, as the degree of freedom parameter increases the partial correlations approach zero (Hsu, Sinay, & Hsu, 2012) through reduction of tail-heaviness (i.e., fewer extreme values). We adjusted the degrees of freedom to correspond to two ranges where 90% of the partial correlations on average fell between ± .35 and between ± .25.

Simulating from a G-Wishart distribution also guarantees a positive definite posterior estimate for $\mathbf{\Theta}$ (Kuismin & Sillanpää, 2016). To achieve a positive definite matrix, the partial correlations must become smaller as the network becomes more densely connected. As such, the same degree of freedom will result in smaller partial correlations when sparsity is at 50% compared to when it is at 80%. To account for this, we determined which degrees of freedom corresponded to the previously mentioned partial correlation ranges for each level of sparsity.

Sparsity varied from .1 to .9 in increments of .2. Sample size conditions were 100, 200, 250, 500, 1,000, 2,000, and 3,000, as in Simulation 1, and p was fixed at 20. In total, we had 70 conditions.

Performance Measures

We were interested in quantifying the performance of these methods with respect to both the accuracy of edge detection and the accuracy of edge weights. For edge detection we calculated the sensitivity (i.e., the true positive rate) and the false positive rate (FPR) as

$$Sensitivity = \frac{TP}{TP + FN} \text{ and } \text{FPR} = 1 - \left(\frac{TN}{TN + FP}\right) \tag{9}$$

where TP is the number of true positives (i.e., the number of estimated edges that are non-zero in the population matrix), FN the number of false negatives (i.e., the number of un-estimated edges that are non-zero in the population matrix), TN the number of true negatives, and FP the number of false positives detected by each method. Sensitivity and the FPR range from 0 to 1. A sensitivity score of 1 indicates that the method is correctly detecting all true positives, and a FPR of 0 means the method is correctly estimating all true negatives as exactly zero. A method with perfect edge detection would have a sensitivity score of 1 and FPR of 0. Finally, we compared the sparsity of the estimated network to population sparsity to assess whether the methods were sensitive to population sparsity.

A method with perfect edge detection may still estimate edge weights incorrectly. To capture the accuracy of the estimated edge, we calculated the correlation between the non-zero partial correlations in the population and the corresponding estimated edge weights. We refer to this performance measure as true edge correlation. Note, if an edge was non-zero in the population matrix but zero in the estimated matrix (i.e., a false negative) it was included in the

estimation of the true edge correlation as we were interested in the correlation between the estimated and population values of *true* edges.

Results

Simulation 1: Empirical Partial Correlations

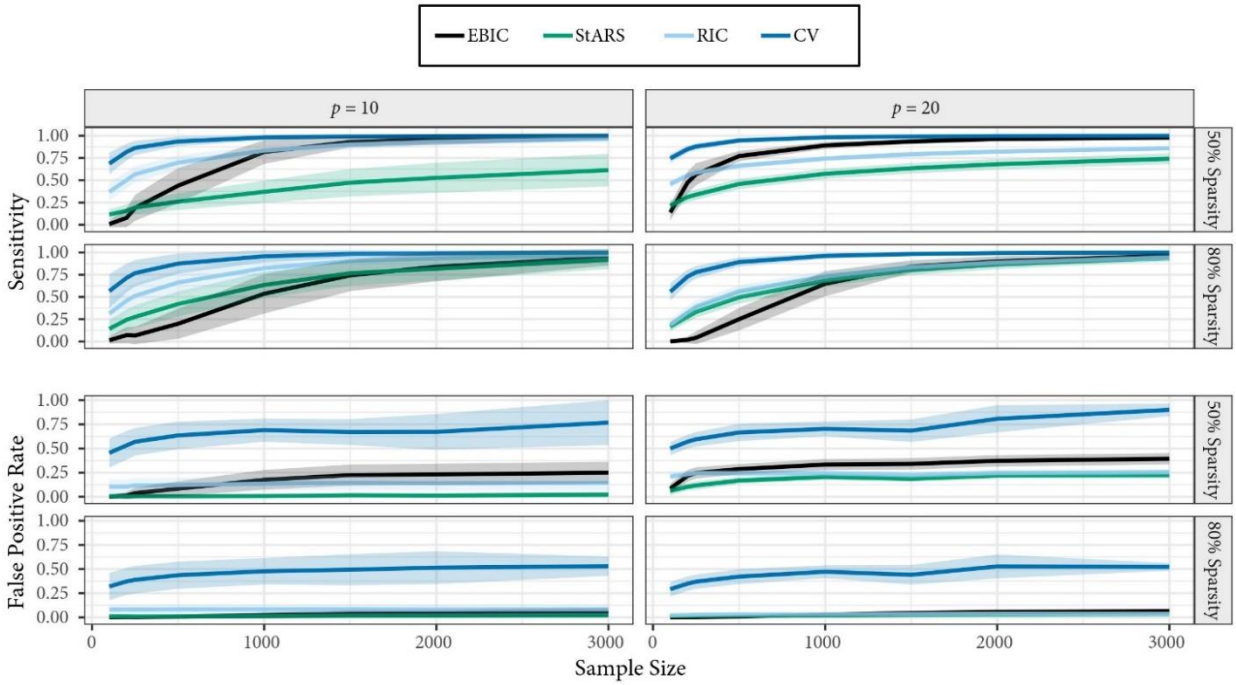
Sensitivity and FPR are presented in Figure 3. The performance measure is denoted on the y-axis. The columns and rows correspond to different simulation conditions (*p* and sparsity), and sample size is denoted on the x-axis. The four methods are depicted as different lines, and the shading around the lines represents ± one standard deviation of the outcome. The variability depicted in the figures represents the average sampling variability within a population. Figure A3 in the supplementary material depicts the between-population variability.

For smaller sample sizes ($n < 500$), EBIC often returns empty networks. This can be improved by decreasing the γ parameter. However, we set $\gamma = .5$ for all conditions per recommendations in the psychological network literature (Epskamp & Fried, 2018). As such, a number of networks with a sample size of 100 or 200 were empty (particularly $n = 100$). The other methods never returned an empty network. However, StARS and RIC in those same conditions often returned exceptionally sparse networks (i.e., < 10% of possible edges were estimated).

Sensitivity. Across both levels of *p* and both levels of sparsity, CV had far higher sensitivity than the other three methods. For sample size 500 and above, CV consistently had sensitivity rates of greater than 80%. EBIC, RIC, and StARS, in contrast, had low sensitivity at small *n*. For these three methods, although sensitivity improved as sample size increased, it did not reach 80% until sample size was greater than 1,000. For many conditions with a sample size of 500 or less,

sensitivity is under 50% particularly for StARS and EBIC. Sparsity did not have a great impact on sensitivity, but the size of the network did. When $p = 10$, all methods had higher sampling variability compared to $p = 20$, meaning the performance of a method was more uncertain. This was particularly true for EBIC. Greater variability when $p = 10$ was not unexpected as the performance metrics are computed as a proportion of the total number of edges. As such, when there are fewer potential edges each single edge has a greater influence on both the mean and variability of a performance index.

Figure 3. *Sensitivity and FPR Results from Empirical Partial Correlations Simulation*



Note. The columns (from left to right) correspond to the number of variables p (10 or 20), and the rows correspond to population sparsity (i.e., the percentage of edges equal to 0 in the population; 50% or 80%). The shading around each of the lines represents the average sampling variability within a population.

False Positive Rate. The FPR results largely mirrored the sensitivity results: methods and conditions with higher (better) sensitivity typically exhibited higher (worse) FPRs. CV had a markedly higher FPR across all conditions compared to the other three methods. Unlike

sensitivity, the FPR of all methods was impacted by sparsity. Specifically, there was an interaction between sparsity and sample size. Within the sparse condition (sparsity = 80%), the FPR was generally low and constant across sample size. Once sparsity decreased to 50% the FPR was higher and increased as sample size increased. For example in the $p = 20$ and $n = 500$ conditions, EBIC on average has an FPR of 5% when sparsity is 80% but the FPR rate increases to 60% when sparsity is 50%. The effect of sample size on the FPR may be explained by the decreased penalization of $\widehat{\boldsymbol{\Theta}}$ (i.e., smaller λs are being selected; see Figure A5 in the supplementary material for a visualization of selected lambda across sample size) as sample size increases, resulting in a denser estimated matrix. Note that StARS and RIC are more robust to this interaction. Like sensitivity, network size affected the variability of the results, but not the trends.

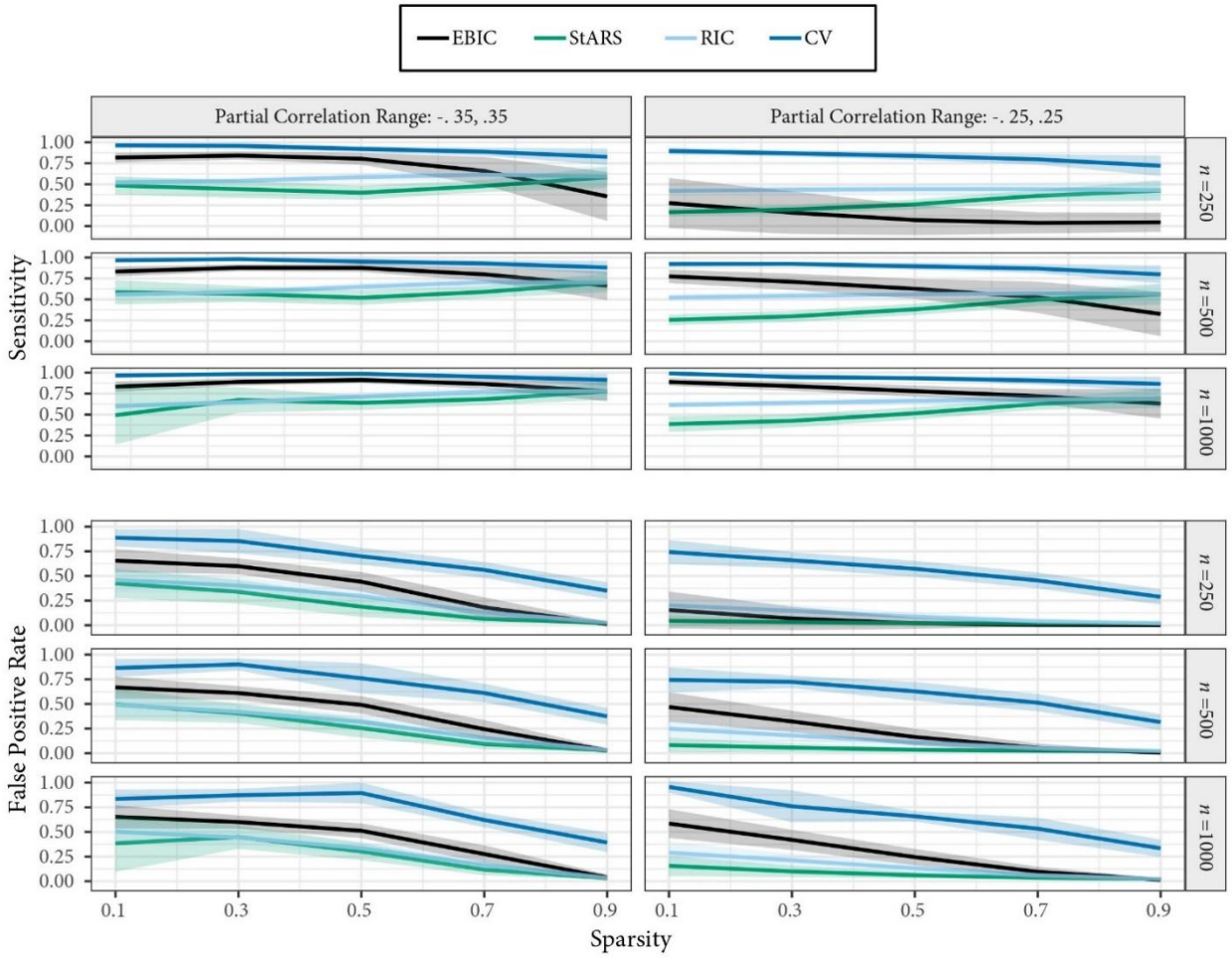
True Edge Correlation. The results for true edge correlation were similar to those for sensitivity (i.e., low correlation when $n > 1000$, greater variability at $p = 10$). Due to this similarity, all True Edge Correlation figures are included in the supplementary material (see Figure A6 in the supplementary material).

Simulation 2: Simulated Partial Correlations

Sensitivity and FPR are presented in Figure 4. Note, since we had five levels, sparsity is now denoted on the x-axis. The rows correspond to sample sizes. We selected three characteristic levels of sample size (250, 500, and 1000) to visualize, and the columns correspond to the average partial correlation ranges (±.35, ±.25) that contain 90% of the values. Each line depicts a different method, and the shading around the lines represents ±one standard deviation of the outcome.

Sensitivity. In agreement with Simulation 1, CV has the highest sensitivity across all conditions compared to the other three methods, and the sensitivity of all methods increases with sample size. Sparsity had an impact on methods in specific conditions but this effect was not consistent in direction across methods or conditions. For example, with larger sized partial correlations and a sample size of 1,000, sensitivity increased with sparsity for StARS and RIC but decreased for EBIC. In this condition, CV was not impacted. However, we noted a large effect of partial correlation size on sensitivity wherein conditions with larger partial correlations had higher sensitivity, particularly when sparsity was low. This is likely because larger effects require less power for detection.

Figure 4. *Sensitivity and FPR Results from Simulated Partial Correlations Simulation*



Note. The columns correspond to the different partial correlation ranges, and the rows correspond to different sample sizes. The x-axis denotes varying levels of population sparsity (i.e., the percentage of edges equal to 0 in the population). The shading around each of the lines represents the average sampling variability within a population.

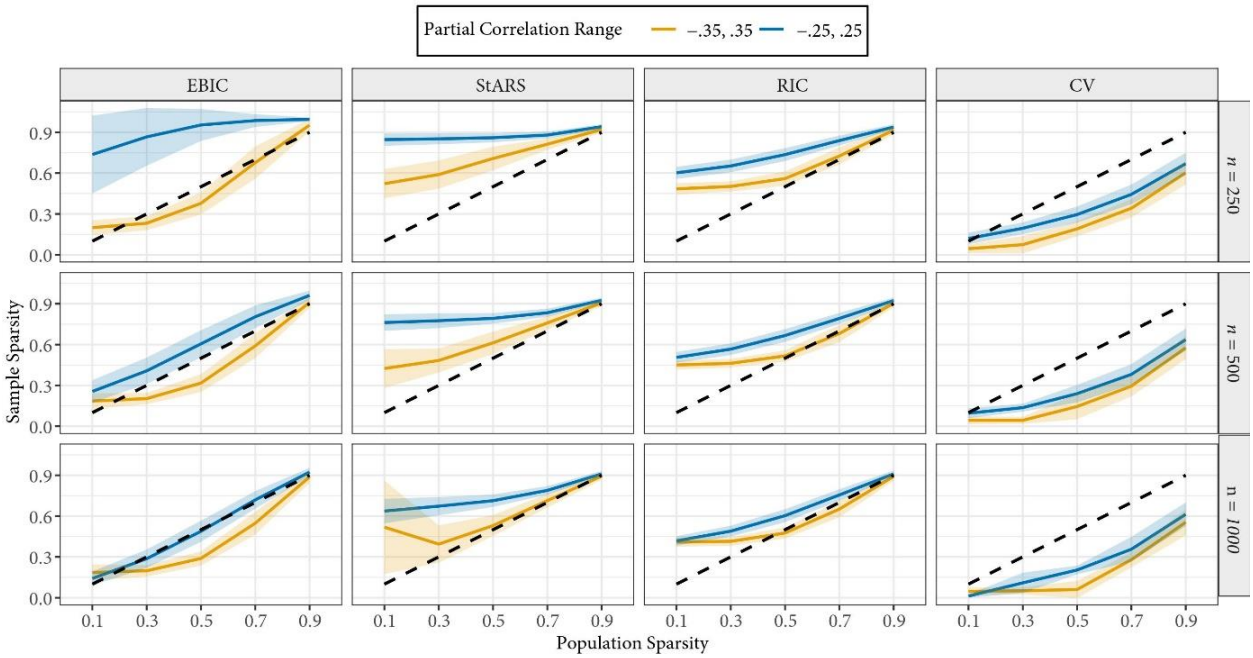
False Positive Rate. A number of the results from Simulation 2 were in agreement with Simulation 1. First, there was a trade-off between sensitivity and FPR wherein conditions with high sensitivity also had a high FPR, and those with lower sensitivity tended to also have a lower FPR. This trade-off extended to methods more generally wherein methods with high sensitivity tended to also have a high FPR (e.g., CV). Second, CV had a higher FPR across all conditions in

comparison with the other methods. Third, both sparsity and sample size had an impact on FPR. Specifically, the FPR increased as sparsity decreased and sample size increased.

Simulation 2 allowed us to investigate the effect of partial correlation size. Importantly, the effect of sample size and sparsity on FPR was moderated by the absolute partial correlation size. Notably, networks with smaller absolute partial correlations had a lower FPR across all conditions compared to those with larger absolute partial correlations. When on average 90% of the partial correlations were between ± 0.25, sample size and sparsity had less impact on the FPR, particularly for RIC and StARS. However, as the average partial correlation range increases, so does the FPR, and the interaction between sample size and sparsity is greater. Although this is the case for all methods, it is most apparent for EBIC and CV.

To explain the effect of partial correlation size, it is important to consider the assumptions and necessary conditions inherent in the glasso (for more details see the Discussion). For example, consistent model selection depends on the irrepresentable condition being fulfilled (Meinshausen & Bühlmann, 2006). This condition is more likely to fail as sparsity decreases. When this condition is not met, the true edges are excessively penalized, and more false positives are estimated increasing the false positive rate (Zhao & Yu, 2006).

Figure 5. *Population versus Estimated Sparsity.*



Note. The dashed line denotes the correct level of sparsity (i.e., equal sparsity between population and estimated sparsity). Lines represent different partial correlation ranges wherein on average 90% of the partial correlations are between the values (e.g., -.24, .25).

True Edge Correlation. True edge correlation (depicted in Figure A7 in the supplementary material) does not seem to be greatly impacted by sparsity. However, conditions with larger partial correlations tended to have higher true edge correlations, likely due to greater variance in the true edge weights. This finding is most pronounced for EBIC but observed across all methods.

Population versus Estimated Sparsity. Figure 5 depicts population sparsity on the x-axis and estimated sparsity on the y-axis. The rows represent sample size, and the columns represent different methods. The dashed line represents the accurate estimation of sparsity for reference (i.e., when the population and estimated network's sparsity are equal). Note that, RIC and StARS tended to underestimate the number of edges, although their performance improved as sparsity

and sample size increased. CV tended to overestimate the number of edges. EBIC tended to overestimate edge number in conditions with larger partial correlations, and underestimate in conditions with smaller partial correlations. Importantly, EBIC had high accuracy when sample size was large and the partial correlations were smaller. However, it is important to take into consideration the FPR. Even though there were conditions where the methods estimated the correct *number* of edges, it is likely many of the wrong edges were estimated.

Overall, it appears as though these methods were often not able to capture population sparsity. Given the impact of sparsity on method performance, it is important for researchers using the glasso to consider the population sparsity of the construct they are investigating (e.g., do I expect this construct to have many true connections?). However, the results comparing population to estimated sparsity show that estimated sparsity cannot be used to infer population sparsity. For example, a researcher cannot assume since the estimated network has 30% sparsity the population network must also be sparse. This is particularly true as these methods do not consistently under or over-estimate sparsity. Rather the magnitude of the population partial correlations, which are also unknown to researchers, impacts whether sparsity is under, over, or accurately-estimated.

Discussion

In this paper, we compared four methods for selecting the penalty parameter to compute the graphical lasso across conditions that are typical in psychology. We found that all methods had concerning performance features. First, all methods, with the exception of cross-validation, had low sensitivity at sample sizes that were most typical to psychology ($n < 1{,}000$). Second, we found that the sparsity of the population network, absolute partial correlation size, and sample size all impact the false positive rate. Notably, for networks with larger absolute partial

correlations, there was an interaction between sparsity and sample size wherein the FPR increased as sample size increased, particularly when the population network was dense. This interaction was not present or was attenuated for networks with smaller absolute partial correlations. Third, we found greater variability both within and between conditions and populations as sparsity decreased. This is important as it suggests that as true sparsity decreases, the uncertainty of the results increases. Finally, we found a consistent trade-off between sensitivity and the FPR when comparing both methods and conditions. For example, cross-validation had extremely high sensitivity (a desirable feature) but a high false positive rate (an undesirable feature). Additionally, conditions with larger absolute partial correlation sizes had high sensitivity but also a high false positive rate. The opposite was found for conditions with smaller absolute partial correlations.

When considering these results, it is important to note that regularization methods, including glasso, were developed for high-dimensional situations (i.e., $n << p$). In these cases, the inverse of the sample covariance matrix cannot be computed due to singularity: $\det(\mathbf{\Sigma}) = 0$. Regularization, by shrinking the variable space, allows for the estimation of a high-dimensional matrix. However, to accurately estimate the underlying network the assumption of sparsity and the irrepresentable and beta-min (i.e., non-zero coefficients must be sufficiently large) conditions must hold (Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006). Previous research has characterized method performance when these conditions are met (Fu & Knight, 2000), but these conditions have been found to rarely hold without explicit specification (Zhao & Yu, 2006). In our simulations, we did not ensure these assumptions were met. Rather we were guided by the sparsity and partial correlation sizes observed in psychological applications of network models.

Our results demonstrate that as true sparsity decreases (i.e., the population network increases in density) the FPR is higher and there is more variability in method performance both within and across conditions. This suggests that the methods are not viable when true networks are densely connected. Thus, it is important to consider population sparsity before using these methods. Unfortunately, none of the methods included in our simulations were consistently sensitive to population sparsity. That is, estimated sparsity is a poor indicator of true sparsity. Our results also indicate that when there are larger absolute partial correlations in the population network, the FPR increases. One potential explanation is the failure of the irrepresentable condition, which can result in an increase in false positives due to an over-penalization of true edges (Zhao & Yu, 2006). Overall, these results strongly suggest that researchers think about the possible properties of the network they are investigating to qualify whether a regularization method would be appropriate for their data (e.g., is it likely the population network is sparse?). This is important as there are many situations, particularly in low dimensions, where the costs of regularization supersede the benefits (Williams & Rast, 2020; Williams et al., 2019). It is also important to note that some of these performance issues can be attenuated. For example, sensitivity can be improved with a large sample size (n > 1,000) and the FPR can be improved with the use of thresholding (i.e., setting smaller edges to 0 after regularization) which has recently been advocated in psychological applications (Epskmap, 2018).

When deciding which penalty parameter selection method to use, it is important to consider the research goal. Given the consistent trade-off observed between sensitivity and the FPR, an important consideration is whether controlling the FPR or increasing sensitivity is more important to the research question at hand. If the goal is to obtain a dense network that contains the true model, then cross-validation, due to its high sensitivity, may be an appropriate method.

In this case, a high FPR is acceptable. Across our simulations, StARS consistently returned the sparsest network, so if the research goal is to return a network containing only the most important edges, then StARS, due to its low FPR, may be an appropriate method. In this case, low sensitivity is acceptable (see Table 3 for summary of each method's strengths and weaknesses). Finally, if both sensitivity and the FPR are equally important then using the RIC may be best as it tends to return the most balanced network. In other words, an RIC-selected network would tend to have greater sensitivity than StARS but lower than CV and a lower FPR than CV but higher than StARS.

Table 3. *Method Strengths and Weaknesses*

	Strength	**Weakness**
EBIC	Moderately High Sensitivity	Inconsistent Results
CV	High Sensitivity	High False Positive Rate
StARS	Low False Positive Rate	Low Sensitivity
RIC	Low False Positive Rate	Moderate Sensitivity

Another consideration for penalty method choice is whether the researcher is most interested in prediction or explanation. For example, a cross-validation method may be appropriate when the goal is prediction. However, when the goal is explanation, CV's high FPR may be prohibitive even though it is highly sensitive.

In this paper, we only assessed the performance of these methods on continuous data. Estimating networks from ordinal data is another ongoing area of research. In current practice, EBIC is sometimes applied to a polychoric correlation matrix computed from ordinal data (Epskamp, 2016). This practice would be tricky to generalize to the other three methods we

investigated, as they all require raw data to carry out their resampling or reshuffling procedures. It may be possible to adapt these procedures to ordinal data (e.g., by computing polychoric correlation matrices after resampling or reshuffling), but these extensions have not been tested.

In summary, EBIC seems to be greatly influenced by the characteristics of the population and data, more so than RIC and StARS, such as sparsity, sample size, and partial correlation size. Although RIC and StARS are influenced by these factors as well, it is to a lesser extent compared to EBIC. Further, RIC tends to strike the best balance between sensitivity and the FPR, compared to the other methods. As such, if glasso is being used to fit a psychological network, using the RIC to select λ may be more appropriate. However, given the concerning performance of all methods we evaluated, a non-regularized approach is likely best (Williams & Rast, 2020; Williams et al., 2019).

Conclusion

Network modeling is an important tool in psychological research. However, the estimation of networks is a rapidly evolving area of study. Here, we found that regularization approaches have concerning performance in conditions that are typical for psychological data, and we recommend using non-regularized estimation approaches instead. Importantly, our results underscore the need to carefully consider the likely population properties and the inferential goal before implementing an estimation method.

Network models are not the only methodological tool that has gained popularity in psychology recently. Causal inference has also become more commonplace due to psychologists' interest in understanding and explaining systems. But, like network models, much is still

unknown about building and estimating causal inference models. As such, the following chapter develops a framework for variable selection when the research goal is causal inference.

Chapter 3. Statistical Control Requires Causal Justification[4]

Psychological research that uses observational or quasi-experimental designs can benefit from statistical control to remove the effect of third variables—variables other than the target predictor and outcome—from an estimate of the causal effect that would otherwise be confounded (Breaugh, 2006; McNamee, 2003)[5]. Statistical control can lead to more accurate estimates of a causal effect (Pearl, 2009), but only when the right variables are controlled for (Rohrer, 2019)[6]. Although controlling for third variables is common practice (Atinc, Simmering, & Kroll, 2012; Bernerth & Aguinis, 2016; Breaugh, 2008), the selection of these variables is rarely justified on causal grounds.

In this chapter, we illustrate that controlling for an inappropriate variable can result in biased causal estimates. We begin by giving a brief introduction to causal inference and regression models, defining statistical control, and examining situations where statistical control is useful for researchers. We then highlight the pervasive issues surrounding how control variables are typically selected in psychology. We outline the assumptions required to justify controlling for a third variable—most importantly, that the control variable is a plausible confounder or lies on the confounding path. Next, we discuss the consequences of controlling for

---

[4] This chapter was slightly adapted from published paper: Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science*, *5*(2).

[5] It should be noted that many statisticians prefer the term "adjust" over "control" in the context of regression; The concern is that statistical control may be mistakenly conflated with (the stronger) experimental control (e.g., Gelman, 2019). Although we agree that "adjust" is more precise, we continue to use the term "control" as it is more prevalent in applied psychological research.

[6] We use the terms “accuracy” and “bias” to describe the relation of a population regression coefficient (*Y* regressed on *X*) to the average causal effect of *X* on *Y*. This use is different from how “bias” is used in statistics, where it describes the relation of an estimated statistic to its population quantity.

other types of third variables, including mediators, colliders, and proxies. We then discuss how longitudinal data can be used to deal with more complex models. Using an applied example, we provide practical recommendations for applied researchers working with observational data who wish to use statistical control to bolster their causal interpretations.

An Brief Introduction to Causal Inference

In this section, we provide a very brief introduction to the field of causal inference, including key concepts and definitions that are relevant for the present paper. The field of causal inference expands far beyond what we are able to cover here. We direct interested readers to (Dablander, 2020) for a slightly lengthier introduction, and to Pearl, Glymour, and Jewell (2016) or Peters, Janzing, and Schölkopf (2017) for book-length introductions.

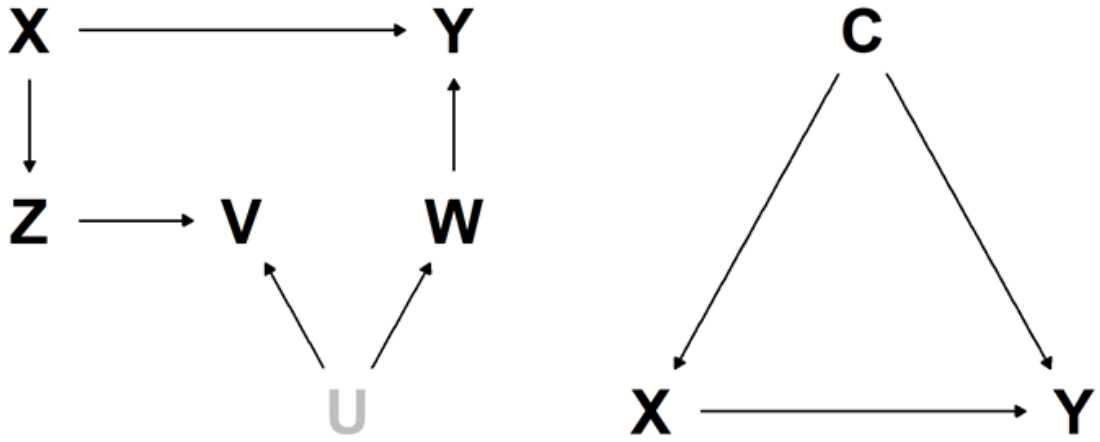
Causal inference involves estimating the magnitude of causal effects, given an assumed causal structure. We use Pearl's (1995) definition of causality, namely that X is a cause of Y when an intervention on X (e.g., setting X to a particular value) produces a change in Y. A causal effect—the expected increase in Y for a one-unit intervention in X—is *identified* when it is possible to derive an unbiased estimate of the causal effect from data. Estimating the magnitude of causal effects is key to understanding psychological phenomena; however, causal inference relies on theoretical assumptions that come from prior knowledge in addition to statistical information. Identifying a single causal effect requires accounting for and removing confounding effects without inducing spurious effects. As such, although it is common practice in psychological research to estimate and interpret multiple coefficients simultaneously (e.g., based on a regression model with six predictors), we will focus on the identification of a single causal effect at a time using a tool called a directed acyclic graph (DAG).

A DAG depicts hypothesized causal relations between variables (see Rohrer, 2018 for a detailed introduction to DAGs in psychology). The DAGs used throughout this paper contain three features: capital letters that represent variables, the letter *U* that represents a set of unmeasured variables, and arrows that represent causal effects. For example, Figure 6 (left panel) depicts that *X* is a cause of *Y*, *Z*, and, indirectly, *V* (*X* affects *V* through the mediating variable *Z*), and there is an unmeasured set of common causes that affects both *V* and *W*[7]. We use DAGs to represent causal relations in the population, and, in the same way, researchers can use DAGs to encode hypothesized causal systems and to inform decisions about which statistical models enable the identification of a causal effect. Importantly, DAGs are non-parametric; that is, they do not require any particular functional form. However, throughout this paper, we assume that all causal effects are linear and that each variable has a normally distributed residual.

In a DAG, a *path* is a sequence of arrows that connects one variable to another. A path connecting two variables will transmit association (i.e., the path results in the two variables being associated) unless there is a *fork* (i.e., two arrowheads coming together; e.g., $X \rightarrow Y \leftarrow W$) anywhere along the path. Each pair of variables may be connected by multiple paths and, if one of these paths transmits an association, then the pair of variables is expected to be associated[8]. In Figure 6 (left panel), *X* and *Y* are connected by two paths: $X \rightarrow Y$ and $X \rightarrow Z \rightarrow V \leftarrow U \rightarrow W \rightarrow Y$. The latter path does not transmit association due to the inverted fork, $Z \rightarrow V \leftarrow U$. But, the path $X \rightarrow Y$ does transmit association, and, as such, we expect *X* and *Y* to be associated in sample data drawn from a population represented by this causal graph.

[7] DAG figures 6, 8, 9, 10, and 15 were created in R using the package **ggDag** (Barrett, 2021). All result figures were created in R Version 3.6.2 (R Core Team, 2021) using the package **ggplot** (Wickham, 2016).
[8] Assuming two (or more) paths do not cancel each other out.

Figure 6. *Directed Acyclic Graphs (DAGs).*



Note. Left Panel: An example DAG with one set of unmeasured variables, U. Right Panel: A simple example of confounding.

The association information embedded in DAGs can help researchers discover potential threats to identifying a causal effect. We use *bias* to mean the discrepancy between a population parameter (e.g., a population regression coefficient) and the causal effect. In Figure 6 (right panel), there are two paths that connect *X* and *Y* and transmit association—both paths contribute to the association between *X* and *Y*. Crucially, one path, $X \leftarrow C \rightarrow Y$, is *noncausal*; that is, it is not part of the causal effect of *X* on *Y* so manipulating *X* does not change *C* nor does it change *Y* through *C*. As such, the association between *X* and *Y* is a biased estimate of the causal effect. In general, a common cause (i.e., a confounder) of a predictor and an outcome results in an association that is biased for the causal effect. To remove this bias, the common cause path (in this case, the path through *C*) must be removed from the estimated association. This can be done through an experimental research design where the predictor is randomized (Greenland, 1990), but experimental manipulation of psychological variables is often unfeasible. Thus,

psychologists have had to find another method to block confounding paths. One such method is statistical control, which can be accomplished via regression (McNamee, 2005).

Linear Regression and Statistical Control

In this section, we review how multiple linear regression produces coefficients that represent the linear association between each predictor and the outcome variable, conditional on the set of other predictors. To simplify the presentation, we assume all variables are standardized (with means = 0 and variances = 1). The linear regression model formulates an outcome variable, Y, as a linear function of a set of p predictor variables, X_1,..., X_p plus error:

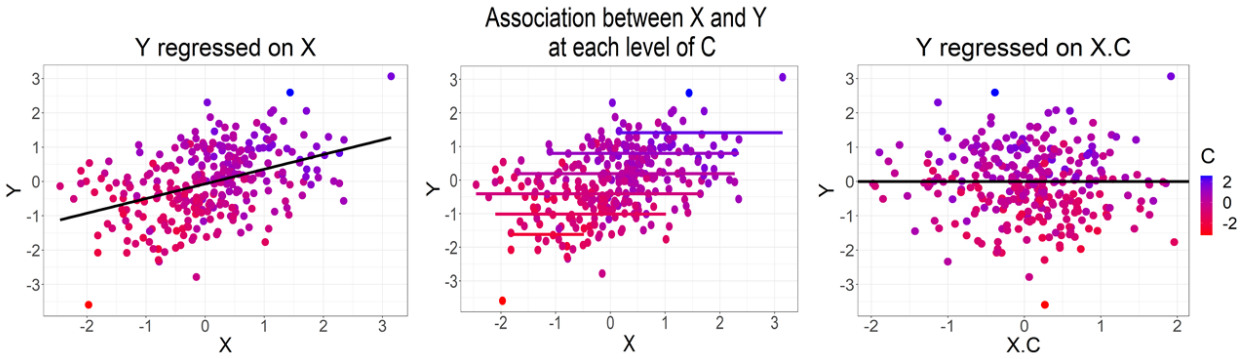
$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \quad (10)$$

In simple regression ($p = 1$), β_1 is equivalent to the correlation between X_1 (which for simplicity we denote as X in the $p = 1$ case) and Y (when all variables are standardized). Similarly, when $p > 1$ and all predictors are perfectly uncorrelated with each other, then the β for each predictor is equal to the correlation between that predictor and the outcome. In these special cases, each β reflects the total linear association between the predictor and Y and can be interpreted as the expected change in Y for a one-unit change in X.

When the predictors are correlated with each other, the estimation and interpretation of regression coefficients is more complicated. In these cases, the variance shared among predictors gets partitioned among the regression coefficients, such that each regression coefficient represents the expected change in Y for a 1-unit increase in X, holding all other predictors fixed. Conceptually, this approach is the statistical equivalent of sampling participants who have the same value on all but one of the predictors, and estimating the association between that one

predictor and *Y* in the sample. Thus, multiple regression coefficients are known as partial regression coefficients because they represent the isolated association between a single *X* and *Y* when none of the other predictors are changing. If the predictors other than *X* represent the full set of variables that confound the *X*-*Y* association and if there is no reverse causality (i.e., *Y* does not cause *X*), then the causal effect ($X \rightarrow Y$) is identified using the partial regression coefficient.

Figure 7. *Visualizing Statistical Control*



Note. Left Panel: the confounded regression of *Y* on *X* reveals a strong linear association. The color of individual points represents their values on the confounding variable, *C*. Middle Panel: Within each level of *C*, the association between *X* and *Y* is 0. Right Panel: When *C* is included as an additional predictor in the regression equation of *Y* on *X*, its confounding influence is removed to reveal only the partial association between *X* and *Y*. *X*.*C* is the residual scores that are obtained when *X* is regressed on *C*.

Figure 7 illustrates statistical control. The scatterplot on the left shows a strong linear association between *X* and *Y*, and the color of the points in this plot represents values of a third variable, *C*. *C* is correlated with both *X* and *Y*, raising the possibility that it may be a confounder. The middle plot shows the population-level regression lines that one would get if it were possible to compute the simple regression coefficient of *Y* on *X* for each subpopulation with a fixed value on *C*. In practice, however, there is not enough information in this small dataset to accurately estimate the regression coefficient within each subpopulation sample. The plot on the right shows the statistical approach to do this; that is, the de-confounded association between *X* and *Y*.

The x-axis now represents the residuals that are obtained when *X* is regressed on *C*; that is, the part of *X* that is independent of *C* ($X = \beta_1 C_1 + X.C$). When *Y* is regressed on this *residualized* predictor, $Y = \beta_2 X.C + \varepsilon$, there is little remaining relation between *X.C* and *Y* and β_2 is close to 0. Adding the control variable *C* to the regression of *Y* on *X* ($Y = \beta_2 X + \beta_3 C + \varepsilon$) is equivalent to regressing *Y* on the residualized *X*—the value of β_2 is the same in both equations.

By statistically controlling for the correct variable, a confounding effect can be removed from an estimate, making statistical control a valuable tool for researchers who are interested in causal inference and have access to observational or quasi-experimental data (Morabia, 2011; Pourhoseingholi, Baghestani, & Vahedi, 2012). Here, we focus on controlling for the correct variable(s), but obtaining an unbiased association depends on several additional assumptions: (1) Any interactions or non-linear effects must be specified correctly (Cui, Guo, Lin, & Zhu, 2009; Simonsohn, 2019), (2) the predictor and control variables must be measured without error or a model that deals with the measurement error must be used (Savalei, 2019; Westfall & Yarkoni, 2016), (3) the relevant variables must be measured at a time when the causal process can be captured.[9]

Causal inference is not the only reason that a researcher may choose to control for a third variable. Controlling for a variable that shares variance with the outcome but not the predictor will decrease the amount of residual variance in the outcome which, in turn, lowers the standard error of the estimated regression coefficient and increases power (Cohen, Cohen, West, & Aiken, 2003). Control variables are also sometimes used to establish that a new measure is uniquely

[9] Some research questions can only be answered when the set of variables are measured at a specific interval. For example, if the research question pertains to how asking questions in class impacts one's final grade, then the 'question' variable should represent the number of questions each student asked across the entire class, rather than just a single week.

predictive beyond some already established measure (Wang & Eastwick, 2020; we discuss this practice in Box 3). Researchers may also be interested in using the partial regression coefficients to describe partial associations, rather than using them to explain psychological processes.

Frequently, though, researchers aim to develop and test theories of psychological processes, and this endeavor almost always involves making and testing causal hypotheses. Although it is rare that causal inference is explicitly acknowledged as the goal in non-experimental studies (Grosz et al., 2020), a key component of theory building is proposing a set of principles that explains a process and then formulating a model based on these principles (Borsboom et al., 2021a)—in other words, positing a set of causal hypotheses (Shmueli, 2010; Yarkoni & Westfall, 2017). When coefficients are interpreted as reflecting the strength of a causal effect, then selecting and controlling for the correct variable is of the utmost importance because controlling for the wrong variable can increase, rather than decrease bias.

Common Practices: How Do Researchers Typically Choose Control Variables?

Reviews of published studies in psychology journals have found that over 50% of the studies reviewed gave no justification for the inclusion of specific control variables (Becker, 2005; Bernerth & Aguinis, 2016; Breaugh, 2008). Moreover, Atinc and colleagues (2012) and Carlson and Wu (2012) noted that, when researchers did justify their control variables, they typically did so by noting the statistical association between the predictor and the control (e.g., the predictor and third variable correlate at .40, so it is appropriate to control for the third variable).

Psychological researchers also receive relatively little helpful advice about what constitutes strong evidence for inclusion of a control variable in their area of research. The

available advice is often too vague (e.g., "offer rational explanations, citations, statistical/empirical results, or some combination"; Becker, 2005), too minimalistic (e.g., use a theoretical model to motivate control variable selection, Breaugh, 2006), or is of little practical use (e.g., provide evidence that control variables are accomplishing their intended purpose; Carlson & Wu, 2012). The handful of articles that do focus more specifically on the need to explicate relations between the control, predictor, and outcome provide some helpful guidance, but it can be difficult to know how to implement this guidance in one's own line of research. For example, Meehl (1970) noted that controls should not be automatically considered exogenous, and instead, researchers must consider the possibility that other important variables in the model—the predictor or outcome—might impact the control. Additionally, others have argued that researchers should outline the theory behind their decision to include/exclude control variables (Berneth & Aguinus, 2016; Edwards, 2008). However, on their own, these calls to integrate theory may be difficult to implement. Fortunately, recent work on the importance of causal language and causal thinking in psychology (Dablander, 2020; Grosz et al., 2020; Rohrer, 2018) suggests that one way to implement these calls for proper control variables is to give more consideration to how control variables are *causally* linked to the other variables in the model.

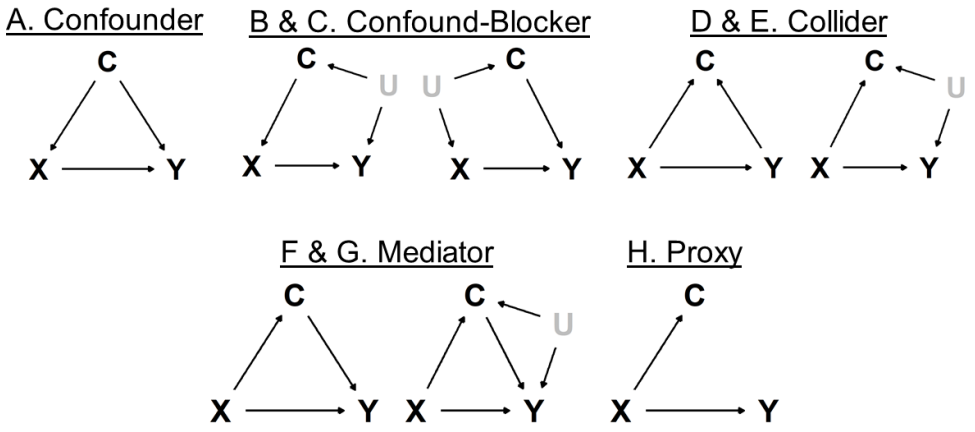
As we will show in the coming sections, when a central goal of an analysis is to learn about a process, the only way to qualify a variable as a good control is to consider the causal model connecting the control(s), predictor, and outcome. In the following sections, we show how, depending on the causal status of the third variable and the strength of its causal relations with both the predictor and outcome, controlling for the third variable can either remove or add substantial bias to the estimate of the causal effect. In doing so, we present a framework for principled control variable selection and justification.

One Step Forward, Two Steps Back: Controlling for the Wrong Variable

Though it is possible to remove bias from an estimate of a causal path by controlling for a confounding variable, it is also easy to *add* bias to an estimate by controlling for a variable that is either not a confounder or does not block a confounding path (VanderWeele, 2019). In the following paragraphs we describe different kinds of variables and discuss the consequences of controlling for each. Our goal is not to give a comprehensive list of all variables that one might possibly control for (see Cinelli et al., 2022 for a more comprehensive list of "good" and "bad" control variables), but rather to underscore how the impact of statistical control depends on the type of control variable. Figure 8 depicts eight types of third variables (*C*) that differ in their causal relation to the predictor (*X*) and outcome variable (*Y*). Panel A shows a confounder and panels B & C show two "confound-blockers", all of which can remove bias when controlled for.[10] Panels D-H show colliders, mediators, and a proxy, all of which are problematic when controlled for (Cinelli et al., 2022; Elwert & Winship, 2014; Pearl 2009; Rohrer 2018).

[10] Bias amplification can occur when, after controlling for a confounder or an instrumental variable (especially one that is strongly correlated with *X* and only weakly correlated with *Y)*, the bias from a second uncontrolled confounding path can increase. If the bias removed by controlling for one confounder is less than the amount of bias induced by bias amplification, the overall bias can increase (Pearl, 2011; Steiner & Kim, 2016).

Figure 8. *Example Causal Models*



Confounder and Confound-Blocker

A confounder is a variable that is a (direct or indirect) cause of both *X* and *Y*, (see Figure 8A).[11] By controlling for a confounder, we can block the confounding path that obscures the causal effect of *X* on *Y*. But it is also possible to block the confounding path by controlling for any other variable that lies on that path. We call such a variable a *confound-blocker*, because it is not itself a confounder but controlling for it nevertheless blocks the confounding path. For example, to estimate the causal effect of coffee on concentration, it may be important to control for the confounding effect of sleep (because less sleep may lead to both greater caffeine consumption and lower concentration). This confounding path can be blocked either by measuring and controlling for the confounder itself—hours of sleep—or by measuring and controlling for another variable along the confounding path (e.g., desire for coffee). In panels B and C, the confounder is unmeasured, but controlling for *C*, a confound-blocker, de-biases the association.

[11] Some other definitions of a confounder (e.g., VanderWeele & Shpitser, 2013) do not draw a distinction between confounders and confound-blockers.

Collider

When two variables share a common effect, the common effect is called a collider between that pair of variables (Figure 8D & E). For example, both IQ and hard work can result in getting accepted to college, so college student status is a collider between IQ and hard work (Elwert & Winship, 2014; Rohrer, 2018). Controlling for a collider will induce a spurious (i.e., non-causal) association between the variables that are causes of the collider. For example, if regressing hard work on IQ produced a simple regression coefficient of zero, controlling for college student status (which is positively affected by both hard work and IQ) would induce a spurious negative effect between these variables. A variable that is a collider for a pair of variables other than the outcome and predictor can still bias the target causal estimate. As controlling for a collider induces a spurious association between its causes, this can transform a path between the predictor and outcome from a path that does not transmit association to a path that does. For example, in Figure 8E, *C* is a collider for *X* and *U*. When *C* is not controlled for, the non-causal path from *X* to *Y* ($X \rightarrow C \leftarrow U \rightarrow Y$) does not transmit an association due to the inverted fork ($X \rightarrow C \leftarrow U$). But controlling for *C* induces a spurious association between *X* and *U*, and the new non-causal path from *X* to *Y* ($X - U \rightarrow Y$; where $X - U$ denotes a spurious association) now transmits association, resulting in a biased causal estimate.

Mediator

A mediator is a variable that is caused by *X* and is a cause of *Y* (Figure 8F & G; Baron & Kenny, 1986; Hayes, 2009; Judd & Kenny, 1981). For example, sleep problems might mediate the relation between anxiety and tiredness, such that sleep problems are a mechanism by which anxiety increases tiredness. If a researcher is interested in the total effect of the predictor ($X \rightarrow Y$

plus $X \rightarrow C \rightarrow Y$) on the outcome (compared to only the direct effect, $X \rightarrow Y$), then controlling for a mediator will undermine this effort by blocking one causal path of interest. Even if a researcher is only interested in the direct effect, controlling for a mediator could induce bias if the mediator and the outcome share a common cause (Figure 8G). When such a common cause exists, the mediator is a collider for the predictor and this common cause, and because the mediator is being conditioned on, the non-causal path, $X \text{—} U \rightarrow Y$, now transmits association and biases the estimate (Rohrer et al., 2022).

Proxy

A proxy is caused by X and has no causal relation to Y (Figure 8H; Pearl, 2009). Note this does not mean the proxy is a 'good' or sensical measure of the predictor. For example, GPA and number of cars owned might be proxies for cognitive ability, such that cognitive ability is a cause of GPA and (indirectly via income) cars owned. The number of cars owned, however, is likely a poor measure of cognitive ability.

If the predictor is a perfectly reliable variable (i.e., X contains no measurement error), controlling for a proxy will not affect the estimate of the $X \rightarrow Y$ path: the regression coefficient of X on Y will capture the causal effect and the coefficient of the proxy regressed on the outcome will be zero (in the population). But if X is in fact an unreliable measure of the true causal variable (e.g., cognitive ability is measured with a test that is not perfectly reliable), then controlling for a proxy will attenuate the estimated causal path of interest. This attenuation effect arises because the proxy can be understood as a second unreliable measure of the same underlying predictor (e.g., GPA and the unreliable cognitive ability test are both measures of cognitive ability). When both the predictor and control variable are unreliable measures of the same construct, the true predictive effect of the construct gets partitioned into two coefficients,

neither of which capture the full causal effect. The magnitude of the attenuation depends on the strength of the paths from the true (latent) predictor to the measured predictor and to the proxy.

Inappropriate Control Leads to Bias: Demonstrating the Importance of the Causal Structure

In the following section we demonstrate how the causal structure influences the partial regression coefficients. Each figure displays the consequences of controlling for a third variable given a range of hypothetical population models. We used path tracing (i.e., Wright's Rules; Alwin & Hauser, 1975; see section *Path Tracing* in the Supplementary Material for an example) to obtain a population correlation matrix for each causal structure, and calculated regression coefficients from each population correlation matrix, using the formula:

$$\beta = \Sigma_{xx}^{-1}\Sigma_{xy} \tag{11}$$

where $\boldsymbol{\Sigma_{xx}}$ is the p x p correlation matrix of predictors and Σ_{xy} is a p x 1 vector containing correlations between each predictor and the outcome. [12] We compare the partial regression coefficient (when a variable is controlled for) to the simple regression coefficient (when no other variables are controlled for). We show the population values of simple and partial regression coefficients under three of the causal structures depicted in Figure 8: when the third variable is a (1) confounder, (2) mediator, and (3) collider for the predictor and outcome (see Figure A9 in the Supplementary Material for results from all models shown in Figure 8). We assume, for the time being, that all variables are measured without error (in the section, *Measurement Error Makes Proxy Variables Problematic*, we relax that assumption and look at the impact of controlling for

[12] The code used to calculate the coefficients is available at https://osf.io/64rfv/?view_only=f49974350af14a7185994eeb00374306.

a proxy). For each population model, the direct effect of *X* on *Y* is set at .15, and the direct effect between *C* and *Y* is set at .5 (these are standardized values). The direct effect between *X* and *C* varies to demonstrate the variability of the bias within a population model.

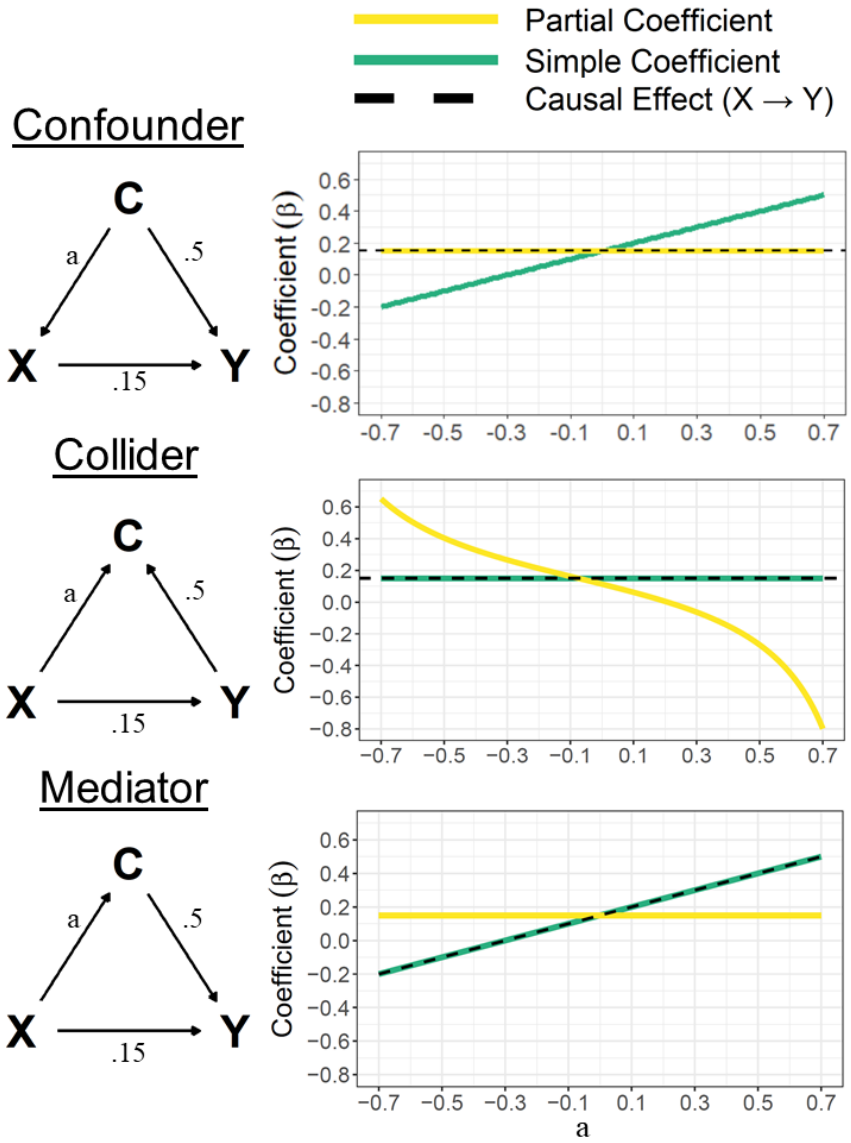
Results are shown in Figure 9. The top row shows the effect of failing to control for a confound: The yellow line depicts the effect of *X* on *Y* (which accurately estimates the causal effect), controlling for the confounder *C*, whereas the green line depicts the overestimated or underestimated value when *C* is not controlled for. The second and third rows of Figure 9 show two situations in which controlling for a non-confounder introduces bias. When the control variable is a mediator, as the absolute strength of the effect of *X* on *C* increases, the total causal effect increases as well, but the partial coefficient remains the same, resulting in an increasing discrepancy between the coefficient and the causal effect. When the control variable is a collider for the predictor and outcome variables, as the absolute strength of the effect of *X* on *C* increases, the discrepancy between the simple and partial coefficients increases as well. It is important to note that the direction of the bias that arises when controlling for a collider or a mediator depends on the parameters of the model. Without knowing the true causal effect values (and we may safely assume that these are unknown!), the impact of controlling for a non-confounder is unpredictable. As such, if a researcher is unsure whether the variable that they plan to control for is either a confounder or a variable that blocks the confounding path, they should not interpret the resulting partial coefficient as a conservative approximation of the true causal effect.

Measurement Error Makes Proxy Variables Problematic

Measurement error can further muddy the interpretation of controlled regression coefficients. In the presence of measurement error, simple regression coefficients (confounded or not) will be attenuated (Shear & Zumbo, 2013), and controlling for an imperfectly-measured

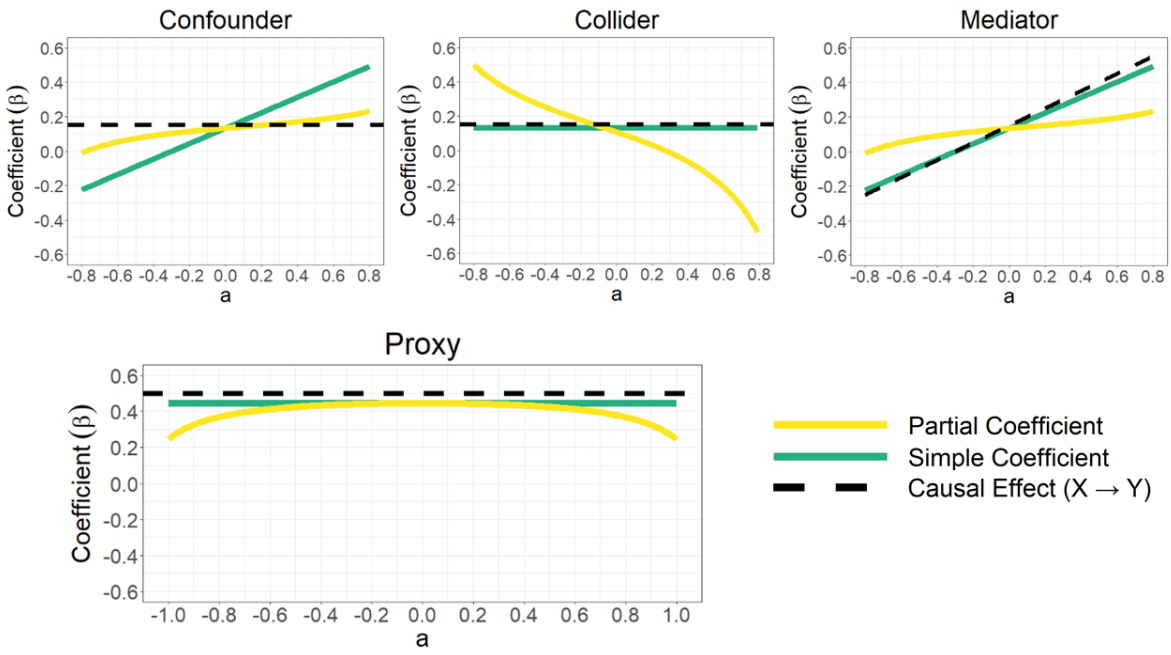
confound may not remove its full confounding effect (Westfall & Yarkoni, 2016). Figure 10 shows results from the same series of three models as Figure 9 plus the proxy model where both *X* and *C* have 80% reliability. The path values connecting each of the variables are kept the same as Figure 9 except that we set the direct causal effect of *X* on *Y* in the proxy model to .5 instead of .15, to more clearly show the attenuation effect that ensues as a result of controlling for a proxy. As Figure 10 shows, controlling for a proxy is problematic when the predictor and proxy are highly correlated. As described earlier, when *X* is measured with error (so that the observed variable is X_m, rather than *X* itself), X_m and *C* can be seen as two imperfect measures of the same true construct (with *C* being a much weaker measure than X_m to the extent that the causal path from *X* to *C* is smaller than the path from *X* to X_m), and they each are able to account for some of the true causal effect.

Figure 9. *Partial and Simple Regression Coefficients under Three Causal Structures*



Note. In each graph, the x-axis depicts the population value of the direct effect connecting the control variable and the predictor (denoted by label a), and the y-axis depicts the value of the regression coefficient of *Y* on *X*. The direct effect of *X* on *Y* and the value of the direct effect connecting *Y* and *C* are held constant across the results. Solid lines represent the partial (yellow line) and simple (green line) regression coefficients. The dashed line represents the total *X*→*Y* causal effect.

Figure 10. *Controlled and Uncontrolled Regression Coefficients When Variables are Measured with Error.*



Note. In each graph, the x-axis depicts the population value of the direct effect connecting the control variable and the predictor (a), and the y-axis depicts the value of the regression coefficient of Y on X_m. The direct effect of X on Y and the value of the direct effect connecting Y and C are held constant across the results. Solid lines represent the partial (yellow line) and simple (green line) regression coefficients. The dashed line represents the total $X \rightarrow Y$ causal effect. Predictor and control variables are measured with error (reliability = .8) so neither the simple nor partial coefficients capture the causal effect.

Each of the causal structures in Figures 9 and 10 produces a correlation between the third variable and both the predictor and the outcome variable. In fact, the very same correlation matrix (and thus, the very same set of regression coefficients) could be produced by every one of these models. As such, these correlations alone cannot reveal whether the third variable is, for example, a mediator or a collider for the predictor and outcome variable (Maxwell & Cole, 2007; Pearl, 1998). Evidence of a statistical association among a third variable, predictor, and outcome merely implies that there is some causal structure that connects these variables (either directly or via a set of unobserved variables)—but the confounder structure is just one possibility among

many. Therefore, it would be a mistake to assume that a variable should be controlled for, merely on the grounds that it is correlated with both the predictor and outcome.

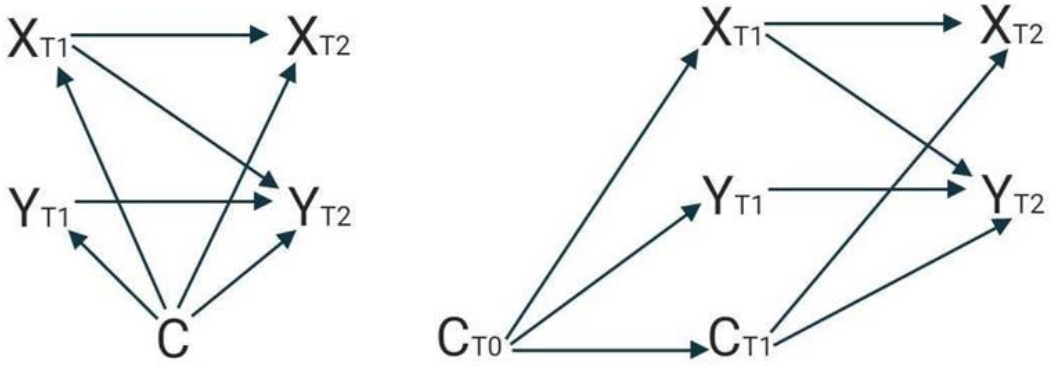
More Complicated Models and Longitudinal Data

In the previous section we used simple causal structures to show how bias arises, but often the true causal diagrams are more complicated. For instance, a causal effect may be confounded by a large set of variables, of which many are unmeasured. Measuring all confounders is not necessary if there is a more proximate variable through which many (or all) of the confounders influence the outcome or predictor. Controlling for such a variable would block all the confounding paths where it functions as a mediator (between the confounder and the outcome or predictor) without having to measure or control for the confounders themselves.

Another complicated situation is when a potential control variable occupies two roles. For example, a variable may act as a confounder between two other constructs if measured at one time point and a mediator if measured at another; however, it is not necessarily possible to make this distinction for the same instantiation of the predictor and outcome (see Box 1). Because the goal of statistical control is to remove a confounding effect without blocking the causal effect, it is important for a researcher to identify and measure the control variable at a time when it serves as a confounder between the predictor and outcome, rather than a time when it serves as a mediator. Box 1 explains in more detail what we mean and gives an example of how to do this. Similarly, if there were bi-directional causality between a control variable and the outcome, the control variable would be a confounder if measured at some time points and a collider for the predictor and outcome variables at others. When such complicated structures exist, it can be difficult to obtain a set of control variables that de-biases a causal effect. Often, longitudinal data can help with this endeavor.

Longitudinal data—when the same variables are measured at multiple measurement occasions in the same individuals—provide information about the temporality of variables. These data, along with the use of longitudinal models, can be used to make propositions about the location and direction of effects. For example, if X_{T1} (where the subscript denotes the measurement occasion) is found to predict Y_{T2} even after controlling for previous measurements of *Y*, then *X* is said to Granger-cause *Y*. Granger causality depends on two criteria: 1) the Granger-cause precedes its effect, and 2) the Granger-cause explains unique variation in its effect over and above what is predicted by a previous measure of *Y* (Granger, 1980; Maziarz, 2015). Establishing Granger causality is not the same thing as establishing causality, however (Eichler & Didelez, 2010). Effects that meet the definition of Granger causality may still be influenced by confounders, because controlling for a previous version of *Y* may not block all confounding paths. For example, in Figure 12 (left panel) controlling for Y_{T1} blocks one of the confounding paths ($X_{T1} \leftarrow C \rightarrow Y_{T1} \rightarrow Y_{T2}$) but not the other ($X_{T1} \leftarrow C \rightarrow Y_{T2}$). Longitudinal data can make it much easier to control for confounders, but it does not negate the need to clearly justify the underlying causal structure (Rohrer, 2019).

Figure 11. *Examples of Time-invariant and Time-varying Confounders.*



Note. Left Panel: The confounder, *C*, does not change across the two measurements. There are two confounding paths: 1) $X_{T1} \leftarrow C \rightarrow Y_{T2}$ 2) $X_{T1} \leftarrow C \rightarrow Y_{T1} \rightarrow Y_{T2}$. Controlling for Y_{T1} only

blocks the second confounding path. Right Panel: The confounder does change across measurements. Because an effect precedes its cause, the subscript for the first version of the confounder is 0. Again, there are two confounding paths: 1) $X_{T1} \leftarrow C_{T0} \rightarrow C_{T1} \rightarrow Y_{T2}$ 2) $X_{T1} \leftarrow C_{T0} \rightarrow Y_{T1} \rightarrow Y_{T2}$. Controlling for Y_{T1} only blocks the second confounding path.

Combined with a justified causal structure, longitudinal data can address some complicated causal structure problems. First, by repeatedly measuring a pair of variables that influence each other ($X \rightarrow C$ & $C \rightarrow X$) at the correct interval, the once bi-directional paths become uni-directional paths (e.g., X_{T1} is a cause of C_{T2} and C_{T1} is a cause of X_{T1}). [13] Second, measuring the same variables across time allows for the removal of a specific kind of unmeasured confounding, namely *time-invariant* confounding. A time-invariant confounder (Figure 12 left panel) is a confounder whose level and effects do not change across the measured time points (e.g., ethnicity—assuming the effect of participants' ethnicity on the other variables in the model is constant across measurement occasions). In contrast, a time-varying confounder (Figure 12 right panel) is a variable whose level or effect changes between measured time points (e.g., positive affect, relationship satisfaction).

One way to remove unmeasured time-invariant confounds is to use a fixed effects model (Allison, 1994; Kim & Steiner, 2021). [14] A fixed effects model estimates an individual-specific intercept, which captures all effects that vary between but not within individuals in a sample (or another unit of analysis such as school or country). Because time-invariant confounds do not

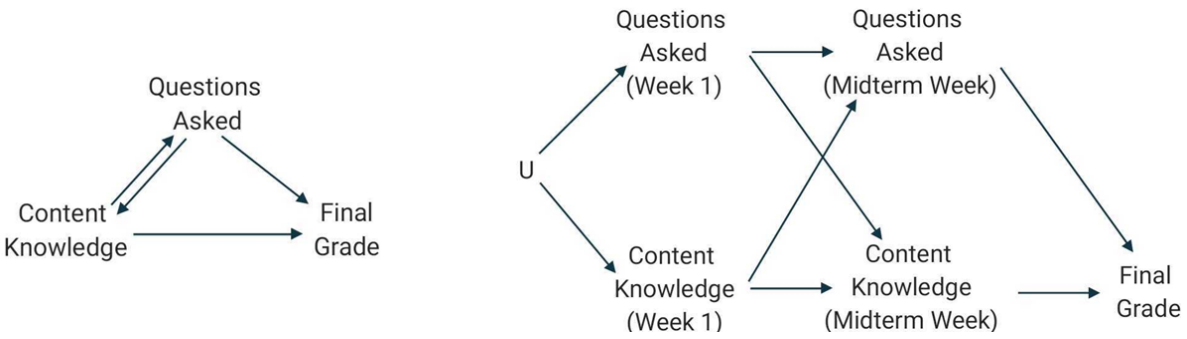
[13] The interval at which a set of variables should be measured depends on the causal effect of interest (Lundberg, et al., 2021). For example, the effect of high school mentorship on post-college income is a different causal effect than the effect of college mentorship on post-college income. Each of these causal effects would require measuring mentorship at a different interval.
[14] Two other methods of removing time-invariant confounding is by estimating mean deviation scores and gain scores. See Kim and Steiner (2020) for a comparison of these methods.

Box 1. Directional Causality

Imagine a researcher is interested in the effect of student content knowledge on final grades, and the researcher posits that the number of questions a student asks is a potential confounder for this effect. However, the researcher recognizes that the number of questions a student asks could also be a mediator. Therefore, it is plausible that (1) how much a student knows could influence how many questions they ask in class, but also (2) how many questions a student asks in class could also influence how much they know about class content. As such, there may be bidirectional causality between the number of questions asked in class and the amount a student knows about class content. Figure 11 (left) shows one depiction of such a causal model. But, this figure is misleading as it suggests that course knowledge could be a cause of questions asked and, in the same instantiation, questions asked could be a cause of course knowledge. This is impossible as causality happens across time. As such, one of these variables—either content knowledge or questions asked—had to occur before the other, excluding the previous variable from being caused by the latter variable. Therefore, an appropriate DAG should depict this bidirectional effect across different instantiations of content knowledge and questions asked (see Figure 11 right panel). For example, knowledge at week 1 of the course may be a cause of the number of questions asked in midterm week, and questions asked at week one may be a cause of knowledge at midterm week. Specifying the DAG with different instantiations of the same variable allows researchers to query which causal effect they are interested in (the effect of content knowledge at week 1 on final grade or the effect of content knowledge at midterm week on final grade) and explore which instantiation of the control variable blocks a confounding path.

Box 1 (Continued). Directional Causality

Figure 12*. Incorporating Multiple Instantiations of a Variable into a DAG*



vary within individuals, the fixed effects model can de-bias a coefficient for unmeasured time-invariant confounders. A fixed effect can be estimated by including a dummy variable in the regression model for each participant (see Hanck et al., 2017 and Colonescu, 2016 for tutorials on estimating a fixed-effects model in R). Rather than controlling for a specific variable, the fixed-effects model simultaneously controls for all the attributes of individuals that do not vary over time. Although fixed-effects models remove the effects of unmeasured time-invariant confounders, they do not remove the effects of *time-varying* confounders. Therefore, considering whether time-varying confounders exist for a pair of variables and making a plan, if needed, to deal with them is still important when estimating a fixed effects model.

In summary, the causal structure must still be proposed and justified even if longitudinal data are available (Imai & Kim, 2016; Hernán et al., 2002) and the only way to provide a valid argument that it is appropriate to control for a particular variable is to discuss the causal structure that includes the control(s), predictor, and outcome. This argument, which should be presented for each control variable, can include empirical motivation (e.g., an estimated association from a

previous study) but it must be justified on a theoretical basis (outlined below). This strategy aligns with advice from psychometricians to be conservative with the number of control variables (Becker, et al., 2016; Carlson & Wu, 2012) and to carefully justify each one (Bernerth & Aguinis, 2016; Breaugh, 2006; Carlson & Wu, 2012; Edwards, 2008). Of course, proposing a full causal map of relations among all the variables in one's model is more difficult than using the default methods of selecting control variables. To assist in this difficult but necessary endeavor, we next use an applied research example to demonstrate how to decide which variables to control for using causal reasoning.

Using the Causal Structure to Justify Control Variables: An Applied Example

In this section, we outline the steps to properly justify control variables based on causal structures. As a demonstration, we use a simplified, applied example from the literature on personality and work.

Selecting Variables

Begin with a pair of outcome and predictor variables, where the predictor variable is hypothesized to be a cause of the outcome.[15] We chose conscientiousness or the disposition to be hard-working, responsible, and organized, as the predictor and career success, including annual income, occupational prestige, and job satisfaction, as the outcome. Past research supports conscientiousness as a predictor of future career success (Dudley et al., 2006; Moffit et al., 2011; Sutin et al., 2009; Wilmot & Ones, 2019). We propose that this relation exists because being a

[15] If a predictor is not a likely cause of the outcome, then the researcher may be interested in prediction, rather than explanation. In that case, they may want to choose a set of predictors that maximizes the variance explained rather than focusing on the interpretation of coefficients.

conscientious worker (e.g., fulfilling work responsibilities, being punctual, working diligently) increases one's career success.

Next, generate a list of variables that could be confounders or confound-blockers—the (potential) controls. This list should include variables that are statistically associated with both the predictor and outcome. We considered variables that might be confounders of personality and career success and identified two–educational attainment and childhood socioeconomic status (SES).

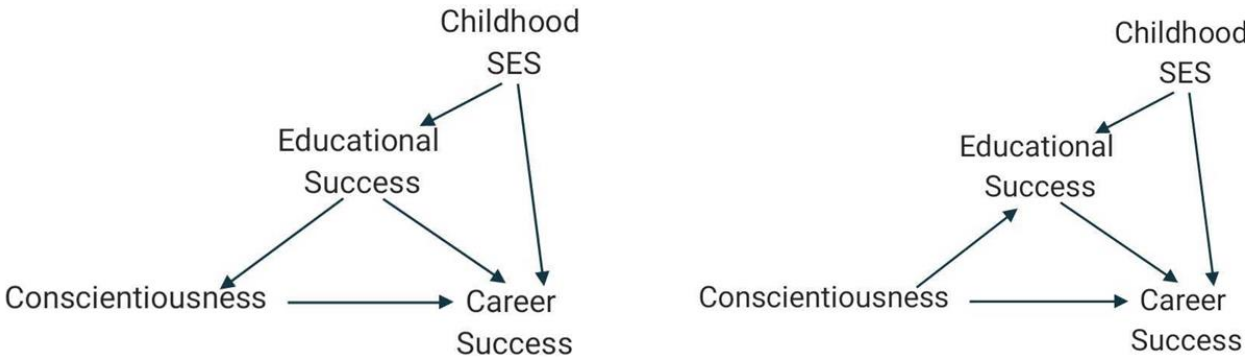
Specifying and Justifying the Causal Structure

With this list of potential controls, the next step is to specify a causal structure that includes the predictor, outcome, potential confounders, and other important third variables (see Figure 13 for our example). For this step, researchers can use *U* to stand in for a set of non-specific common causes. Using *U* as a stand-in variable allows researchers to reason about potential confounding paths and consider how to block them even when the full set of confounders are unknown. After outlining a causal structure, each part of the structure should be justified. Importantly, for the list of potential controls, researchers must justify why these variables are either confounders or confound-blockers of the predictor and outcome. For example, based on the *social investment principle* of personality, which posits that age-graded social roles serve as one mechanism of personality development (Roberts & Wood, 2006), we hypothesized that educational attainment influences conscientiousness; that is, engaging in the structured context of higher education that demands individuals to act responsibly should promote individuals to become more organized, hardworking, and responsible (i.e., more conscientious; Roberts et al., 2004). We also theorized that higher educational attainment would lead to greater career success given robust associations between educational attainment and

income, unemployment, job satisfaction, occupational prestige, and control over work (Gürbüz, 2007; Ross & Reskin, 1992; Slominski et al., 2011; U.S. Bureau of Labor Statistics, 2020). Together, these two lines of research support our hypothesized structure that educational attainment is a plausible confounder of the causal effect of conscientiousness on career success.

Additionally, childhood SES may also be a confounder or confound-blocker for conscientiousness (mediated by educational attainment) and career success because children from families with higher incomes and more highly educated parents are, themselves, more likely to achieve higher levels of educational attainment and have higher paying and more prestigious jobs (Gürbüz, 2007; Ross & Reskin, 1992; Slominski et al., 2011; U.S. Bureau of Labor Statistics, 2020). These causal hypotheses are depicted in Figure 13 (left panel). Of course, any of these hypotheses about the causal structure may be wrong. The value of this framework, however, lies in the clear outlining of assumptions and hypotheses and, in turn, the consequences that arise if these assumptions are wrong.

Figure 13. *Considering Alternative Structures: Educational Attainment as a Confounder or a Mediator*



Researchers should also justify why each control variable will not bias the estimate by blocking a causal path (i.e., it is not a mediator) or inducing a spurious path. In most research

studies there will be uncertainty about parts of the causal structure (e.g., a variable may be a plausible confounder and also a plausible collider for a pair of variables). These uncertainties can be depicted by having multiple competing models—a set of plausible causal structures. In our example, there is reason to believe that educational attainment might be a mediator between conscientiousness and career success, although it is unlikely to be a proxy or a collider.[16] That is, conscientiousness may be a cause of educational attainment—because being dispositionally hard-working and reliable causes people to do better in school and receive more opportunities to further their education, and and higher educational attainment, in turn, causes career success. Childhood SES, however, could neither be a mediator nor a collider for conscientiousness and career success because events that occur in adulthood do not change events that happened in childhood. Therefore, we propose that we have a clear confounder—childhood SES—and one variable that could be a confounder and/or a mediator—educational attainment.

Comparing Alternative Models and Selecting Control Variables

With a set of plausible causal structures, the next step is to select an appropriate set of control variables, which should block all confounding paths without blocking any causal paths or inducing any spurious associations between the predictor and outcome. This appropriate set of control variables need not contain every confounder–sometimes all confounding paths can be blocked with a subset of the confounders.

With a set of appropriate controls for each plausible model, researchers will end up in one of three positions. First, there may be a set of control variables that are appropriate across all

[16] Educational attainment is unlikely to be a proxy for conscientiousness, given its clear and well-substantiated relation to career success. Additionally, it is unlikely to be a collider between conscientiousness and work success, given the typical temporal relation between education and career success (i.e., most individuals complete their education before entering in the workforce).

plausible models. In this case, the researcher can argue that the value of the partial coefficient is a more accurate estimate of the hypothesized causal path than the simple coefficient. Researchers who interpret their estimated association causally must make it clear to readers that the causal interpretation of the results is predicated on the specified model being true and the assumptions outlined previously (see section *Statistical Control: How and When it Works*; e.g., all non-linear and interaction effects are correctly specified). Second, researchers may find that the appropriate set of control variables varies across the set of plausible models. In this case, researchers could select a single model and control for its corresponding control set and discuss how the interpretation of the results depends on the chosen structure being true. Here, the alternative models should be included in the paper along with their associated control sets. Alternatively, researchers could run separate models that control for each of the appropriate control sets and present the partial associations from each model along with a discussion of the assumptions that have to hold for each of these coefficients to be unbiased for the causal effect. Researchers may also consider conducting a sensitivity analysis to investigate which effects hold up to control by various possible confounders, keeping in mind that, without making assumptions about the causal model, there is *no basis to claim* that partial effects are more or less "conservative" than simple effects. Finally, researchers could be in a position where one or more of the models have no appropriate control set. When this is the case, researchers may consider blocking the confounding paths through some other method (e.g., instrumental variables or front-door criterion; Pearl, 1995) reporting the simple coefficients, or reporting the partial coefficients with an acknowledgement that not all of the confounding paths are blocked. These recommendations are summarized in Table 3.

When reporting partial coefficients, there are some practices researchers should always follow. First, both the simple and partial coefficients should be made accessible to the reader; not providing access to both sets of coefficients leaves the reader without valuable information about how statistical control impacts the estimates. Additionally, it is important to stress that the coefficient relating the outcome to the predictor and the coefficients relating the outcome to the controls cannot generally be interpreted in the same way, because control → outcome coefficients represent direct (rather than total) effects and they may themselves be confounded (Westreich & Greenland, 2013).[17]

Table 3. *Guidelines for Controlling after Justification Process*

Results of Causal Structure Reasoning	**Suggested Approach**
A single control set is appropriate across the plausible models	• Control for the control set • List assumptions that the causal interpretation is predicated on
Different control sets are appropriate across plausible models	• Control for one of the control sets • Discuss how the interpretation of the coefficient depends on the selected model being true Or • Run multiple models with different control sets • Present results from different models as competing estimates
	• Use another method to de-bias effects (e.g., instrumental variables) Or

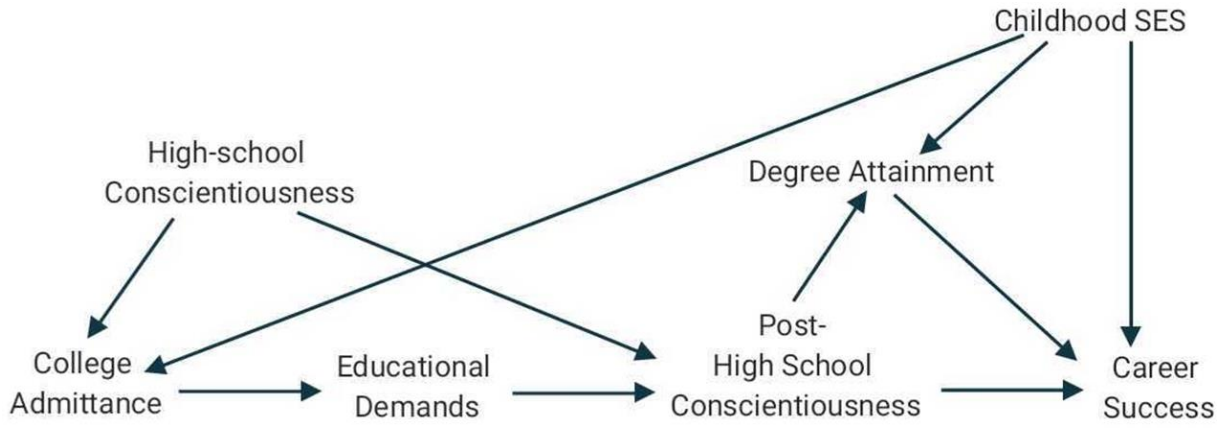
[17] The tendency to interpret these coefficients in the same way is known as the Table 2 Fallacy.

No control set	• Report simple coefficients Or • Report partial coefficients with an acknowledgement that it is a biased coefficient

In our example, because educational attainment may serve as either a mediator or a confounder, we must consider and discuss the consequences of both causal models (see Figure 13). If educational attainment is a confounder, then controlling for educational attainment would block the two confounding paths. Instead, if educational attainment is a mediator, then controlling for educational attainment not only blocks a causal path but also induces a spurious path (due to Conscientiousness → Educational Attainment ← Childhood SES). As such, if educational attainment is a mediator, controlling for childhood SES is the correct approach. If longitudinal data were available, then it could be possible to distinguish educational attainment as a confound from educational attainment as a mediator. Longitudinal data would also allow fixed effects to be estimated, thereby removing time-invariant confounding by controlling for all fixed person-level attributes. Another way to disentangle whether a variable is a confounder or a mediator is to take a more nuanced view of each of the variables (see Box 2).

Box 2. Breaking Down Variables: Figure 14 depicts DAGs that include variables that are amalgams of complex processes that unfold across time (e.g., educational attainment). Collapsing over a variable can be useful, but it can also create ambiguity when de-biasing a causal effect. One way to disambiguate this process is by specifying each variable's content and time-scale. For example, educational attainment could be broken down into multiple parts–being a college attendee, engaging in the educational demands of college (e.g., completing schoolwork), and receiving a college degree. Additionally, conscientiousness can be assessed multiple times–once during high school and once after high school. Adding this nuance helps us to clarify the causal effects; in particular, early conscientiousness influences college admittance and the educational demands of college influences later conscientiousness. It also clarifies how childhood SES impacts multiple steps in the educational path, including college attendance and degree attainment. Assuming this more nuanced structure is accurate, controlling for childhood SES would allow for the effect of post-high school conscientiousness on later career success to be identified.

Figure 14. *Breaking Down Variables in a DAG*



Discussion

In this paper, we demonstrated the importance of carefully selecting control variables. In particular, we highlighted how controlling for the wrong variable can lead researchers to results and interpretations that are less accurate than if no variables had been controlled. Further, we showed that the underlying causal structure determines whether controlling for a variable adds or removes bias. Additionally, we clarified that statistical associations are not sufficient justification for selecting a control variable because these associations could arise from a number of different causal structures.

Throughout this paper, we discussed estimating the weights of causal paths and how these weights can be biased by controlling for the wrong variable. But this framework is important even for researchers who are only interested in assessing whether a causal effect exists (rather than estimating the weight of that causal effect). In this case, the researcher may not be concerned whether the weight of the causal path is an underestimate (or overestimate) as long as the results indicate that there is a non-zero causal path between the two variables. There are two reasons why the causal structure is important even if the existence, rather than the weight, of a causal effect is the focus. First, controlling for the wrong variable can, in some situations, entirely remove the effect of interest. In particular, if the impact of X on Y is mainly mediated by a third variable and that mediator is controlled for, the association between X and Y can be reduced to (or very near to) zero, leading to the relation being mistakenly dismissed as unimportant. The same thing can happen if a spurious association that biases the effect of interest is induced when controlling for a collider, and this association is of the opposite sign and sufficiently strong. Additionally, although researchers may not be concerned with the exact weight of the path of interest, they likely care that it is at least in the right direction (e.g., if

higher Machiavellianism *increases* aggressive behavior then it would not be helpful to have results indicating that higher Machiavellianism *decreases* aggressive behavior). As Figures 9 and 10 show, controlling for an inappropriate third variable can inaccurately flip the sign of the coefficient, leading to an estimate that indicates a causal effect in the opposite direction. Therefore, we reiterate that, for controlled results to be meaningfully interpreted to explain a process, a causal structure must be proposed and defended. Without a causal structure, neither the researcher nor the reader is able to make sense of discrepancies between the partial and simple coefficients.

Box 3. What about Controlling to Assess Incremental Validity?

Researchers will sometimes cite incremental validity, rather than explanation, as the motivation for statistical control. Incremental validity means that a new predictor accounts for substantial variance in an outcome variable over and above the variance accounted for by other, known predictors s (Sechrest, 1963), and it is typically established using multiple regression with the previously known predictors included as controls. If the new predictor accounts for significant variance over and above the other predictors, then the new predictor is said to have incremental validity.

There are three common ways in which incremental validity is used in psychological research. The first is to establish the clinical usefulness of a test to predict a specific criterion (Smith et al., 2003). In this use, the research goal is improved *prediction*, and it is appropriate to choose control variables that represent whatever measures are currently most popular or predictive in the field. Incremental validity for the sake of establishing predictive utility does not require causal justification.

The second common use of incremental validity is to establish that a new theoretically motivated predictor *explains* a phenomenon over and above previously established predictors. In this case, incremental validity allows researchers to "[demonstrate] the empirical novelty of a conceptual contribution" while "simultaneously [acknowledging] a precedent established in earlier work" (p. 157, Wang & Eastwick, 2020). This use of incremental validity can lead to greater "understanding of the relations between a focal predictor, a covariate, and an outcome variable" (p. 157, Wang & Eastwick, 2020). We have endeavored to make clear by now that any method that aims to build up a theoretical understanding of variable relations needs to consider the causal direction of those relations when choosing which predictors to control for.

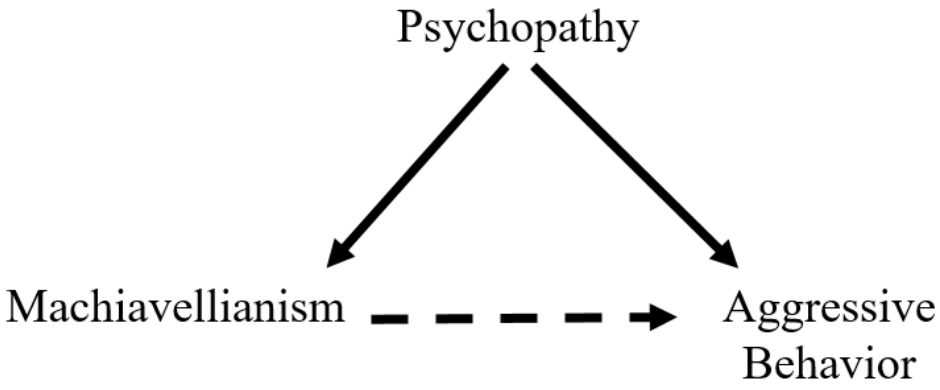
But Wang and Eastwick (2020) describe a third use of incremental validity testing, which may initially seem not to require causal theory, namely the "isn't-it-just" argument. The isn't-it-just argument uses incremental validity to dispute a criticism that a new measure only predicts an outcome because it is a proxy for some already established predictor (e.g., "you found an effect of Machiavellianism on

Box 3 (Continued). What about Controlling to Assess Incremental Validity?

aggressive behavior but isn't Machiavellianism just psychopathy?").

This "isn't it just…" argument, though using different language, is equivalent to ruling out an alternative explanation that Machiavellianism is a proxy for psychopathy (i.e., variability in psychopathy causes or produces variability in scores on a Machiavellianism scale). Someone arguing this point might claim that the estimated effect is due to a causal path from psychopathy to aggressive behavior rather than such a path from Machiavellianism to aggressive behavior (see Figure 15). In other words, the "isn't it just…" argument is another way to posit a confounder. As such, it makes sense to control for psychopathy to estimate the remaining (i.e., unique) effect of Machiavellianism on aggressive behavior. But, before testing for incremental validity, it is still important to establish that psychopathy is a potential confounder by defending the underlying causal structure. If it were more plausible that psychopathy was a proxy for Machiavellianism, rather than the other way around, or that psychopathy was a plausible mediator of the effect of Machiavellianism on aggressive behavior, then it would not make sense to investigate the incremental validity of Machiavellianism over and above psychopathy. Therefore, researchers making the "isn't it just…" argument to test for incremental validity should carefully consider the plausible causal ordering of variables to appropriately control for third variables.

Figure 15. *Incremental Validity as a Question of Confounding*



Conclusion

Although the causal structure holds the key to identifying confounders, it is uncommon for psychological researchers to present causal justification for their choice of control variables. The absence of explicit causal reasoning may be due to an unspoken ban on causal language within psychology (Dablander, 2020; Grosz et al., 2020). In contrast to fields such as epidemiology and economics, psychologists who report on non-experimental findings often claim to be interested in prediction or association even though the interpretations made within a paper are more compatible with causal inference (Grosz et al., 2020). In this paper, we have argued that statistical control for the goal of causal inference is incompatible with causal agnosticism about how the control variable relates to the predictor and outcome: Whether controlling for some variable increases or decreases bias is a function of how the variables in a model causally relate to each other.

There are three main benefits to justifying the causal structure. First, control variables are selected in a more appropriate and careful manner. This will improve results from analyses with control variables and increase the alignment between the estimate and the theoretical quantity of interest. Second, there will be an improvement in the theoretical models motivating the study. Most researchers have, at least implicitly, a hypothesized causal structure. Taking the time to (literally) draw the causal model may give researchers an opportunity to think carefully about their causal assumptions and to consider alternative plausible causal structures. Third, when the hypothesized causal structure is made explicit, readers will more easily glean the causal framework the authors are working from. Readers can then be aware of the assumptions that are inherent in a model and can reason about how those assumptions may influence the results if they are violated. In summary, psychological research stands to benefit substantially from

researchers thinking and communicating carefully about why they selected specific control variables.

As we have underscored in this chapter, variable selection is crucial for valid inference. But one aspect of variable selection that we have only briefly discussed is measuring the set of variables at the *correct time interval*. Generally, measuring the correct set of variables at the wrong time will result in missing the effect of interest (Gollob & Reichardt, 1987). In the next chapter, we develop a statistical model that allows researchers to use existing empirical information to improve estimates calculated from the correct set of variables that are measured at the wrong time.

Chapter 4. Incorporating Stability Information into Cross-Sectional Estimates

Many psychological theories and research questions involve processes that unfold over time, and are thus inherently longitudinal (e.g., impact of education on income). But psychological researchers commonly use data collected at a single time point (cross-sectional data) to make inferences about such processes (Gollob & Reichardt, 1987). It is well-known that regression and correlation coefficients based on cross-sectional data are not unbiased estimates of their corresponding longitudinal coefficients (Maxwell & Cole, 2007; Maxwell, Cole & Mitchell, 2011; O'Laughlin, Martin, & Ferrer, 2018). Because cross-sectional coefficients do not account for the longitudinal stability of either predictors or outcomes, the resulting estimates reflect a mixture of cross-lagged associations (i.e., associations between two different variables) and auto-regressive associations (i.e., association between a variable and a future version of itself). A coefficient that is a combination of both associations with no clear way to disentangle the two is problematic when a cross-lagged effect is of interest.

The standard advice to researchers, then, is that longitudinal research questions require longitudinal data (Maxwell & Cole, 2007; Maxwell et al., 2011; O'Laughlin, et al., 2018). But collecting longitudinal data with the relevant time lag is not always feasible. In such cases, it would be useful to have a method that can improve cross-sectional estimates. To this end, we outline an approach that combines cross-sectional data with existing knowledge about variable stability, which allows researchers to estimate longitudinal coefficients with data collected at a single time point. We also explore, via simulation, the effect of misspecification on the bias of stability-informed estimates.

Terminology

In this article, the term *cross-sectional coefficient* refers to a regression coefficient derived from a model in which all variables are measured at a single time point, and *longitudinal coefficient* refers to a regression coefficient derived from a longitudinal model in which all variables are measured at two time points. An *autoregressive* coefficient (e.g. AR_X) is a longitudinal regression coefficient of a single variable on itself at a previous time point. A *cross-lagged* coefficient (e.g., CL_{YX} is the longitudinal regression of a variable (e.g., *Y*) on a different variable (e.g., *X*) measured at a previous time point, *controlling for the outcome variable measured at the earlier time point* (e.g., *Y* measured at T_0). To be clear, these coefficients are not necessarily equivalent to *causal effects*. For a longitudinal coefficient to be an unbiased estimator of a causal effect, several additional assumptions must be met, such as accounting for all confounding variables and specifying the model correctly (e.g., correctly specifying a non-linear effect; Simonsohn, 2015).

The Use of Cross-Sectional Estimates in Psychology

Recent research has illuminated a disconnect between psychological theories, which tend to focus on explaining causal mechanisms, and the preponderance of psychological research articles that purport to examine those theories, which often avoid explicit causal inference (Grosz et al., 2020). In particular, Grosz and colleagues (2020) found that many studies that use regression or path analysis models do not explicitly state that their goal is causal inference but describe model decisions and inferences that are only justifiable under a causal interpretation of the results. One implication of this disconnect between theory and research method is that researchers often avoid the difficult (but possible) work of delineating, justifying, and testing the assumptions required for causal inference (e.g., the assumption that no omitted third variables act as a common cause that may explain the observed associations). Another implication is that

researchers do not delineate the time interval over which purported mechanisms are theorized to occur, and thus do not acknowledge the discrepancy between the theorized time interval and the measured time interval. In fact, a great deal of published regression and path analysis research uses data in which the hypothesized predictors and outcomes were measured at the same time. For example, Maxwell and Cole (2007) evaluated 72 studies that mentioned mediation in their abstract and found that 53% of those studies used cross-sectional data to estimate the mediation model. When the research goal is causal inference, all research questions are longitudinal and variables measured at a single timepoint can rarely capture these effects (Gollob & Reichardt, 1987).

Two recent papers exemplify the use of cross-sectional designs to answer longitudinal questions. In one, Devine and Apperly (2022) were interested in how theory of mind and social motivation impact social competence in children. The authors found positive cross-sectional associations between both predictors (theory of mind and social motivation) and the outcome (social competence), leading them to the causal-sounding inference that "both theory of mind ability and social motivation contribute to successful social interaction at school" (Devine & Apperly, 2022, p. 1). In another paper (Kim, Sommet, Na, & Spini, 2022), researchers were interested in how subjective social class impacts both social trust (How much do you trust most people?) and institutional trust (How much do you trust various institutions?). These authors interpreted positive cross-sectional associations between social class and social and institutional trust to indicate that "[...]individuals from higher social classes are usually motivated to maintain social hierarchy and support institutions" (Kim et al., 2022, p. 193). Their conclusion implies a mediated causal effect of social class on trust. In both cases, the authors interpreted a cross-sectional association as a longitudinal effect. But the cross-sectional coefficient incorporates

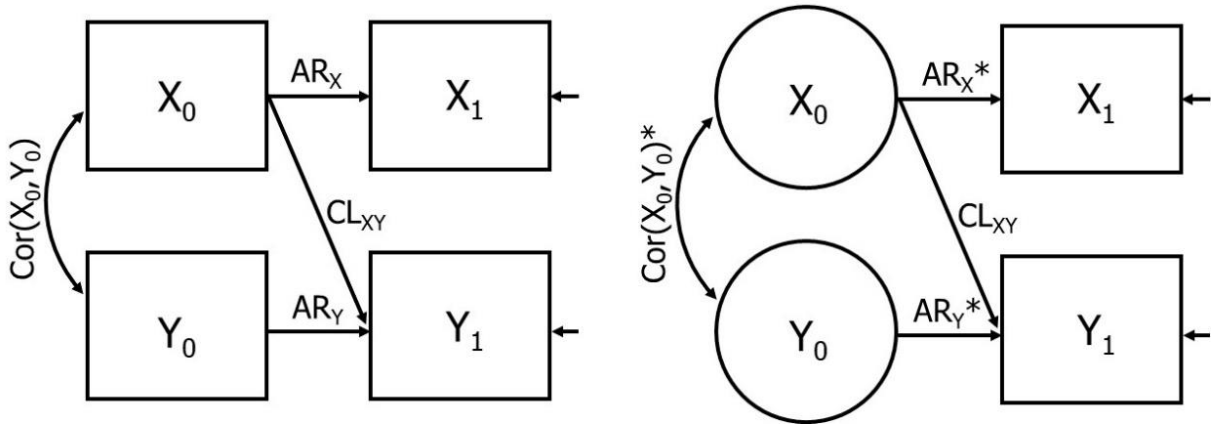
information from only one time point and it is unclear what time lag this effect unfolds over. As such, the cross-sectional association is likely to be biased for the effect of interest.

By using these two studies as an example, we do not intend to single out either group of authors; rather we mean to highlight a ubiquitous practice that stems from (1) researchers' desire to understand mechanisms and develop theories and (2) the challenge of obtaining longitudinal data. Below, we explain why using cross-sectional estimates to answer longitudinal questions can complicate the matter of obtaining an unbiased estimate.

The Problem with Cross-sectional Estimates

Estimating a parameter whose expected value is equal to the weight of the population coefficient (i.e., an unbiased estimate) is an important goal. And getting an unbiased estimate for a longitudinal effect, when only cross-sectional data are available, is exceptionally difficult.

Figure 16. *A Path Model and the Corresponding Stability Informed Model.*



Note. Left Panel: Depicted is a stationary model with two variables, *X* and *Y*, at two time points. AR_X and AR_Y represent the weights of the auto-regressive effects and CL_{XY} represents the weight of the cross-lagged effect. Right Panel: A phantom variable model where a variable's stability can be set at a specific value to enable the estimation of the longitudinal (cross-lagged) effect. AR_X*, AR_Y*, and $Cor(X_0, Y_0)$* denote constrained parameters.

Suppose a researcher is interested in the effect of *X* on *Y*. If the true model is as depicted in Figure 16 (left panel), then the weight of the effect of interest is equal to CL_{XY}. An unbiased estimate for the effect of interest could be obtained by regressing Y_1 on X_0, controlling for Y_0. But only measuring X_1 and Y_1—while remaining interested in the longitudinal effect—complicates matters. The standardized regression coefficient of Y_1 on X_1 is equal to:

$$\beta_{Y_1X_1} = AR_X CL_{XY^+} AR_X AR_Y Cor(X_0, Y_0) \quad (12)$$

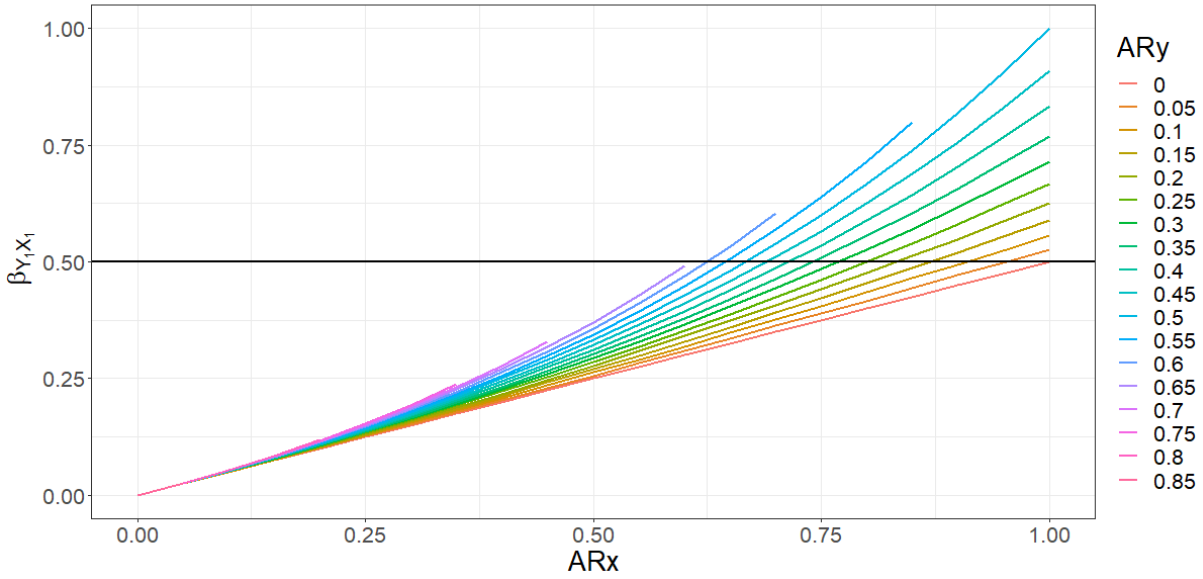
As shown in Equation 11, under most circumstances, the cross-sectional coefficient will be biased for the longitudinal coefficient. Part of the reason for this bias is that the cross-sectional coefficient is also a function of the auto-regressive coefficients (AR_X and AR_Y). Cross-sectional data do not contain the longitudinal information that is necessary to disentangle auto-regressive from cross-lagged coefficients. As a result, the weight of the auto-regressive associations can drastically impact cross-sectional estimates which is problematic when the cross-lagged coefficient is of interest.

In Figure 17, we demonstrate this by plotting the change in the simple regression coefficient of Y_1 on X_1 for varying auto-regressive coefficients.[18] The value of AR_X varies on the x-axis and the values for AR_Y are depicted with different colored lines. The black line represents the true value of the cross-lagged coefficient (i.e., the parameter of interest) which is set to .5. For this model, $\beta_{Y_1X_1}$ is unbiased for the longitudinal parameter when $AR_X = \frac{1}{1+AR_Y}$ (Gollob & Reichardt, 1985). It is clear from Figure 17 that, although this condition is occasionally met (namely, where the colored lines intersect .5), the cross-sectional coefficient is almost always

[18] These results were generated by solving for $\beta_{Y_1X_1}$ (using Equation 11) with a range of weights for the auto-regressive effects. The generating model is stationary.

biased. Moreover, the direction of the bias is not constant, in that sometimes the cross-sectional coefficient over-estimates the cross-lagged coefficient and other times it underestimates it.[19]

Figure 17. *The Impact of Auto-regressive Effects on Cross-sectional Estimates.*



Note. The horizontal black line at .5 depicts the weight of the true cross-lagged effect. The generating model is stationary (i.e., $Cor(X_0, Y_0) = Cor(X_1, Y_1)$)

There is one more complication. Often, psychological researchers are interested in estimating multiple effects simultaneously (e.g., mediation models, path models). If a variable is both a predictor and outcome (e.g., the mediator in a mediation model), then it becomes even more unlikely that the set of auto-regressive weights will result in unbiased estimates, and often such a set does not exist (Maxwell & Cole, 2007; Maxwell et al., 2011).

The fact that few conditions will render cross-sectional estimates unbiased for longitudinal parameters vastly undercuts the value of cross-sectional data for many research questions. Accurately estimating cross-lagged effects requires information about auto-regressive

[19] Similar results can be found in Maxwell and Cole (2007).

parameters. Although cross-sectional data cannot provide this information, longitudinal information may exist in the literature from previously collected longitudinal data. In these cases, having a method to integrate longitudinal information (from another data set) into cross-sectional estimates could assist in getting an estimate that accounts for the auto-regressive effects. In the following section, we describe the *Stability Informed Model*, which combines cross-sectional data with existing knowledge about a variable's stability—the degree to which a variable correlates with itself across two time points.

The Stability Informed Model

We developed a model that we call the Stability Informed Model to augment cross-sectional estimates with longitudinal information. Later, we discovered that such a model had been proposed by Gollob and Reichardt (1985). Although this model was introduced into the literature many years ago, few methodologists or substantive researchers appear to be familiar with it. Additionally, to our knowledge, it has never been used. As such, we believe there is value in re-introducing and extending the Stability Informed Model, testing its robustness, and providing usable tools and explanation for its estimation.

The Stability Informed Model combines cross-sectional correlational data with pre-specified stability coefficients and produces estimates of longitudinal coefficients. If the process being modeled is stationary and the correct model is specified, this approach allows for the estimation of longitudinal coefficients with data collected at a single time point. We capitalize on the fact that information about the stability of variables may exist in the literature (e.g., from previous longitudinal studies), and, as such, it may be more plausible for researchers to obtain cross-sectional correlations and independent stability estimates than to obtain longitudinal data.

The Stability Informed Model uses phantom variables within an SEM framework: unmeasured previous time points of the measured variables are specified as standardized phantom latent variables and auto-regressive paths are specified between each measured variable and its associated phantom variable (e.g., Measured: X_1, Y_1, Unmeasured and included as phantom variables: X_0, Y_0; see Figure 16 right panel). The Stability Informed Model shares similarities with other phantom variable approaches (e.g., Harring, McNeish, & Hancock, 2017), and is similar to single-indicator measurement models in which external reliability estimates are used to derive fixed measurement error variance (Savalei, 2019). What sets the Stability Informed Model apart from these other models is the added model constraints and the motivation behind the inclusion of these constraints.

The Stability Informed Model is identified by imposing three sets of constraints. First, the auto-regressive paths of each variable are constrained to be a nonlinear function of the user-specified stability of the variable and other model parameters. Second, the stationarity assumption is imposed, such that the variance-covariance matrix of phantom variables at Time 0 is equal to the variance-covariance matrix of the measured variables at Time 1.[20] Third, the within-time-point residual covariances are fixed to 0 (although, as we will discuss later, these can be fixed to non-zero values or estimated). With these constraints imposed, the model has the freedom to estimate $\frac{p(p-1)}{2}$ additional cross-lagged paths from phantom variables at Time 0 to observed variables at Time 1 (see section *Degrees of Freedom for the Stability Informed Model* in the Supplementary Material for more information).

[20] These constraints can be imposed using software packages like *lavaan* (Rosseel, 2012)

Figure 16 (right panel) shows the Stability Informed Model corresponding to a two-variable model in which *Y* is regressed on *X*. To estimate the cross-lagged path CL_{XY} when only Y_1 and X_1 are measured, the parameters AR_X*, AR_Y*, and $Cor(X_0, Y_0)$* are constrained to:

(13) $$AR_x^* = Cor(X_0, X_1)$$

(14) $$AR_y^* = Cor(Y_0, Y_1) - CL_{xy} Cor(X_0, Y_0)^*$$

(15) $$Cor(X_0, Y_0)^* = \frac{AR_x^* CL_{xy}}{1 - AR_x^* AR_y^*}$$

where $Cor(X_0, X_1)$ and $Cor(Y_0, Y_1)$ would be replaced with plausible stability values for these variables. These plausible stability values would be informed by the existing longitudinal literature on these variables.

Assuming the correct model was specified, the correct stability values were used, and the process was stationary, the stability-informed estimate of CL_{XY} would be unbiased for the cross-lagged coefficient. Although stability-informed estimates can be unbiased for the longitudinal coefficient when the 3 key assumptions are met, it is likely that these assumptions are violated in real data. Additionally, the Stability Informed Model will often be saturated which means that model fit indices cannot be used to identify misspecification. As such, it is important to investigate how this model responds to violations of these assumptions. To this end, in the following section we explore the effects of both stability misspecification and structural model misspecification.

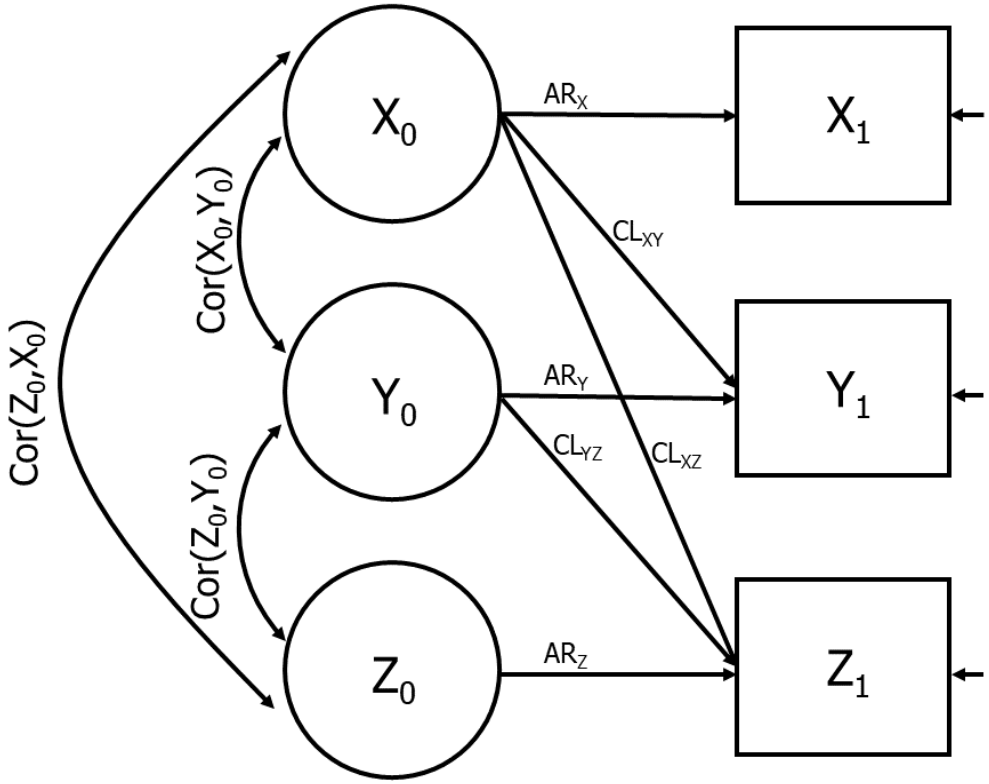
Exploring Misspecification within the Stability Informed Model

Stability Misspecification

In this section, we explore the impact of stability misspecification, which occurs when the specified stability values are not equal to the true stability values in the generating model. We generated data from the model depicted in Figure 18 with all auto-regressive coefficients having the same value of either .3 or -.3, and all cross-lagged coefficients having the same value of either .3 or -.3.[21] We then fit the Stability Informed Model to the generated population covariance matrix (i.e., no sampling variability was imposed), with one variable's stability fixed to an incorrect value in the fitted model. We manipulated the specified stability of one variable at a time so that the impact of each variable's stability misspecification would be clear. The stability misspecification varied from -50% of its true value to +50% of its true value in increments of 10%. For example, in conditions where *X*'s stability misspecification is being varied, the stability of *X* would be some percentage greater or less than its true value while the stability of *Y* and *Z* would be set at their true values.

[21] These weights were used because they are within the interval of weights that we may plausibly encounter for psychological estimates.

Figure 18. *Generating Model for Misspecification Simulations.*

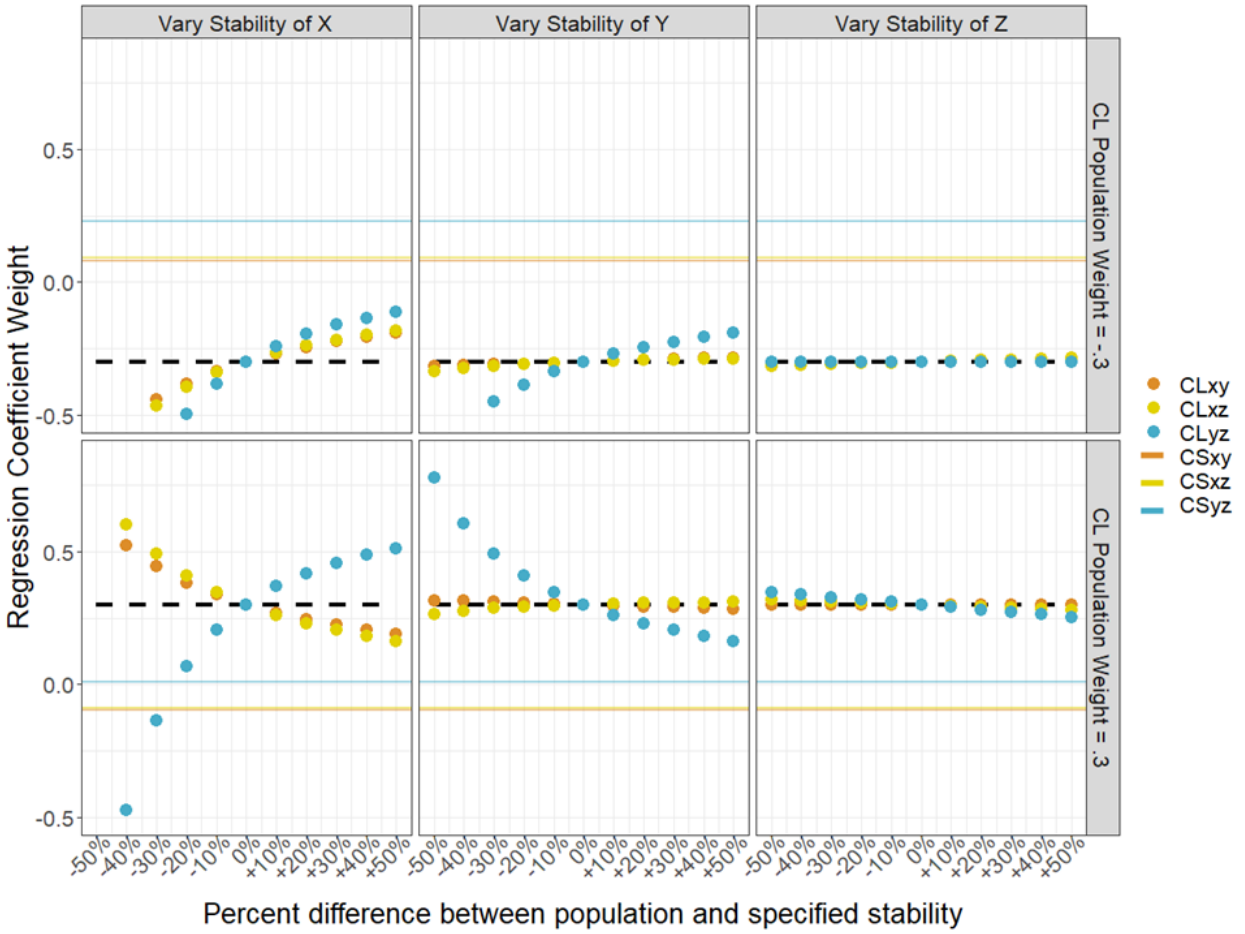


Note. For conditions exploring misspecification due to correlated residuals, one of the residual covariances between the observed variables was non-zero.

Figure 19 shows the stability-informed estimates (depicted by points) produced by fitting the Stability Informed Model to the generated Time 1 data (i.e., X_1, Y_1, and Z_1). These results are from conditions where the auto-regressive coefficients are set at .3 (see the Figures A10-A12 in the Supplementary Material for the results when auto-regressive paths are set at -.3). For reference, solid lines depict the corresponding estimates from a cross-sectional model in which Y_1 is regressed on X_1, and Z_1 is regressed on both Y_1 and X_1. The dashed line depicts the true value of the cross-lagged coefficients in the longitudinal generating model. Only conditions that resulted in a positive definite covariance matrix and where a solution was found are shown in Figure 19.

When the specified stabilities are equal to those in the generating model (i.e., at 0% on the x-axis), the stability-informed estimates are unbiased for the longitudinal parameters. But when a specified stability is inflated or deflated relative to the true value, the stability-informed estimates show considerable bias. In some cases, this bias is greater than that of the cross-sectional estimates. Comparison between the three column panels in Figure 19 shows that the consequence of misspecifying the predictor stability is more costly than misspecifying the outcome stability. For example, misspecifying the stability of *X* has negative consequences for all three of the cross-lagged coefficients. But misspecifying the stability of *Y* only has a negative consequence on CL_{YZ}. Finally, misspecifying the stability of *Z*, which is strictly an outcome variable, has little effect on any of the stability-informed estimates (assuming that the stability of *X* and *Y* are correct). As such, this model may be particularly useful for research questions where researchers are confident in their estimate of the predictor's stability, such as when the predictor is an unchanging variable like race or childhood SES. Additionally, the impact of under-estimating a variable's stability appears to be worse than that of over-estimating it. As such, if there is uncertainty about a variable's stability, we suggest picking a value in the upper range of the plausible stabilities.

Figure 19. *Assessing the Impact of Stability Misspecification.*



Note. The values of each of the auto-regressive paths in the generating model are set at .3. The black dashed line represents the population weight for each of the estimated cross-lagged paths, and the solid lines represent the cross-sectional estimates for each path (colors correspond between the same cross-lagged and the cross-sectional estimate). The lines and dots depicting the unestimated cross-lagged and cross-sectional paths overlap at zero in each panel. For each condition, the stability in the estimated Stability Informed Model was set at its true value.

Structural Model Misspecification

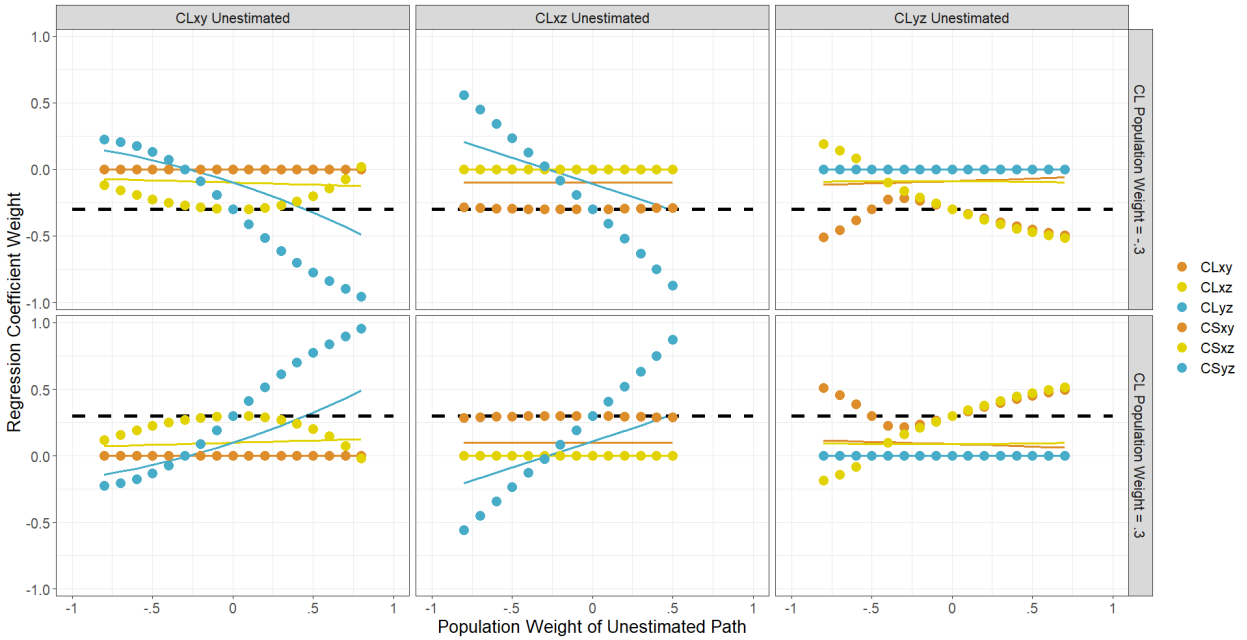
Structural model misspecification concerns whether the correct paths are estimated in the structural model (e.g., two residuals covary but their covariance is constrained to be 0). We investigated two kinds of structural model misspecification: missing cross-lagged paths and missing covarying residuals.

Missing Cross-lagged Paths. First, we explored the impact of estimating an incorrect set of cross-lagged paths. Recall that the model allows only $\frac{p(p-1)}{2}$ paths to be estimated; as a result, researchers may wish to estimate more cross-lagged paths than they are able to. We tested how a cross-lagged path being inaccurately set to zero impacted the other estimates. We did this by calculating the generating covariance matrix for the model depicted in Figure 18. For each condition, one of the cross-lagged paths (CL_{XY}, $CL_{XZ,}$ or CL_{YZ}), was fixed to zero in the Stability Informed Model. The true weight of the unestimated path ranged from -1 to 1. The true weight of the other two cross-lagged paths were equal to each other and set at either .3 or -.3. Across conditions, the auto-regressive effects were all equal to each other and set at either .3 or -.3.

Our goal was to explore the impact of structural—not stability—misspecification and, as such, the stability values for each variable were specified at their true values. But the auto-regressive estimates could be biased since they are a function of both stability and cross-lagged parameters. Only conditions that resulted in a positive definite matrix and where a solution was found are depicted in Figure 20. In Figure 20, the true weight of the unestimated path varies along the x-axis. The stability-informed estimates are depicted by different colored dots, and the cross-sectional estimates (for comparison) are depicted by different colored lines. In the figure, each column represents the results when a specific cross-lagged path is inaccurately fixed to zero, and each row depicts a different true cross-lagged weight for the estimated coefficient. The black dashed line depicts the true value for the cross-lagged coefficients that are included in the Stability Informed Model specification. Only the results from when the auto-regressive values are set at .3 are shown in Figure 20.

We can see that as the absolute value of the unestimated path increases, the stability-informed estimates of the other cross-lagged coefficients become more biased. The CL_{YZ} path is the most impacted by either the CL_{XY} or the CL_{XZ} path being erroneously fixed to zero. This is because the association between *Y* and *Z* comes from two sources—*Y* and *Z* share a common cause (*X*) and *Y* is a cause of *Z*. If one of the paths from *X* to either *Y* or *Z* is erroneously constrained to zero, then the CL_{YZ} path will be biased to account for the association between *Y* and *Z* that is due to their common cause. Additionally, fixing the CL_{YZ} path to zero affects both CL_{XY} and CL_{XZ}, because the common cause paths are forced to account for the association that, in the generating model, is due to the longitudinal regression of *Z* on *Y*. More generally, setting a non-zero effect in the population to zero biases the other estimates as there is now extra association that they have to account for. A similar issue would occur if an important variable (e.g., *X*) was not included in the model (this is known as omitted variable bias; Wilms, Winenn & Lanwehr, 2021). Misspecified cross-lagged paths could be particularly problematic for the Stability Informed Model as the limited degrees of freedom of the model strongly restrict how many of these paths may be estimated. One option is to constrain some of the cross-lagged effects to be equal, which would allow for more cross-lagged paths to be estimated.

Figure 20. *Assessing the Impact of Cross-lagged Path Misspecification.*



Note. The values of each of the auto-regressive paths in the generating model are set at .3. The black dashed line represents the population weight for each of the estimated cross-lagged paths, and the solid lines represent the cross-sectional estimates for each path (colors correspond between the same cross-lagged and the cross-sectional estimate). The lines and dots depicting the unestimated cross-lagged and cross-sectional paths overlap at zero in each panel. For each condition, the stability in the estimated Stability Informed Model was set at its true value.

Missing Residual Covariance.

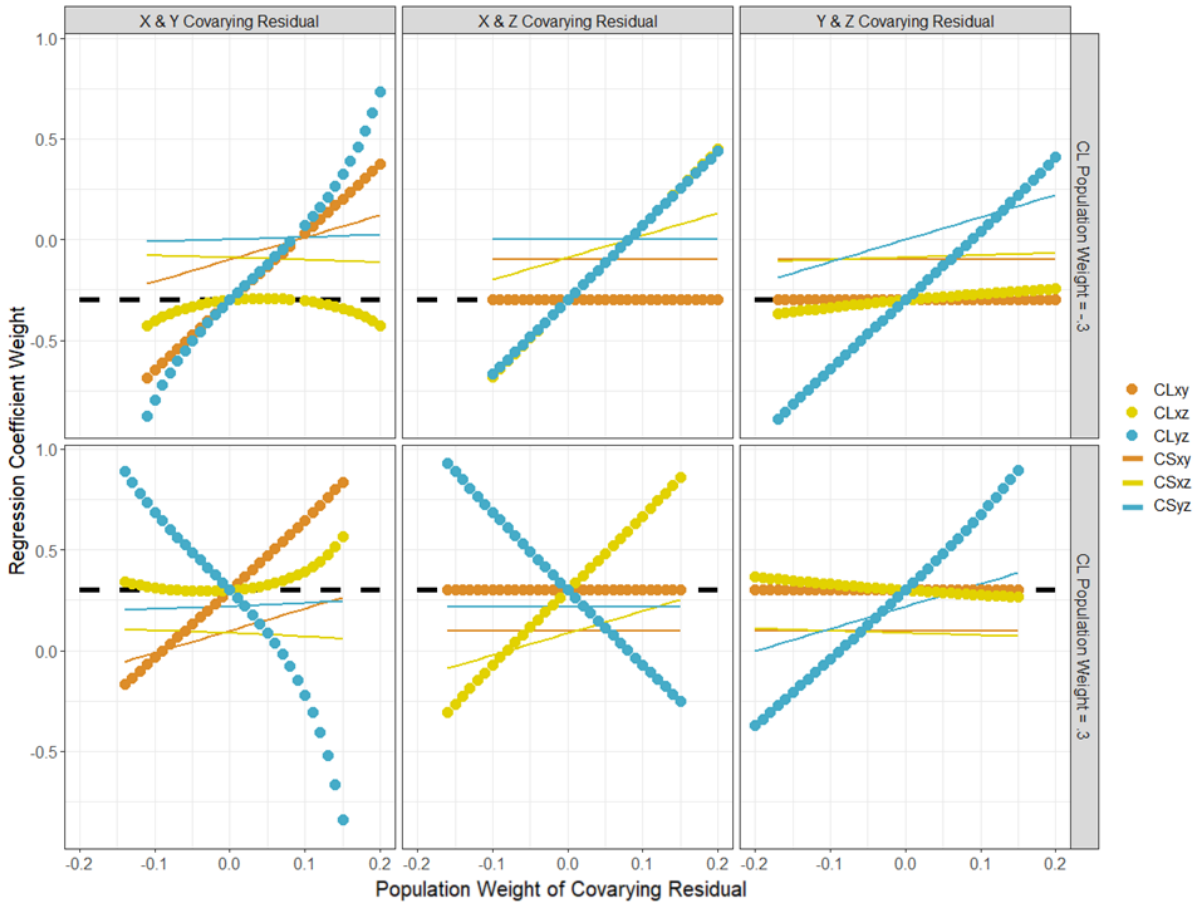
The second kind of structural model misspecification we explored is a non-zero residual covariance in the data generating model (e.g., the residual covariance for *X* and *Y* was set at .2). To do this, we estimated the Stability Informed Model with the correct stability specification and the correct set of cross-lagged paths, but a residual covariance inaccurately constrained to 0. Across conditions, we varied the non-zero residual covariance between -.2 and .2 in increments of .01. Note that, because all variables in the generating model have variance of 1, the residual covariances can be interpreted as being on a standardized metric corresponding to the total correlation between variables. For example, if *X* and *Y* have a residual covariance of .2, this means that .2 of the correlation between *X* and *Y* can be attributed to a covariance between their

residuals. As in the previous sections, the set of cross-lagged coefficients were set to be equal to each other; the set of auto-regressive coefficients were set to be equal to each other; the auto-regressive and cross-lagged coefficients were set to be either -.3 or .3. Only conditions that resulted in a positive definite matrix and where a solution was found are depicted in Figure 21.

In Figure 21, the true weight of the residual covariance varies along the x-axis, with 0 in the center. The stability-informed estimates are depicted by different colored dots, and the cross-sectional estimates are depicted by different colored lines. Each column represents the results when a different pair of variables have a non-zero residual covariance. The black dashed line depicts the population value for each of the cross-lagged coefficients.

Figure 21 shows that the bias from residual covariances being inaccurately set to zero can be severe. This is particularly true for the cross-lagged path between the two variables that have a covarying residual. For example, when *X* and *Y* have a non-zero residual covariance, the path CL_{XY} now has to account for either more or less of the covariance between these two variables than it does in the generating model (whether the effect is over or underestimated will depend on the values of both the residual covariance and other population parameters). But other cross-lagged paths can be biased as well. For example, CL_{YZ} is also biased when *X* and *Y* have covarying residuals. Other cross-lagged coefficients are impacted because when one cross-lagged coefficient is biased the remaining cross-lagged coefficients will have to adjust as well. This is particularly true for `downstream' coefficients in which the predictor is not exogenous. For example, CL_{XY} and CL_{XZ} are only strongly biased when there is residual covariance between *X* and *Y* and *X* and *Z*, respectively. On the other hand, CL_{YZ} (a downstream coefficient) is heavily impacted regardless of which two variables have a covarying residual.

Figure 21. *Assessing the Impact of Covarying Residual Misspecification.*



Note. The values of each of the auto-regressive paths in the generating model are set at .3. The black dashed line represents the population weight for each of the estimated cross-lagged paths, and the solid lines represent the cross-sectional estimates for each path (colors correspond between the same cross-lagged and the cross-sectional estimate). For each condition, the stability in the estimated Stability Informed Model was set at its true value.

Structural model misspecification is particularly a problem for the Stability Informed Model as there will likely be more paths a researcher would like to estimate than there are degrees of freedom available. This is an important limitation to keep in mind. One potential solution is to constrain some estimated paths to be equal which allows more paths to be estimated. In Table 4, we demonstrate this by constraining CL_{XY} and CL_{XZ} to be equal in the Stability Informed Model. This provides an extra degree of freedom and the ability to estimate

one more parameter, in situations where the assumption of equal cross-lagged paths is tenable. The model could then estimate the non-zero covariance between *X* and *Y*. Table 4 shows the parameter values returned by the Stability Informed Model when the residual covariance between *X* and *Y* is and is not fixed to zero.

Table 4. *Estimating Residual Covariances*

	Model 1: X and Y residual covariance constrained to 0	Model 2: X and Y residuals allowed to covary
AR_X	.3	.3
AR_Y	.3	.3
AR_Z	.22	.3
CL_{XY}	-.03	.3*
CL_{XZ}	.31	.3*
CL_{YZ}	.69	.3
$Cov(\varepsilon_x, \varepsilon_y)$	0†	-.10

Note. The true weights for the cross-lagged and auto-regressive parameters is .3. The true weight for the residual covariance between X and Y is -.10
** CLXY and CLXZ are constrained to be equal in Model 2.*
† Residual covariance between X and Y is fixed to zero in Model 1.

Table 4 demonstrates that, if a researcher has enough knowledge about the model, this knowledge can be used to free up more degrees of freedom. This in turn can result in a better fitting model. But, if this knowledge is unavailable and the researcher believes that there may be model misspecification, then the estimates returned from the Stability Informed Model may be biased. But researchers can also use the Stability Informed Model to conduct a sensitivity analyses to assess the degree of variation in the estimates across specifications. This practice can help researchers calibrate their trust in both the cross-sectional and stability-informed estimates.

Introducing the stim Package

To facilitate the estimation of the Stability Informed Model, we developed a software package—**stim—**using the R coding language. However, the Stability Informed Model can be estimated in any other latent model estimation software. Names of R packages are bolded, and names of both R functions and arguments are italicized. We assume basic knowledge of the coding language R and the R package **lavaan** (Rosseel, 2012).[22] [23]

stim Package Overview

The Stability Informed Model can be estimated by calling the function *stim()* which has the following arguments:

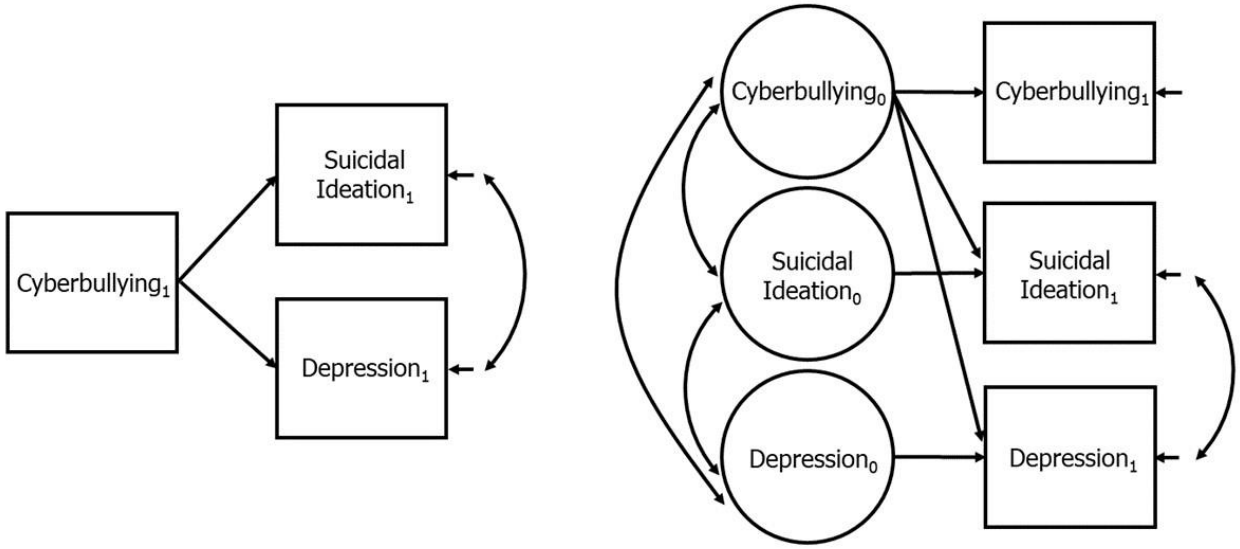
- *data*: A data frame with the measured variables. Not needed if input for *S* is provided.
- *S*: A covariance matrix for the measured variables. Not needed if input for *data* is provided.
- *n*: Number of observations in data set. Not needed if input for *data* is provided.
- *model*: A vector object with the cross-sectional model specification (in **lavaan** syntax).
- *stability*: A vector object with the stability value for each variable.

The *stim()* function converts the specified cross-sectional model into its corresponding Stability Informed Model (see Figure 22), incorporates the stability values into the parameter constraints, and estimates the model using the R package **lavaan**. The function's output is an object of type *stim*.[24] When the *summary()* function is used on a *stim* object, the stability-informed estimates are printed along with their p-values and standard errors. Additionally, the

[22] There are many tutorials and example coding scripts available for **lavaan**. For example, https://lavaan.ugent.be/tutorial/tutorial.pdf
[23] An Rmarkdown version of this section is available at https://osf.io/s6xk7
[24] More information on the output object can be found https://github.com/AnnaWysocki/stim/blob/main/README.md

lavaanSummary() function prints the **lavaan** output for each estimated model. The *stim* output object also has the syntax for the Stability Informed Model—model syntax for the **lavaan** function. This contains the syntax to specify the structural part of the Stability Informed Model as well as the parameter constraints for the auto-regressive paths and the latent correlations. Researchers can re-use this syntax if they wish to run the Stability Informed Model directly through the **lavaan** function.

Figure 22. *Converting a Cross-sectional Model into a Stability Informed Model.*



Note. A cross-sectional model (left panel) and its corresponding Stability Informed Model (right panel).

Below, we include an example for estimating the Stability Informed Model using empirical data from a study on the effect of experiencing cyberbullying on suicidal ideation and depression.

An Applied Example

The *stim()* function requires three pieces of information: 1) a cross-sectional data set (in the form of raw data or a covariance matrix), 2) stability estimates for each variable, and 3) a model specification.

For the following example, we use a covariance matrix of three variables—cyberbullying, suicidal ideation, and depression—from Mitchell and colleagues (2018). Since the data input is a covariance matrix, the sample size of the data set (n = 348) will need to be provided.

```
S <- matrix(c( 1,  .18, .19,
              .18,  1,  .38,
              .19, .38,  1), nrow = 3, ncol = 3)

colnames(S) <- rownames(S) <- c("c", "d", "s")
```

For the stability estimates, we searched the literature for longitudinal studies that measured at least one of the variables of interest. All stability estimates had a time lag between 1 year and 8 months.[25] Since we collected multiple stability estimates for each variable, we averaged them resulting in a single value for each variable.

```
# Create a vector containing the stability estimates
# Labels in the stability object should match column/row names
# in the S/data input
stability <- c(c = .69, d = .66, s = .7)
```

We were interested in the effect of cyberbullying on both depression and suicidal ideation. Since depression and suicidal ideation are likely impacted by factors other than cyberbullying, we allowed the residuals of depression and suicidal ideation to covary (See Figure 22 right panel).

[25] see https://osf.io/cu2dx for an overview of the stability literature search

Estimating a Stability Informed Model. The **stim** package is available on CRAN and can be installed and loaded with the following code:

```
install.packages('stim')
library(stim)
```

The input for the *model* argument must be an object with the cross-sectional model specified in **lavaan** syntax. The *stim()* function then translates the cross-sectional model into a Stability Informed Model and incorporates the stability estimates into the parameter constraints.

```
# Specify a model to estimate the stability-informed estimates of
# cyberbulling on depression and suicidal ideation

# Allow the residuals of depression and
# suicidal ideation to covary

model <- ' s ~ c   # outcome ~ predictor
           d ~ c

           s ~~ d # residual covariance
         '
```

After specifying inputs for the *data*, *stability*, and *model* arguments, the Stability Informed Model is estimated using the *stim()* function.

```
modelFit <- stim(S = S, n = 348, model = model,
                 stability = stability)
summary(modelFit)
```

```
## StIM: Stability Informed Models
## ------------------------------------
## ------------------------------------
##
## Variables (p): 3
## Sample Size (n): 348
## Estimated Parameters (q): 3
## Degrees of Freedom: 0
##
## ------------------------------------
## Model 1
##
## Stability:
##     c   s    d
##  0.69 0.7 0.66
##
##  Autoregressive Effects:
##       ARc       ARs       ARd
## 0.6909935 0.6730729 0.6346404
##
##  Cross Lagged Effects:
##  Effect Estimate Standard.Error P.Value
##    CLcs    0.147          0.045   0.001
##    CLcd    0.146          0.047   0.002
##
##  Residual Covariances:
##  Effect Estimate Standard.Error P.Value
##  RCovsd    0.161          0.034       0
##
## ------------------------------------
##
```

These results are consistent with there being a longitudinal cross-lagged coefficient of .146 of cyberbullying on suicidal ideation and .147 of cyberbullying on depression. These estimates are similar but slightly lower than the cross-sectional estimates, which were .19 and .18, respectively.

For the stability-informed estimates to be unbiased for their respective longitudinal effects, the stability values must be unbiased, the model must be specified correctly, and the assumption of stationarity must hold. We can assess how much an estimate varies within a plausible range of stability values by doing a sensitivity analysis in which we estimate a set of models where each model has a different set of stability values.

Conducting a Sensitivity Analysis. The *stim()* function estimates a model for each row in the *stability* argument input. For the sensitivity analysis, we created a range (+/- .1) around the stability estimates used in the previous model. This results in three plausible stability values for each variable, and 27 models (each with a different combination of stability values) to estimate. For example, the stability values for cyberbullying will be .57, .67, and .77.

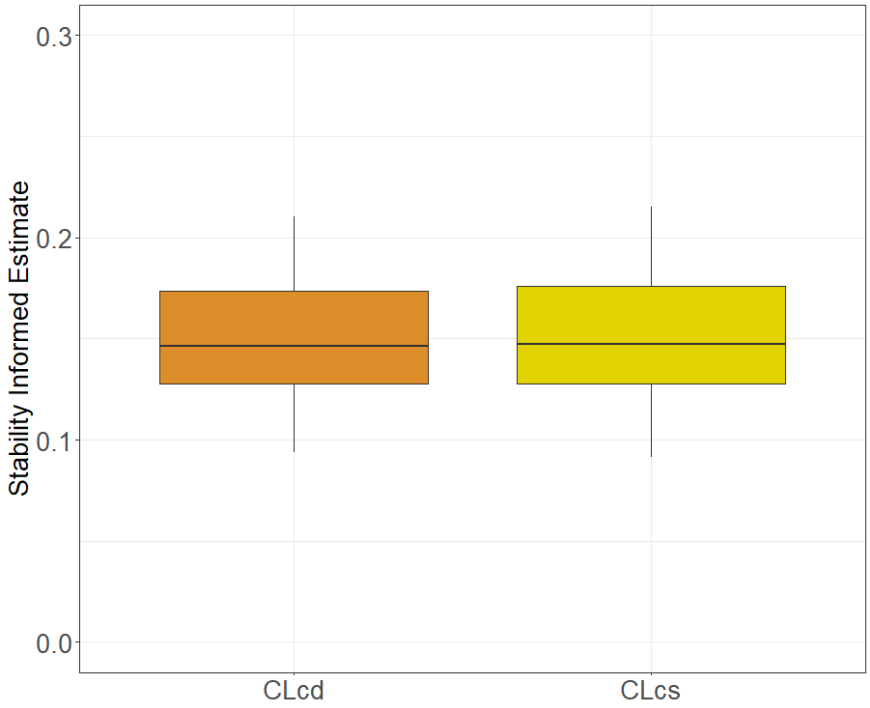
```
stabilityRange <- rbind(stability,
                        stability - .1,
                        stability + .1)

StabilityDf <- expand.grid(stabilityRange[,"c"], stabilityRange[,"d"], stabilityRange[,"s"])
colnames(StabilityDf) <- c("c", "d", "s")

head(StabilityDf) # show first 6 rows from stabiility data frame
```

```
##      c    d   s
## 1 0.69 0.66 0.7
## 2 0.59 0.66 0.7
## 3 0.79 0.66 0.7
## 4 0.69 0.56 0.7
## 5 0.59 0.56 0.7
## 6 0.79 0.56 0.7
```

```
sensitivityFit <- stim(S = S, n = 348, model = model, stability = StabilityDf)
```

The output object from the *stim()* function now has the results from 27 models which can be challenging to visualize in a table format. Instead, we can summarize the stability-informed estimates across the 27 models by plotting them (see Figure 22).

Figure 23. *Stability-informed Estimates across the 27 Models*



Based on our sensitivity analysis, we can see that both of the stability-informed estimates range from roughly .10 to .23 depending on the stability value inputs. However, note that regardless of the stability inputs both effects remain positive. This bolsters our conclusion that there is a positive longitudinal coefficient of cyberbulling on suicidal ideation and depression, over a time lag of about one year. To interpret this coefficient causally, we would have to consider the tenability of other assumptions such as whether there are any unaccounted common causes.

Conclusion

In this chapter we introduced the Stability Informed Model for estimating longitudinal coefficients from cross-sectional data, investigated the impact of several sources of

misspecification on the resulting coefficients, and introduced a software package to estimate the model.

We believe this model is a valuable contribution to the field of psychology and can provide better estimates of longitudinal processes, particularly when a process is stationary and researchers have reliable information about the structural model and variable stability. But the stability-informed estimates can also be heavily biased when the model assumptions are violated or when the model is structurally misspecified. Structural misspecification is an important source of bias as often the population model will be more complicated than the model that can be estimated within the proposed framework. We proposed two solutions: 1) constraining estimated coefficients to be equal to each other, and 2) setting fixed parameters to a value other than zero. But these are only viable solutions when a researcher has correct information about what the non-zero constrained value should be or which parameters should be constrained.

Another source of bias is stability misspecification. Currently, the Stability Informed Model allows a single stability value to be specified for each variable, and, if that value is not the true stability, the resulting estimates will be biased. But there will likely be a range of plausible stability values for each variable. An ideal solution would allow not only a range of stability values to be specified but also allow probabilities to be assigned to each of the values—values in the middle of the range may be more plausible than values in the extremes. As a next step, we intend to develop a Bayesian extension that estimates the Stability Informed Model using informative priors rather than rigid parameter constraints for the non-estimated parameters. As such, a prior distribution could be specified for each variable's stability where the distribution shape (e.g., a normal distribution) would specify the probabilities of the values in the range. A Bayesian model could also replace exact zero constraints with approximate zero constraints (e.g.,

a prior distribution with a mean of 0 and a small variance for a residual covariance). In this way, a Bayesian approach could ameliorate the issues of both stability and structural model misspecification.

The Stability Informed Model is not intended to replace longitudinal models, but rather intended as a model that can provide more information and in certain situations less biased estimates than cross-sectional models. More generally, we believe the Stability Informed Model motivates a framework for incorporating different sources of information into a single model. As such, we think this model may be able to address the more general problem in causal inference of missing variables. Measuring every variable involved in a psychological process is infeasible, but incorporating information about a relevant variable from previous studies may be more achievable. Although this approach is not common in psychology, it has the potential to allow researchers to build more complex and nuanced models that are better representations of the world.

Chapter 5. Discussion & Conclusion

Each of the chapters in this dissertation provide guidance or insight into an active question about the match between an estimate and its estimation approach or a theoretical target and its linked estimate. More specifically, these studies have informed how network models should be estimated and how causal models can be built. In Chapter 2, I explored how best to estimate network models by evaluating the performance of different estimation approaches. I found that all regularized approaches had concerning performance—low sensitivity, false positive rate increases as sample size increases, estimated sparsity seemingly unrelated to population sparsity. Ultimately, I suggest non-regularized methods be used to estimate low-dimensional psychological networks. In Chapter 3, I demonstrated how control variables can either help or harm estimates depending on the causal relations between the control(s), outcome, and predictor. I then developed a control variable selection framework to guide applied researchers as they select their set of variables to measure. In Chapter 4, I discussed the challenge when cross-sectional estimates are used to learn about longitudinal parameters. I then outlined a model that integrates longitudinal information into cross-sectional estimates for situations where longitudinal parameters are of interest.

But beyond these specific contributions, each study develops tools for a larger framework where methodology is evaluated for alignment with the research question across multiple steps. Each study underscores the importance of making methodological decisions based on existing theory and also provides tools and guidelines to do so. In Chapter 2, I disentangled estimate and estimator selection, and focused on comparing multiple estimators of the same estimate. The proposed framework in Chapter 3 encourages researchers to explicitly delineate the motivating model behind their estimate choice. By building such a model, researchers are encouraged to

make bold falsifiable claims that future researchers can build upon—a foundation for incremental progress and theory development. In Chapter 4, I proposed an approach to augment estimates with existing information. When studying complex systems, one study cannot measure and estimate all important parts of the system. In these cases, integrating estimates from multiple studies is vital, and Chapter 4 develops one approach to do so.

Broadly, each of these chapters center on variable selection and its impact on model building and selection. Chapter 2 demonstrates that when sample size exceeds the number of variables measured (i.e., a low-dimensional setting), the costs of regularization exceed the benefits. Chapter 3 delineates the cost of incorrect variable selection for causal inference and proposes a framework for improved control variable selection and justification. Chapter 4 points out that variables measured at a single timepoint can rarely recover longitudinal parameters and develops a statistical model to augment the set of measured variables. In summary, variable selection is a foundational step in research, and the chapters in this dissertation investigate the impact of a mismatch between the variables selected and the research goals and propose solutions to ameliorate these consequences.

Limitations and Future Directions

Although I believe these chapters meaningfully add to each area's body of research, there are a number of limitations that are important to note. Although the purpose of Chapter 2 was to evaluate different penalty selection methods, there are many more penalty selection methods than the ones compared here (see Kuismin and Sillanpää, 2017 for an over-view of different penalty selection methods). The included methods were chosen based on previous research demonstrating they were competitive with the default method in psychology (Liu et al., 2010;

Mohammadi & Wit, 2015), EBIC, and based on their ease of implementation for applied researchers (i.e., the availability of an R-package or code to fit these models).

In Chapter 3, the framework and examples did not discuss composite variables—a variable created by combining another set of measured variables. Composites are common in psychology (e.g., summing items on an extraversion questionnaire to obtain an individual's extraversion score) and have implications for causal inference. Crucially, the causal relation between the items being collapsed determines whether the causal estimate between a composite and another variable can be recovered. A future direction is to outline different kinds of composites, discuss how they impact causal inference, and provide guidelines on when and how causal inference can be achieved.

Finally, the Stability-Informed Model, outlined in Chapter 4, can produce unbiased longitudinal estimates from cross-sectional data. But this model relies on the external information being correct and the model being correctly specified, and the model is sensitive to misspecifications. A future direction is to estimate this model in a Bayesian framework using informative priors—a distribution of plausible values—for the non-estimated parameters rather than rigid parameter constraints. As such, a prior distribution could be specified for each variable's stability where the distribution shape (e.g., a normal distribution) would specify the probabilities of the values in the range. It is also possible to replace exact zero constraints with approximate zero constraints (e.g., a prior distribution with a mean of 0 and a small variance for a residual covariance) for the unestimated parameters that we would normally set at 0. In this way, a Bayesian approach could ameliorate the issues of both stability and model misspecification and return more robust and informative estimates.

Implications and Conclusion

Overall, these studies not only contribute to a general movement in psychology to explicitly and thoughtfully link methodology with research goals, but also contributes to the specific areas of network modeling and causal inference. My research comparing penalty parameter methods (Wysocki & Rhemtulla, 2019) alongside others (Willams & Rast, 2020; Williams, Rhemtulla, Wysocki & Rast, 2019) has influenced the field to move away from regularized methods when estimating network models. Additionally, deliberately choosing and justifying control variables should be a central part of every published analysis, and the wide-spread adoption of the framework proposed in Chapter 3 could both improve and clarify the theoretical models in psychology. Finally, the Stability Informed Model motivates a framework for incorporating different sources of information into a single model which will be foundational for causal inference in psychology.

Bibliography

Allison, P. D. (2005, August). *Causal Inference with Panel Data* [Paper presentation]. Proceedings of the Annual Meeting of the American Sociology Association.

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, *40*(1), 37–47. https://doi.org/10.2307/2094445

Atinc, G., Simmering, M. J., & Kroll, M. J. (2012). Control variable use and reporting in macro and micro management research. *Organizational Research Methods*, *15*(1), 57–74. https://doi.org/10.1177/10944281103977

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Barrett, M. (2021). *ggdag: Analyze and create elegant directed acyclic graphs.* R package version 0.2.10, https://CRAN.R-project.org/package=ggdag

Beard, C., Millner, A. J., Forgeard, M. J. C., Fried, E. I., Hsu, K. J., Treadway, M., . . . Björgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological Medicine*, *42*(2), 407–420. https://doi.org/10.1017/S0033291716002300

Becker, T. E. (2005). Potential problems in the statistical control of variables in organizational research: A qualitative analysis with recommendations. *Organizational Research Methods*, *8*(3), 274–289. https://doi.org/10.1177/1094428105278021

Becker, T. E., Atinc, G., Breaugh, J. A., Carlson, K. D., Edwards, J. R., & Spector, P. E. (2016). Statistical control in correlational studies: 10 essential recommendations for organizational researchers. *Journal of Organizational Behavior, 37*(2), 157–167. https://doi.org/10.1002/job.2053

Bernerth, J. B., & Aguinis, H. (2016). A critical review and best-practice recommendations for control variable usage. *Personnel Psychology*, 69(1), 229–283. https://doi.org/10.1111/peps.12103

Bien, J., & Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4), 807–820. https://doi.org/10.1093/biomet/asr054

Borsboom, D. (2008). Latent variable theory. *Measurement, 6*, 25-53. https://doi.org/10.1080/15366360802035497

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. https://doi.org/10.1002/wps.20375

Borsboom, D., & Cramer, A. O. (2013). Network Analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A., Wysocki, A. C., van Borkulo, C. D., van Bork, R. & Waldorp, L. J. (2021b). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primer*. *1*(1). 1-18. https://doi.org/10.1038/s43586-021-00055-w

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203. https://doi.org/10.1037/0033-295X.110.2.203

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021a). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756-766. https://doi.org/10.1177/1745691620969647

Breaugh, J. A. (2006). Rethinking the control of nuisance variables in theory testing. *Journal of Business and Psychology*, *20*(3), 429–443. https://doi.org/10.1007/s10869-005-9009-y

Breaugh, J. A. (2008). Important considerations in using statistical procedures to control for nuisance variables in non-experimental studies. *Human Resource Management Review, 18*(4), 282–293. https://doi.org/10.1016/j.hrmr.2008.03.001

Briganti, G., Kempenaers, C., Braun, S., Fried, E. I., & Linkowski, P. (2018). Network analysis of empathy items from the interpersonal reactivity index in 1973 young adults. *Psychiatry Research*, *265*, 87–92. https://doi.org/10.1016/j.psychres.2018.03.082

Bryant, R. A., Creamer, M., O'Donnell, M., Forbes, D., McFarlane, A. C., Silove, D., & Hadzi-Pavlovic, D. (2017). Acute and chronic posttraumatic stress symptoms in the emergence of posttraumatic stress disorder a network analysis. *JAMA Psychiatry*, 74(2), 135–142. doi: 10.1001/jamapsychiatry.2016 .3470

Carlson, K. D., & Wu, J. (2012). The illusion of statistical control: Control variable practice in management research. *Organizational Research Methods, 15*(3), 413–435. https://doi.org/10.1177/1094428111428817

Chen, J., & Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces model selection. *Biometrika*, *95*(3), 759–771. https://doi.org/10.1093/biomet/asn034

Chen, J., & Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica*, 555-574. http://www.jstor.org/stable/24310025

Chetverikov, D., Liao, Z., & Chernozhukov, V. (2021). On cross-validated lasso in high dimensions. *The Annals of Statistics*, *49*(3), 1300-1317. https://doi.org/10.1214/20-AOS2000

Cinelli, C., Forney, A., & Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research*, 0(0). https://doi.org/10.1177/0049124122109955

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Erlbaum.

Colonescu, C. (2016). Principles of econometrics with R. https://bookdown.org/ccolonescu/RPoE4/

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, *54*, 13–29. https://doi.org/10.1016/j.jrp.2014.07.003

Cui, X., Guo, W., Lin, L., & Zhu, L. (2009). Covariate-adjusted nonlinear regression. *The Annals of Statistics*, *37*(4), 1839–1870. https://doi.org/10.1214/08-AOS627

Dablander, F. (2020). An introduction to causal inference. *PsyArXiv*. https://doi.org/10.31234/osf.io/b3fkw

Devine, R. T., & Apperly, I. A. (2022). Willing and able? Theory of mind, social motivation, and social competence in middle childhood and early adolescence. *Developmental Science, 25*(1), 1–14. https://doi.org/10.1111/desc.13137

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*(1), 40–57. https://doi.org/10.1037/0021-9010.91.1.40

Edwards, J. R. (2008). To prosper, organizational psychology should . . . overcome methodological barriers to progress. *Journal of Organizational Behavior, 29*(4), 469–491. https://doi.org/10.1002/job.529

Edwards, D. (2012). *Introduction to graphical modeling*. Springer Science & Business Media.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics, 32*(2), 407–499. https://doi.org/10.1214/009053604000000067

Efron, B., & Tibshirani, R. (1993). An introduction to the bootstrap. *Journal of the American Statistical Association, 89*, 436. doi:10.1007/978-1-4899-4541-9

Eichler, M., & Didelez, V. (2010). On Granger causality and the effect of interventions in time series. *Lifetime Data Analysis, 16*(1), 3–32. https://doi.org/10.1007/s10985-009-9143-3

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology, 40*, 31–53. https://doi.org/10.1146/annurev-soc-071913-043455

Epskamp, S. (2016). Brief report on estimating regularized gaussian networks from continuous and ordinal data. *arXiv*. http://arxiv.org/abs/1606.05771

Epskamp, S. (2018, 5). *New features in qgraph 1.5*. [Blogpost]. http://psychosystems.org/qgraph_1.5

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50(1), 195–212. https://doi.org/10.3758/s13428-017-0862-1

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*, 48(4). https://doi.org/10.18637/jss.v048.i04

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*(4), 617. https://doi.org/10.1037/met0000167

Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2018). Network psychometrics. *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*, 953-986.

Foster, E. M. (2010). Causal inference and developmental psychology. *Developmental Psychology*, *46*(6), 1454. https://doi.org/10.1037/a0020204

Foygel, R., & Drton, M. (2010). Extended bayesian information criteria for gaussian graphical models. *Advances in Neural Information Processing Systems*, 23.

Fried, E. I., van Borkulo, C. D., Cramer, A. O., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders networks of problems: A review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*, 52(1), 1–10. https://doi.org/10.1007/s00127-016-1319-z

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–24. https://doi.org/10.18637/jss.v033.i01

Friedman J, Hastie T, Tibshirani R. (2019). *glasso: Graphical Lasso: Estimation of Gaussian Graphical Models*. R package version 1.11, https://CRAN.R-project.org/package=glasso

Fu, W., & Knight, K. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*(5), 1356–1378. https://doi.org/10.1214/aos/1015957397

Gauraha, N., & Swapan, P. (2018). Constraints and conditions: The lasso oracle-inequalities. *arXiv*. https://arxiv.org/abs/1603.06177

Gelman, A. (2019, January 25). When doing regression (or matching, or weighting, or whatever), don't say "control for," say "adjust for." *Statistical Modeling, Causal Inference, and Social Science*. https://statmodeling.stat.columbia.edu/2019/01/25/regression-matching-weighting-whatever-dont-say-control-say-adjust/

Gollob, H. F., & Reichardt, C. S. (1985). Building time lags into causal models of cross-sectional data. In A. S. Association (Ed.), *Proceedings of the social statistics section of the american statistical association* (pp. 165–170). Washington, D.C..

Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development*, *58*(1), 80–92. https://doi.org/10.2307/1130293

Göllner, R., Damian, R. I., Rose, N., Spengler, M., Trautwein, U., Nagengast, B., & Roberts, B. W. (2017). Is doing your homework associated with becoming more conscientious? *Journal of Research in Personality*, *71*, 1–12. https://doi.org/10.1016/j.jrp.2017.08.007

Granger, C. W. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, *2*, 329–352. https://doi.org/10.1016/0165-1889(80)90069-X

Greenland, S. (1990). Randomization, statistics, and causal inference. *Epidemiology*, *1*(6), 421–429. https://doi.org/10.1097/00001648-199011000-00003

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 37-48. https://doi.org/10.1097/00001648-199901000-00008

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

Grosz, M., Ayaita, A., Arslan, R. C., Buecker, S., Ebert, T., Müller, S., ... Rohrer, J. M. (2023). Natural experiments: Missed opportunities for causal inference in psychology. *PsyArXiv*. https://doi.org/10.31234/osf.io/dah3q

Gürbüz, A. (2007). An assessment of the effect of educational level on the job satisfaction from the tourism sector point of view. *Dogus University Dergisi, 8*, 36–46.

Harring, J., McNeish, D., & Hancock, G. (2017). Using phantom variables in structural equation modeling to assess model sensitivity to external misspecification. *Psychological Methods, 22*(4), 616. https://doi.org/10.1037/met0000103

Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2019). *Introduction to econometrics with R.* University of Duisburg-Essen.

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102. https://doi.org/10.1037/a0038889

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*(4), 408–420. https://doi.org/10.1080/03637750903310360

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., & Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models-A review. *BioSystems*, 96(1), 86–103. https://doi.org/10.1016/j.biosystems.2008.12.004

Hernán, M. A., Hernández-Díaz, S., Werler, M. M., & Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defect epidemiology. *American Journal of Epidemiology*, *155*(2), 176–184. https://doi.org/10.1093/aje/155.2.176

Hernán, M. A., Hernández-Díaz, S., Robins, J. M. (2004) A structural approach to selection bias. *Epidemiology* 15, 615-625. https://doi.org/10.1097/01.ede.0000135174.63482.43

Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Stampfer, M. J., ... & Robins, J. M. (2008). Observational studies analyzed like randomized experiments: An application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*, *19*(6), 766. https://doi.org/10.1097/EDE.0b013e3181875e61

Hitchcock, C. (2001). A tale of two effects. *The Philosophical Review, 110*(3), 361-396.

Homrighausen, D., & McDonald, D. J. (2014). Leave-one-out cross-validation is risk consistent for lasso. *Machine Learning, 97,* 65-78. https://doi.org/10.1007/s10994-014-5438-z

Homrighausen, D., & McDonald, D. J. (2017). Risk consistency of cross-validation with lasso-type procedures. *Statistica Sinica*, 1017-1036.

Hsu, C. W., Sinay, M. S., & Hsu, J. S. (2012). Bayesian estimation of a covariance matrix with flexible prior specification. *Annals of the Institute of Statistical Mathematics, 64*, 319-342. https://doi.org/10.1007/s10463-010-0314-5

Imai, K., & Kim, I. S. (2016). *When should we use linear fixed effects regression models for causal inference with panel data?* Princeton University.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5*(5), 602–619. https://doi.org/10.1177/0193841X810050050

Kim, Y., Sommet, N., Na, J., & Spini, D. (2022). Social Class—Not Income Inequality—Predicts Social and Institutional Trust. *Social Psychological and Personality Science, 13*(1), 186–198. https://doi.org/10.1177/1948550621999272

Kim, Y., & Steiner, P. M. (2021). Causal graphical views of fixed effects and random effects models. *British Journal of Mathematical and Statistical Psychology*, *74*(2), 165–183. https://doi.org/10.1111/bmsp.12217

Kossakowski, J. J., Epskamp, S., Kieffer, J. M., van Borkulo, C. D., Rhemtulla, M., & Borsboom, D. (2016). The application of a network approach to Health-Related Quality of Life (HRQoL): Introducing a new method for assessing HRQoL in healthy adults and cancer patients. *Quality of Life Research*, 25(4), 781–792. https://doi.org/10.1007/s11136-015-1127-z

Kuismin, M., & Sillanpää, M. J. (2016). Use of Wishart Prior and Simple Extensions for Sparse Precision Matrix Estimation. *PloS one*, *11*(2), e0148171. https://doi.org/10.1371/journal.pone.0148171

Kuismin, M. O., & Sillanpää, M. J. (2017). Estimation of covariance and precision matrix, network structure, and a view toward systems biology. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(6), e1415. https://doi.org/10.1002/wics.1415

Kim, Y., & Steiner, P. M. (2021). Causal graphical views of fixed effects and random effects models. *British Journal of Mathematical and Statistical Psychology*, *74*(2), 165-183. https://doi.org/10.1111/bmsp.12217

Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Clarendon Press.

Liu, H., Roeder, K., & Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. *Advances in Neural Information Processing Systems*, 1432–1440.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics, 41*(6), 2948–2978. https://doi.org/10.1214/13-AOS1169

Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review, 86*(3), 532-565. https://doi.org/10.1177/000312242110041

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods, 12*(1), 23–44. https://doi.org/10.1037/1082-989X.12.1.23

Maxwell, S. E., Cole, D. A., & Mitchell, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research, 46*(5), 816–841. https://doi.org/10.1080/00273171.2011.606716

Maziarz, M. (2015). A review of the Granger-causality fallacy. *The Journal of Philosophical Economics: Reflections on Economic and Social Issues, 8*(2), 86-105. https://doi.org/10.46298/jpe.10676

McElreath, R., & Smaldino, P. E. (2015). Replication, communication, and the population dynamics of scientific discovery. *PloS one, 10*(8), e0136088. https://doi.org/10.1371/journal.pone.0136088

McNally, R. J. (2016). Can network analysis transform psychopathology? *Behaviour Research and Therapy, 86*, 95–104. https://doi.org/10.1016/j.brat.2016.06.006

McNally, R. J., Robinaugh, D. J., Wu, G. W., Wang, L., Deserno, M. K., & Borsboom, D. (2015). Mental disorders as causal systems: A network approach to posttraumatic stress disorder. *Clinical Psychological Science*, *3*(6), 836-849. https://doi.org/10.1177/2167702614553230

McNamee, R. (2003). Confounding and confounders. *Occupational and Environmental Medicine, 60*(3), 227–234. http://dx.doi.org/10.1136/oem.60.3.227

McNamee, R. (2005). Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine, 62*(7), 500–506. http://dx.doi.org/10.1136/oem.2002.001115

Meehl, P. E. (1970). Nuisance variables and the ex post facto design. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science* (Vol. 4, pp. 373–402). University of Minnesota Press

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, *34*(3), 1436–1462. https://doi.org/10.1214/009053606000000281

Mitchell, S. M., Seegan, P. L., Roush, J. F., Brown, S. L., Sustaíta, M. A., & Cukrowicz, K. C. (2018). Retrospective cyberbullying and suicide ideation: The mediating roles of depressive symptoms, perceived burdensomeness, and thwarted belongingness. *Journal of Interpersonal Violence, 33*(16), 2602–2620. https://doi.org/10.1177/0886260516628291

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, W. M., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences, USA, 108*(7), 2693–2698. https://doi.org/10.1073/pnas.101007610

Mohammadi, A., & Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis, 10*(1),109–138. https://doi.org/10.1214/14-BA889

Morabia, A. (2011). History of the modern epidemiological concept of confounding. *Journal of Epidemiology & Community Health, 65*(4), 297–300. http://dx.doi.org/10.1136/jech.2010.112565

National Center for Education Statistics. (2019). Young adult educational and employment outcomes by family socioeconomic status. *Condition of Education.* https://nces.ed.gov/programs/coe/indicator/tbe

O'Laughlin, K. D., Martin, M. J., & Ferrer, E. (2018). Cross-Sectional Analysis of Longitudinal Mediation Processes. *Multivariate Behavioral Research, 53*(3), 375–402. https://doi.org/10.1080/00273171.2018.1454822

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. https://doi.org/10.1093/biomet/82.4.669

Pearl, J. (1998). Why there is no statistical test for confounding, why many think there is, and why they are almost right. *Escholarship*. https://escholarship.org/uc/item/2hw5r3tm

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Pearl, J. (2011). Invited commentary: Understanding bias amplification. *American Journal of Epidemiology, 174*(11), 1223-1227. https://doi.org/10.1093/aje/kwr352

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer.* John Wiley & Sons.

The Pell Institute. (2018). *Indicators of higher educational equity in the United States.*

Pereira-Morales, A. J., Adan, A., & Forero, D. A. (2019). Network analysis of multiple risk factors for mental health in young Colombian adults. *Journal of Mental Health*, 28(2), 153-160. https://doi.org/10.1080/09638237.2017.1417568

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms.* The MIT Press.

Pourhoseingholi, M. A., Baghestani, A. R., & Vahedi, M. (2012). How to control confounding effects by statistical analysis. *Gastroenterology and Hepatology From Bed to Bench*, *5*(2), 79–83.

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163. https://doi.org/10.2307/271063

Ravikumar, P., Wainwright, M. J., & Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ 1-regularized logistic regression. *Annals of Statistics, 38*(3), 1287-1319. https://doi.org/10.1214/09-AOS691

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods, 25*(1), 30. https://doi.org/10.1037/met000022

Roberts, B. W., & Wood, D. (2006). Personality development in the context of the neo-socioanalytic model of personality. In D. Mroczek & T. Little (Eds.), *Handbook of personality development* (pp. 11–39). Erlbaum.

Roberts, B. W., Wood, D., & Smith, J. L. (2004). Evaluating Five Factor Theory and social investment perspectives on personality trait development. *Journal of Research in Personality, 39*(1), 166–184. https://doi.org/10.1016/j.jrp.2004.08.002

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

Rohrer, J. (2019, April 16). Longitudinal data don't magically solve causal inference. *The 100% CI.* the100.ci/2019/04/16/longitudinal-data

Rohrer, J. M., Hünermund, P., Arslan, R. C., & Elson, M. (2022). That's a lot to PROCESS! Pitfalls of popular path models. *Advances in Methods and Practices in Psychological Science, 5*(2), 25152459221095827. https://doi.org/10.1177/25152459221095827

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Ross, C. E., & Reskin, B. F. (1992). Education, control at work, and job satisfaction. *Social Science Research, 21*(2), 134–148. https://doi.org/10.1016/0049-089X(92)90012-6

Savalei, V. (2019). A comparison of several approaches for controlling measurement error in small samples. *Psychological Methods, 24*(3), 352–370. https://doi.org/10.1037/met0000181

Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement, 23*(1), 153–158. https://doi.org/10.1177/001316446302300 11

Shear, B. R., & Zumbo, B. D. (2013). False positives in multiple regression: Unanticipated consequences of measurement error in the predictor variables. *Educational and Psychological Measurement, 73*, 733–756. https://doi.org/10.1177/0013164413487738

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-STS330

Simonsohn, U. (2019, November). Interaction effects need interaction controls. *Datacolada*. http://datacolada.org/80

Slominski, L., Sameroff, A., Rosenblum, K., & Kasser, T. (2011). Longitudinal predictors of adult socioeconomic attainment: The roles of socioeconomic status, academic

competence, and mental health. *Development and Psychopathology*, *23*(1), 315–324. https://doi.org/10.1017/S0954579410000829

Smaldino, P. E. (2016). Not even wrong: Imprecision perpetuates the illusion of understanding at the cost of actual understanding. *Behavioral and Brain Sciences*, *39*, e163. https://doi.org/10.1017/S0140525X1500151X, e163

Smaldino, P. E. (2017). Models are stupid, and we need more of them. *Computational Social Psychology*, 311-331.

Smith, G. T., Fischer, S., & Fister, S. M. (2003). Incremental validity principles in test construction. *Psychological Assessment*, *15*(4), 467–477. https://doi.org/10.1037/1040-3590.15.4.467

Steiner, P. M., & Kim, Y. (2016). The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. *Journal of Causal Inference*, *4*(2), 20160009. https://doi.org/10.1515/jci-2016-0009

Sutin, A. R., Costa, P. T., Jr., Miech, R., & Eaton, W. W. (2009). Personality and career success: Concurrent and longitudinal relations. *European Journal of Personality*, *23*(2), 71–84. ttps://doi.org/10.1002/per.704

Tibshirani, R., & Wasserman, L. (2015). Sparsity and the lasso. *Statistical Machine Learning*, 1-15.

U.S. Bureau of Labor Statistics. (2020). *Unemployment rates and earnings by educational attainment.* https://www.bls.gov/emp/chart-unemployment-earnings-education.htm

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

VanderWeele, T. J., & Shpitser, I. (2013). On the definition of a confounder. *Annals of Statistics*, *41*(1), 196–220.

Wang, Y. A., & Eastwick, P. W. (2020). Solutions to the problems of incremental validity testing in relationship science. *Personal Relationships*, *27*(1), 156–175. https://doi.org/10.1111/pere.12309

Wansbeek, T., & Meijer, E. (2001). Measurement error and latent variables. *A Companion to Theoretical Econometrics. Oxford: Basil Blackwell*, 162-179.

Williams, D. R., & Rast, P. (2020). Back to the basics: Rethinking partial correlation network methodology. *British Journal of Mathematical and Statistical Psychology*, *73*(2), 187-212. https://doi.org/10.1111/bmsp.12173

Williams, D. R., Rhemtulla, M., Wysocki, A. C., & Rast, P. (2019). On nonregularized estimation of psychological networks. *Multivariate Behavioral Research*, *54*(5), 719-750. https://doi.org/10.1080/00273171.2019.1575716

Wilms, R., Mäthner, E., Winnen, L., & Lanwehr, R. (2021, 12). *Omitted variable bias: A threat to estimating causal relationships* (Vol. 5). Elsevier B.V. https://doi.org/10.1016/j.metip.2021.100075

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLOS ONE*, *11*(3), e0152719. https://doi.org/10.1371/journal.pone.0152719

Westreich, D., & Greenland, S. (2013). The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, *177*(4), 292–298. https://doi.org/10.1093/aje/kws412

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Wilmot, M. P., & Ones, D. S. (2019). A century of research on conscientiousness at work. *Proceedings of the National Academy of Sciences, USA*, *116*(46), 23004–23010. https://doi.org/10.1073/pnas.1908430116

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yu, Y., & Feng, Y. (2014). Modified cross-validation for penalized high-dimensional linear regression models. *Journal of Computational and Graphical Statistics*, *23*(4), 1009–1027. https://doi.org/10.1080/10618600.2013.849200

Wysocki, A. C. & Rhemtulla M. (2021). Incorporating Stability Information into Cross-sectional Estimates. *Multivariate Behavioral Research.* 57(1), 168-169. (Abstract).

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, 299–313.

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, *7*, 2541–2563. https://doi.org/10.1109/TIT.2006.883611

Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, *13*, 1059–1062. https://doi.org/10.1002/aur.1474.Replication

Zhu, Y., & Cribben, I. (2018). Graphical models for functional connectivity networks: Best methods and the autocorrelation issue. *Brain Connectivity*, 8(3), 139–165. https://doi.org/10.1089/brain.2017.0511