That's a lot to PROCESS! Pitfalls of Popular Path Models

Julia M. Rohrer,[1] Paul Hünermund,[2] Ruben C. Arslan,[3] & Malte Elson[4]

Path models to test claims about mediation and moderation are a staple in psychology. But applied researchers sometimes do not understand the underlying causal inference problems and thus endorse conclusions that rest on unrealistic assumptions. In this article, we aim to provide a clear explanation for the limited conditions under which standard procedures for mediation and moderation analysis can succeed. We discuss why reversing arrows or comparing model fit indices cannot tell us which model is the right one, and how tests of conditional independence can at least tell us where our model goes wrong. Causal modeling practices in psychology are far from optimal but may be kept alive by domain norms which demand that every article makes some novel claim about processes and boundary conditions. We end with a vision for a different research culture in which causal inference is pursued in a much slower, more deliberate and collaborative manner.

Psychologists often do not content themselves with claims about the mere existence of effects. Instead, they strive for an understanding of the underlying processes and potential boundary conditions. The PROCESS macro (Hayes, 2017) has been an extraordinarily popular tool for these purposes, as it empowers users to run mediation and moderation analyses—as well as any combination of the two—with a large number of pre-programmed model templates in a familiar software environment.

However, psychologists' enthusiasm for PROCESS models may sometimes outpace their training. Even though most psychological researchers are aware that a simple correlation does not equal causation, there is confusion about how this affects more complex models fitted to (mostly) cross-sectional data. As a result, there seems to be a general lack of awareness of the assumptions underlying these models, an excess of unjustified

[1] University of Leipzig. julia.rohrer@posteo.de
[2] Copenhagen Business School
[3] Max Planck Institute for Human Development, Berlin
[4] Ruhr University Bochum

We thank Stefan Schmukle and Nick Brown for their helpful feedback on this manuscript. The subheading "Everything in Moderation" was stolen from a tweet by Stuart Ritchie.

conclusions, and the mistaken impression that a great deal of sophistication and deep understanding have been reached.[5]

With this article, we aim to improve practices involving path models such as the ones frequently fitted with the help of PROCESS. We start with a so-called conditional process model, which combines mediation and moderation, and work our way back to the underlying assumptions. We will provide non-technical explanations so that readers can develop an intuition for the intricacies of such path models. We will then discuss matters of model selection—how can we know that we got the model right?—and conclude with our vision for a different research process resulting in better causal claims.

Due PROCESS

The points that we highlight throughout this manuscript consider common modeling practices within psychology. These practices are frequently realized with the help of PROCESS, and we put emphasis on this particular software in the hope that its numerous users will feel addressed. However, these concerns generalize to other implementations of the same conceptual models. Structural Equation Models (SEM) differ from PROCESS models in some specifics (e.g., they use different estimators and allow for latent variable modeling) but the same causal inference problems apply.

Furthermore, many of our points mirror those made by Andrew F. Hayes, author of the PROCESS textbook (2017). The book is an introduction to regression analysis rather than an introduction to causal inference, and any conflation of the two can lead to confusion (Chen & Pearl, 2013). Nonetheless, causal inference issues are brought up repeatedly in Hayes (2017). Yet many articles relying on PROCESS still make claims that seem hard to defend, and thus we suspect that users have not fully understood or absorbed the warnings in the textbook (or they may have simply not read them). In any case, we think that applied researchers will profit from a straightforward and concise explanation of the assumptions that go into popular path models.

[5] These concerns are only made worse if, additionally, researchers search for and only report effects that reach the threshold for statistical significance (Götz et al., 2020). However, here we will not discuss such questionable research practices, preferring instead to focus on causal inference problems.

The Conditional Process Model

Consider the conceptual diagram in Figure 1. It depicts a path model in which an independent variable has an effect on a dependent variable via a mediator, and one of the mediation paths is moderated.[6] For example, researchers may hypothesize that social support has an effect on task performance among athletes. Supposedly, this effect is partly mediated via self-efficacy: Social support leads athletes to believe in themselves, and this improves performance. But this mediation may only apply to highly stressed athletes. Among more relaxed athletes, social support may not be a salient source of self-efficacy. We adapted this substantive illustration from Rees and Freeman (2009) because it was highlighted in Hayes (2017).

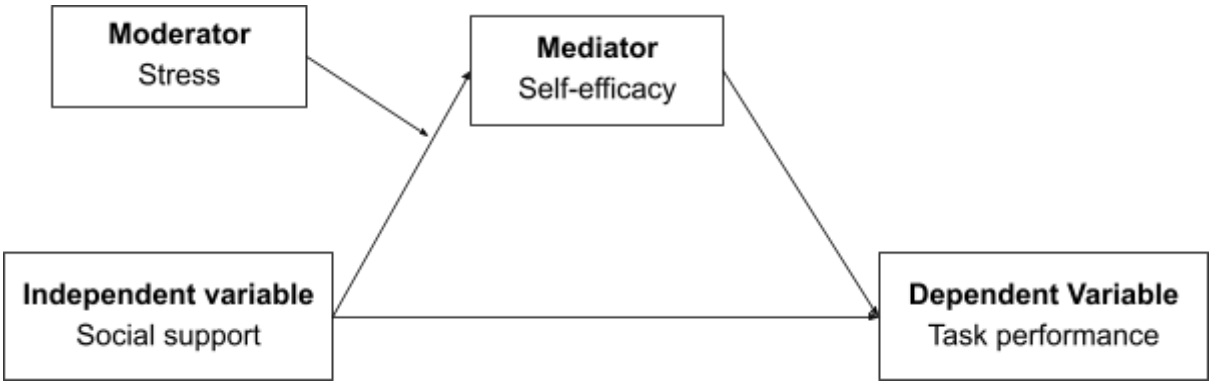

Figure 1. A conceptual diagram of a conditional process model in which a mediated path is affected by a moderator.

When we are testing such hypotheses with the help of path models applied to observational data, we are in the business of causal inference on the basis of observational data (see, e.g., Rohrer, 2018 for an introduction). One may object that such path models are employed for other purposes, such as description or prediction—but unfortunately, those models are *too complex* to result in useful descriptions (see Foster, 2010b for a similar argument), and *not complex enough* to result in useful predictions (see Westfall & Yarkoni, 2016 for an introduction to machine learning). Further, reviews of path models in different

[6] We will apply this standard mediation terminology throughout the manuscript. However, if the independent variable was not experimentally manipulated, it may not be independent at all (i.e., it may be confounded with other variables).

literatures from the social and behavioral sciences conclude that causal inferences are made or implied routinely (e.g., Fiedler et al., 2018; Wood et al., 2008).

Causal inference on the basis of observational data may make psychologists uncomfortable, as the field tends to emphasize randomized experiments as the best if not only way to arrive at valid causal conclusions. Undoubtedly, there is a lot that speaks in favor of randomized experiments and other predominantly design-based approaches to causal inference (such as natural experiments, e.g. Dunning, 2012). However, once we have settled on a model such as the one depicted in Figure 1, or in general once we are interested in mediation, our approach will necessarily be more strongly model-based.

And it only makes sense to talk about and interpret mediation from a causal perspective; from a strictly statistical perspective the phenomenon is indistinguishable from confounding (MacKinnon et al., 2000). To make the most of causal inference on the basis of observational data, it is best to take the bull by the horns while remaining transparent about the underlying assumptions, rather than resorting to ambiguous language that obscures the goal of the analysis (Grosz et al., 2020). Somewhat ironically, explicit causal language may prompt readers to be *more* careful when evaluating whether conclusions are appropriate (Alvarez-Vargas et al., 2020).

Under what conditions can we fit the model depicted in Figure 1 and give a substantive interpretation to the resulting coefficients? How do we successfully identify the causal effects of interest? Existing formalized frameworks allow for a precise articulation of the underlying assumptions (e.g., directed acyclic graphs, the Rubin causal model, see Morgan & Winship, 2015 for a helpful introduction). Here, we will use a more informal approach—starting from a single arrow and moving on to more complex claims about mediation and moderation. Many of the assumptions going into the model will be irrefutable; they cannot be disproved (let alone proved) by observable information (Manski, 2009), and thus demand that researchers take a stance and either accept or refuse them.

An Arrow is an Arrow is an Arrow

One can determine the assumptions under which such a model successfully identifies the causal effects by considering every single arrow it contains. For example, let us start with the arrow pointing from the independent variable to the mediator, Social support -> Self-efficacy. This single-headed arrow represents a *causal* effect of social support on

self-efficacy, and we can estimate the corresponding effect if we can rule out (1) confounding and (2) reverse causality.

To rule out confounding, we need to ensure that any possible variable that causally affects *both* social support and self-efficacy, any confounder, is taken into account (see Figure 2, Panel A for an example of one confounder). In PROCESS, this can be done by including it as a covariate. It is important that the association with the covariate is modeled appropriately (i.e., if the effect of the covariate is non-linear, it needs to be modeled non-linearly); otherwise, so-called residual confounding remains a problem. Furthermore, the covariate should be measured without measurement error, otherwise, residual confounding once again remains a problem (Westfall & Yarkoni, 2016). Of course, perfect or even just precise measurement may often be unachievable in psychology, so it may be necessary to apply latent variable modeling (see again Westfall & Yarkoni, 2016 although it should be noted that this cannot be implemented in PROCESS).[7] Imperfect measurement can also be directly incorporated into the causal graph (Kuroki & Pearl, 2014).

[7] One workaround that we have encountered in the literature is a two-step approach in which, in a first step, researchers extract estimates of the latent variable (e.g., through factor analysis), and, in a second step, these estimates are entered into a PROCESS model. Unfortunately, this approach ignores the uncertainty in the estimated values of the latent variable, which means that the standard errors will be too small—thus, the resulting conclusion may still be a false positive due to residual confounding.

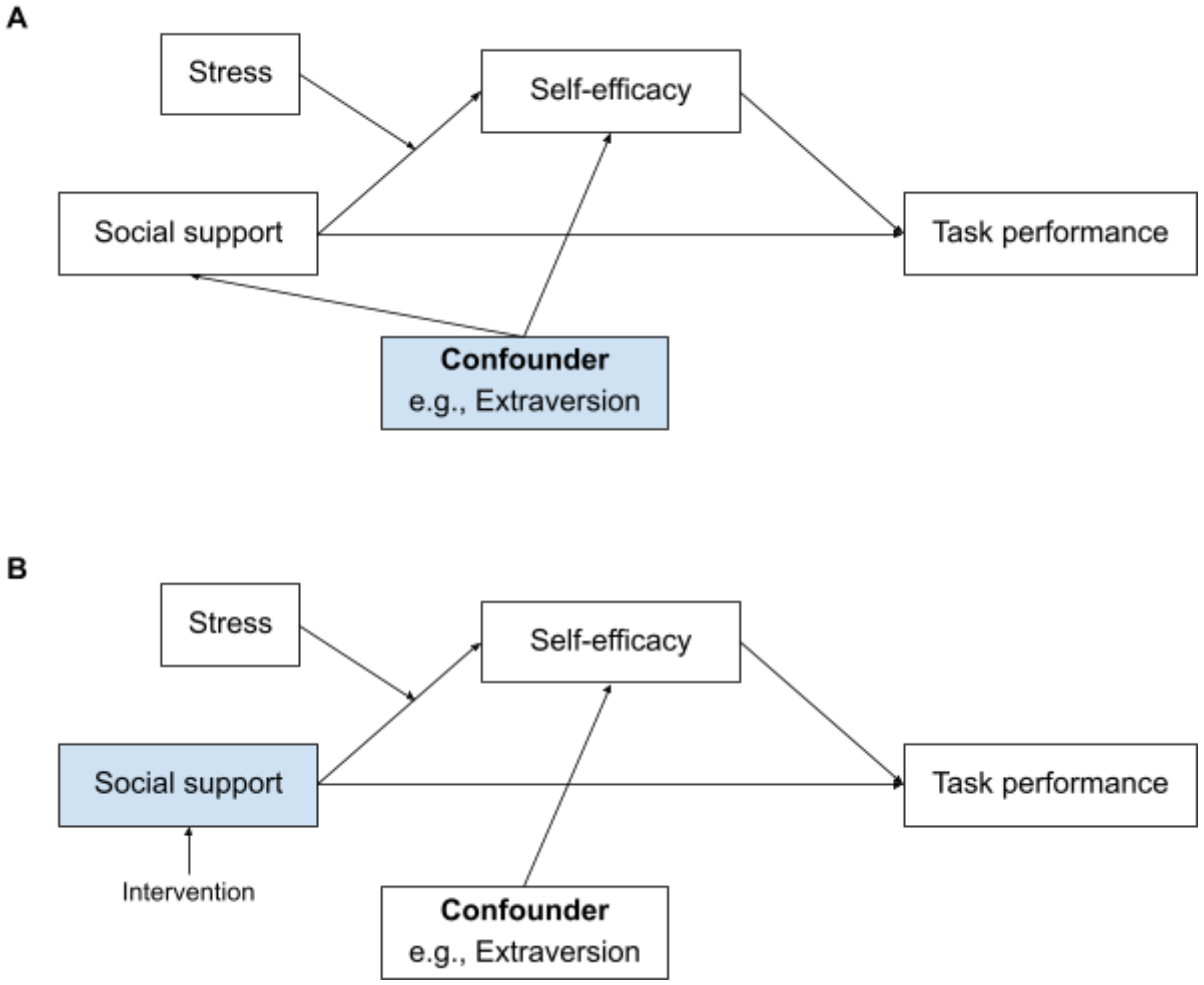

Figure 2. Modifications of the conditional process model from Figure 1. Panel A: A confounder between independent variable and mediator will bias the estimated path. Panel B: If we can intervene on the independent variable, the confounder is no longer a problem and we can estimate the causal effects of the independent variable on other variables.

Reverse causality can sometimes be ruled out by temporal order. Tomorrow's self-efficacy cannot have a causal effect on today's social support—but note that temporal order does *not* rule out confounding. If yesterday's self-efficacy had a causal effect on today's social support as well as an effect on tomorrow's self-efficacy, this would result in a spurious association between today's social support and tomorrows self-efficacy (i.e., yesterday's self-efficacy is a confounder). In other cases, substantive knowledge may help rule out reverse causality, in particular if stable demographic variables are involved. However, when drawing on empirical studies to rule out reverse causality, one should keep in mind that such studies may suffer from their own causal inference problems.

One way to rule out both confounding and reverse causality is an experimental manipulation of the independent variable, with subsequent measurement of any dependent variable. For example, we could randomly assign athletes to receive high or low social support prior to an event, and then later measure their self-efficacy and their task performance. Figure 2, Panel B, provides a graphical interpretation of such an intervention: any path pointing into the independent variable is deleted, as randomly assigned social support is determined by chance (e.g., the flip of a coin) only.[8] Such an experimental manipulation allows a causal interpretation of the *total* effect of the manipulation on any outcome; for example, we could make causal claims about how our social support manipulation affects self-efficacy, or about how it affects task performance.

Unfortunately, being able to identify the *total* effect of one variable on another does not mean that we can automatically identify path-specific effects (see Avin et al., 2005 for technical details on such effects), such as indirect or direct effects. This leads us to problems of mediation analysis.

Mediation: Double Trouble

Claims about mediation, within the "causal chain" approach of PROCESS (and standard SEM), are claims about the *product* of two causal effects. For example, the indirect effect of social support on task performance via self-efficacy (social support -> self-efficacy -> task performance) would be the causal effect of social support on self-efficacy combined with the causal effect of self-efficacy on task performance. Thus, we must be able to identify two causal effects to identify an indirect effect. If either of the two estimates are confounded, the estimate of the indirect effect will be confounded as well. Additionally, if we are using PROCESS, we need to assume that both effects are linear and that the independent variable does not interact with the mediator (i.e., the effect of social support does not change depending on the level of self-efficacy).

[8] This presumes that researchers will use the randomly assigned social-support condition as an independent variable in subsequent analyses. We have encountered studies in which a randomized intervention took place, but the "independent" variable used in the data analysis was a subsequent measure of the construct of interest (e.g., a manipulation check). However, this subsequent measure is affected by *both* the randomized intervention and any relevant confounder, and so its effects cannot be identified easily—we now no longer have an experiment, but a surrogate experiment (Bareinboim & Pearl, 2012). One may conceptualize the subsequent measure as a mediator of the intervention; or alternatively consider the intervention an instrument for the subsequent measure. Rohrer and Lucas (2020) provide a discussion of these two conceptualizations.

These are quite strong assumptions. Even in standard experimental designs, problems arise when the mediator has not been randomized (e.g., Bullock, Green, & Ha, 2010).[9] For example, even if we were able to randomize social support, we still would not be able to estimate the causal effect of self-efficacy on task performance unless we assume that we are able to control for all common causes of these two variables (i.e., no unmeasured confounding; see Figure 3, Panel A, for an example of a confounder).

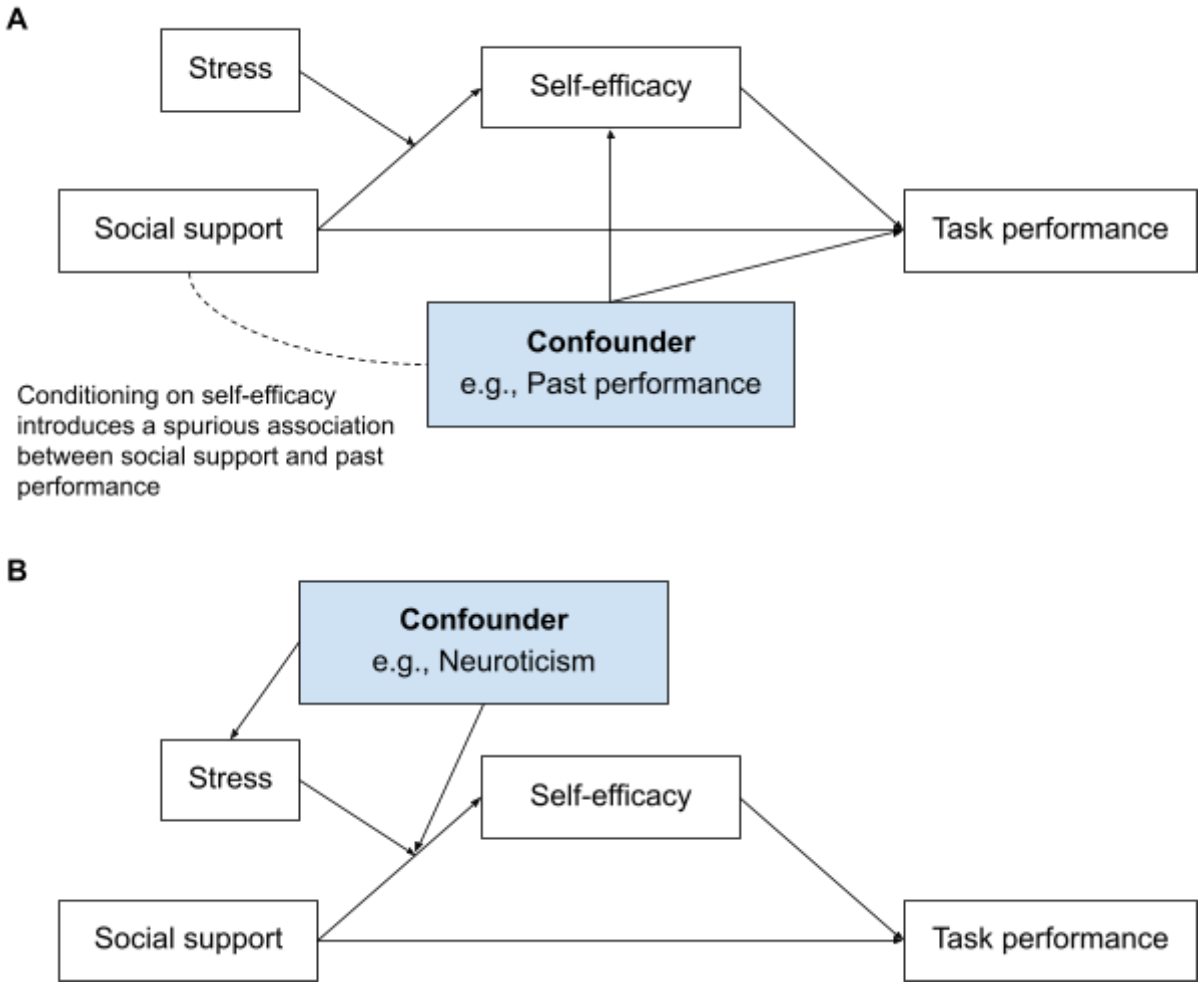

Figure 3. More modifications of the conditional process model. Panel A: A confounder between mediator and the dependent variable will bias the estimate of the indirect effect. Furthermore, statistical control for the mediator will induce a spurious association between the confounder and the independent variable, which will bias the estimate of the direct effect. Panel B: A moderator may be confounded and thus not the variable that actually causally interacts with the independent variable.

[9] And even if the mediator has also been randomized, this is still not sufficient for causal identification without additional assumptions (Imai et al., 2011, p. 770).

In standard mediation analysis, the total effect equals the sum of the indirect effect plus the direct effect. Couldn't that help us identify the indirect effect? After all, we noted above that with the help of randomization, we can identify the *total* effect of social support on task performance. If we additionally knew the *direct* effect, we could simply calculate the indirect effect as the difference between the two. The standard procedure for estimating the direct effect is statistical control for the mediator, which is meant to "shut off" the indirect path and thus only leave the direct path. Unfortunately, this does not work here. The mediator (self-efficacy) is causally affected by social support *and* other factors; it is a "collider" in which the effects of multiple variables come together. If we statistically control for a collider variable, we introduce spurious associations between its causes (Elwert & Winship, 2014; see also Rohrer, 2018 for more explanation geared towards psychologists). For example, here, conditioning on self-efficacy may introduce a spurious association between social support and previous task performance (Figure 3, Panel A).[10] Previous task performance affects current task performance, and so we have actually introduced additional confounding: Social support is now confounded with previous task performance which affects current task performance. Thus, we cannot give a causal interpretation to the coefficient of the direct effect of social support on task performance. We may fix this issue by statistically controlling for previous task performance and any other variable that affects both self-efficacy and task-performance, which leads us back to the strong assumptions that we have successfully measured and controlled for all common causes of the mediator and the dependent variable.

MacKinnon and Pirlott (2015) summarize some steps that researchers can take to increase the plausibility of mediation claims, such as different ways to adjust for confounders, and sensitivity analyses that probe to which extent estimates are robust to unobserved confounding (developed by Imai, Keele, & Tingley, 2010; VanderWeele, 2010). Unfortunately, these methods are not implemented in PROCESS. More generally, it may be worth highlighting that modern approaches to causal mediation analysis (such as e.g., Imai,

[10] To illustrate the case, let us assume that social support and past performance are unrelated in the overall population and that both have positive (additive, linear) effects on self-efficacy. If we now look at people with high self-efficacy, there will be a mix of different "types" of people. Some will have (1) high self-efficacy thanks to both solid social support and good past performance, others will have (2) high self-efficacy thanks to outstanding social support (despite mediocre performance), or (3) high self-efficacy thanks to outstanding performance (despite lacking social support). However, people with both low social support and bad past performance will be rare in this group, and their self-efficacy tends to be low. Thus, across the group of people with high self-efficacy, there may arise a spurious negative association between social support and past performance. The same logic applies to people low in self-efficacy, where we will once again observe the same (spurious) negative association. Hence, conditional on self-efficacy, past performance and social support are negatively correlated.

Keele, & Tingley, 2010) are far more complex than the simple causal chain method that is typically favored in psychology. Nonetheless, they may be necessary to accurately assess mediation under more realistic assumptions. When it comes to mediation, things are much more complicated than they seem.

Compounding Complexities with Multiple Mediators

So far, we have only considered a model with a single mediator. Of course, it is plausible (if not self-evident) that, in reality, any causal chain can be broken down into increasingly fine steps (i.e., serial mediators), and any remaining "direct" effect is transmitted via other intermediary variables (i.e., parallel mediators). The mere existence of multiple mediators is not a problem per se—*if* the crucial assumption of mediation analysis, sequential ignorability[11] is fulfilled, we can still identify a particular causal mechanism of interest (Imai et al., 2011). However, certain constellations with multiple mediators make it harder to achieve sequential ignorability. And if one wants to estimate multiple mediated paths at once, assumptions add up. The appendix of Imai et al. (2011) discusses a number of scenarios and highlights how different types of mechanisms result in different problems, some of which can only be addressed with the help of experimental designs.

Everything in Moderation: Causal Interaction vs. Effect Modification

Moderation refers to a situation in which the effect of one variable on another variable depends on the level of a third variable, the moderator. For example, social support may increase self-efficacy, but only among people who experience a lot of stress at home. From a causal inference perspective, such moderation can refer to two different phenomena.

An actual *causal* interaction would imply that a hypothetical intervention on the moderator would causally affect the magnitude of the effect of interest (see, e.g., VanderWeele, 2009). For example, if stress indeed causally interacts with social support, then an intervention on stress would change the effects of social support. Such a causal interaction is symmetrical: We may say that higher stress leads to a higher effect of social support on self-efficacy, but also that higher social support leads to a higher effect of stress

[11] The precise definition of this assumption is a bit more technical. First, given pre-treatment covariates, treatment assignment needs to be ignorable (like in any standard causal analysis to identify a total effect). And then second, again given pre-treatment covariates *and* observed treatment status, the mediator needs to be ignorable.

on self-efficacy. To correctly estimate such a causal interaction, we need to be able to properly identify the effect of the moderator. Randomization is once again the most direct way to do this, but in case this is not feasible, covariates may be included to rule out confounders. For example, as depicted in Figure 3, Panel B, it may actually be neuroticism that *causally* interacts with social support, not stress. To rule this out, we would need to statistically control for neuroticism. Here, it is important that we also include the interaction between any relevant covariate and the independent variable (Simonsohn, 2019; Yzerbyt et al., 2004)—note that this is unfortunately neither the default setting in PROCESS, nor routinely done in psychology.

Another type of moderation that is non-causal, and which VanderWeele (2009) refers to as "effect modification", means that the effect of a certain variable on another one *covaries* with a third variable,[12] regardless of whether or not an intervention on that third variable would actually result in a larger or smaller effect. Such effect modification can be asymmetrical; the third variable may covary with the effect of X on Y, but X need not covary with the effect of the third variable (which may also be precisely zero). For example, in a clinical setting, one may want to determine subgroups of patients for which a treatment works particularly well. Analyses may indicate that effects are particularly large among individuals with comorbid depression. Even if we do not know whether depression is indeed causally interacting with the treatment, or whether instead some confounding factor (e.g., socio-economic status) is at work, this information could still be helpful to guide treatment decisions.

If mere effect modification is the phenomenon of interest, presenting and interpreting findings should be done carefully. For example, a diagram such as Figure 1 should be avoided, because the arrow pointing away from the moderator begs to be interpreted as a causal interaction. More often than not, for psychologists, effect modification may be less interesting than causal interaction. For example, a clinical researcher asking *how* the treatment works will be much more interested in whether or not there is a *causal* interaction with depression, as it can potentially inform her about treatment mechanisms (e.g., the treatment may be particularly effective against cognitive patterns that are common in depressed patients).

[12] If covariates are included, effect modification occurs conditional on said covariates—and a change in the set of covariates can change whether or not effect modification occurs.

Investigating interactions and effect modification is quite challenging, not only because of causal concerns, but also because of additional statistical stumbling blocks. For example, moderation can crucially hinge on the scaling of the outcome variable, and the statistical model may not accurately speak to the research hypothesis of interest (see Rohrer & Arslan, 2020 for an introduction to these issues). Thus, when it comes to moderation, things are once again more complicated than they seem.

Finding the Right Model

As we have seen above, correctly estimating causal effects is challenging, and it always hinges on getting the model right (even in experiments). But how do we know that we have got the model right? Researchers may want to evaluate a particular model they have fitted to their data, or decide between multiple alternative models. The latter may rarely happen in practice (e.g., Chan et al., 2020) and often go wrong.

In mediation analysis, researchers sometimes aim to compare alternative mediation hypotheses by switching the direction of arrows (see Figure 4) and comparing the size and statistical significance of the estimated indirect effects. In particular, the model with the non-zero or larger indirect effect is thought to be supported by the data. Imagine the following scenario: Running Model A, we find a large indirect effect of X on Y via M. Running Model B, we find a smaller indirect effect of M via X on Y. A researcher may now conclude that Model A is correct, because the indirect effect is larger. This logic, however, is flawed. Why would we presuppose the existence of a (large) indirect effect if mediation analysis is supposed to tell us *whether* there is an indirect effect? And the estimate of the indirect effect can only be interpreted if we assume that we got the model right to begin with. A misspecified model may detect a large indirect effect that is entirely spurious.

But if the magnitude of the estimated indirect effects is not informative, maybe at least we can compare the fit of the models to figure out which one is preferable? Unfortunately, reversing arrows results in models that are equally supported by the data at hand; they belong to the same equivalence class. This means that they share the same implied covariance matrix (Thoemmes, 2015); they are *observationally equivalent*. On a substantive level, these models may look quite different. For example, in Figure 4, Panel A, self-efficacy mediates the effects of social support on task performance (and social support confounds the association between self-efficacy and task performance). In Figure 4, Panel B, social support mediates the effects of self-efficacy on task performance (and self-efficacy

confounds the association between social support and task performance). But no matter which model from the equivalence class we assume to be the actual data-generating process, we will always expect to observe the same empirical associations between the variables. This means that the empirical data alone cannot possibly distinguish between these models.

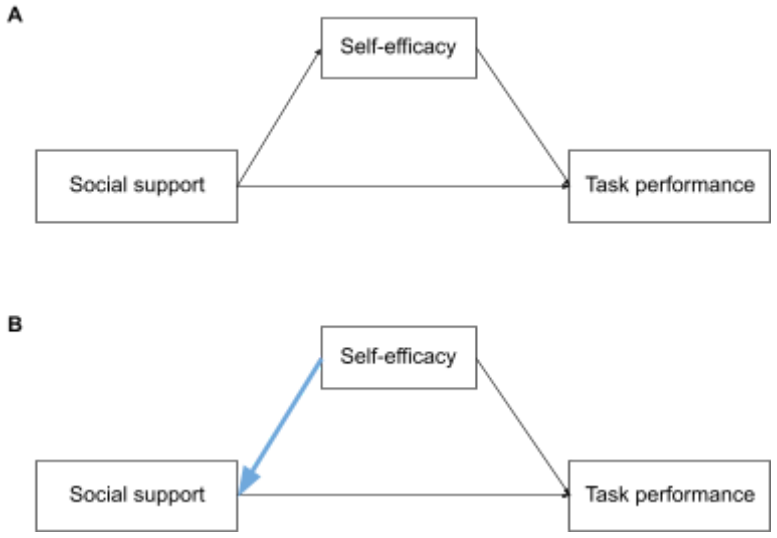

Figure 4. Panel A: Self-efficacy mediates the effect of social support on task performance. Panel B: The arrow from social support to self-efficacy has been reversed, so now social support mediates the effect of self-efficacy on task-performance. Both models belong to the same equivalence class and are thus statistically indistinguishable.

Equivalence classes are also the reason why the evaluation of model fit measures (such as the mean squared error, R^2, strictly speaking a measure of predictive performance, or, in an SEM context, χ^2, RMSEA, CFI and the like) alone can never tell us whether our model is correct: Each model from an equivalence class will produce identical fit indices. However, considerations of model fit may at least enable us to *discard* certain models as implausible. For this, we need to move analyses from PROCESS into an SEM context where we can properly assess and compare model fit. Here, global assessments of model fit may not be the most helpful because they cannot tell us *why* a model does not fit well (Pearl et al., 2016, p. 50), but a more local approach is possible.

If we assume that a certain causal model generated our data, we can derive testable implications. Testable implications are about the *independence* of pairs of variables—casually speaking, the fewer arrows between variables, the more things we can

test. Testable implications take the form of “controlling for C, A and B are statistically independent”—if this is not the case in our empirical data, we can reject the assumed model, and we also know *where* it went wrong (e.g., we missed a factor that causes an association between A and B). The directed acyclic graph framework provides clear rules for how to derive all testable implications of a given model (Elwert, 2013, pp. 252–254; Pearl et al., 2016, Chapter 2.5), and there is even software that automates the process (e.g., dagitty.net, see also Textor et al., 2011). In Box 1, we give a brief introduction on how to derive testable implications. Models from the same equivalence class share the same testable implications and thus firmly remain empirically indistinguishable. Furthermore, it should be noted that some PROCESS-style models do not have testable implications because they are saturated: the model has so many parameters that it can perfectly reproduce the empirically observed associations; in an SEM context, model fit would necessarily be perfect—the model thus cannot possibly fail and no testable implications remain.

Box 1: Spotlight on Testable Implications

To derive testable implications, we can break up a causal graph into three elementary causal structures (Elwert, 2013) that do (or do not) transmit associations between variables.

Chains: A → B → C. This chain transmits a causal association between A and C. If we control for the third variable (B, the mediator), the chain ceases to transmit an association. Considering this chain in isolation, this means that, conditional on B, A and C are independent, which we can write as: A ⊥ C | B.

Forks: A ← C → B. This fork transmits a non-causal association between A and B. If we control for the third variable (C, the confounder), the fork ceases to transmit an association: A ⊥ B | C

Inverted fork: A → C ← B. This inverted fork *does not* transmit an association between A and B, A ⊥ B. However, if we control for the third variable (C, the collider), then the inverted forks transmits a non-causal association between A and B.

We can break up all paths in Figure 3, Panel A, into these elementary structures to arrive at the testable implications of the model. Note that here, we will assume that the graph fully represents the assumed model, which is generally not the case for analyses conducted in PROCESS (more on that below).

Some variables should not be associated in the overall data:

Past performance ⊥ Social support
Past performance ⊥ Stress
Social support ⊥ Stress

Furthermore, Task performance and Stress should not be associated when we control for past performance, self-efficacy and social support:

Task performance ⊥ Stress | Past performance, Self-efficacy, Social support

However, PROCESS assumes that variables that jointly cause another variable are correlated (unless one of them causes the other). This means that, in most PROCESS graphs, there are a number of bidirectional arrows which are not depicted (Hayes, 2017, p. 22) and which reduce the number of testable implications.[13] If we include these arrows and once again deduce all testable implications, we are left with only one of them:

Task performance ⊥ Stress | Past performance, Self-efficacy, Social support

What if we find that task performance and stress are still correlated after controlling for past performance, self-efficacy, and social support? We may reject the underlying substantive model, or modify it. For example, it is possible that conditional independence is only violated because of measurement error in past performance, self-efficacy, or social support, which could be explicitly incorporated into the graph (Kuroki & Pearl, 2014).

Conclusion: Rethinking the Research Process

Running a PROCESS model may be a matter of a few clicks, but as we have seen above, interpreting the output requires a lot of strong assumptions. We need to rule out reverse causality and unobserved confounding (both of which may frequently be highly plausible in psychology) and additionally make assumptions about the functional forms of effects (about which we tend to know little), and whether or not those effects vary between individuals (with variation often being more plausible but leading to estimation complication). If assumptions are violated, the estimated coefficients end up being a mix of spurious and causal associations that can hardly be interpreted.

These issues have been highlighted before (e.g., Bullock et al., 2010; Chan et al., 2020; Fiedler et al., 2011; Thoemmes, 2015) and a large number of methods papers discuss them in great detail. Yet implementations of PROCESS-style models are often reported with little awareness, let alone critical reflection of the underlying assumptions. We may thus be confronted with normative methods that have been selected to further publication instead of discovery (Smaldino & McElreath, 2016), and such suboptimal methods can be quite persistent, in particular if there is little interdisciplinary exchange (Smaldino & O'Connor, 2020). This unfortunate situation can occur without any ill intention on the part of

[13] As a result, many PROCESS models are observationally equivalent. According to our analysis, of the 58 models included in the current version, 86% are equivalent to at least one other model. Overall, the 58 models belong to only 17 different equivalence classes.

researchers, and we do not mean to imply that researchers who use these models are bad at their job or (even worse) do not care about the truthfulness of their claims—they are simply implementing opaque practices which they have been taught, and which may often result in interesting-sounding empirical claims that are readily publishable given current norms.

We thus believe that to improve practices, some fundamental rethinking of what we consider a publishable scientific contribution may be necessary. Currently, researchers seem to feel compelled to do "everything" in a single paper—summarize and synthesize the existing literature, suggest a new theory or at least modify an existing one, hypothesize moderation and/or mediation and provide (preferably positive) empirical evidence through statistical analyses that they run themselves, maybe even across multiple studies they conducted themselves. It is perhaps unsurprising that they end up cutting corners when it comes to causal inference—a hard topic, about which they often receive little training—and rely on out-of-the-box statistical models.

Here is an alternative vision of what the research process *could* look like. An empirical investigation starts with conceptual considerations. Which causal effect is of interest in the first place, and why? Would it inform our theories, or does it have practical relevance? Can we come up with a well-defined counterfactual? What assumptions are we willing to make? These questions neatly tie in with recent calls for more rigorous theory (Muthukrishna & Henrich, 2019) and more formal modeling (Guest & Martin, 2020; Smaldino, 2017), but also with concerns about the utility of psychological research in times of crisis (Lewis, 2020). Such conceptual considerations may warrant their own publication, which allows others to build on them, but also reduces the pressure to immediately skip to data, out of some ill-conceived notion that only empirical studies count as science.

During this first stage, we may realize that we are not (yet) at the point in the research process at which we should try to estimate causal effects. For example, we may notice that open-ended exploration, description (Rozin, 2001) or prediction (Yarkoni & Westfall, 2017) are more suitable endeavors for the matter at hand; or that more basic questions regarding measurement need to be settled first. All of these types of investigations, if conducted rigorously, are relevant scientific contributions in their own right—researchers should not feel pressured to disguise them as hypothesis-testing confirmatory studies making some (explicit or implicit; Grosz et al., 2020) causal claim.

But if we finally get to the step of causal effect estimation, we should fully dedicate ourselves to the task. We need to clearly define the effects of interest, and venture to find a suitable identification strategy (see Foster, 2010a for an accessible introduction to the steps of causal inference). Which identification strategy works will strongly hinge on the assumptions that we are willing to make. Here, we might realize that a (field) experiment is the best way forward; or maybe we can find a suitable natural experiment (see Dunning, 2012 for a great introduction), such as a genetically-informative study (e.g., Briley et al., 2018); or maybe we will indeed settle for estimation based on structural models, such as PROCESS-style models. Of course, our decision will be partly constrained by concerns of feasibility (such as the funding available), and causal inference is not a monolithic endeavor—diverse perspectives and different strands of evidence produced by myriad methods can contribute (Krieger & Davey Smith, 2016). However, this is not a justification for selling a design as more convincing than it is, and for hiding assumptions—quite the opposite.

Any empirical study (including a randomized experiment) will make a multitude of assumptions, and the most crucial ones should be listed transparently. By this, we do not mean the type of boilerplate often tacked onto articles with PROCESS-style models (e.g., "future experimental studies should...", see also Chan et al., 2020). Instead, authors should list the *actual* specific assumptions under which their central estimate of the causal effects can be interpreted: "This estimate corresponds to the causal effect of *X* and *Y* under the assumption that, apart from *A*, *B*, and *C*, there are no common causes between the two of them"; or, for example, in a longitudinal study: "Results provide evidence for a causal effect of *X* on *Y* under the assumption that there are no time-varying confounding factors that affect both with different time lags" (Rohrer & Lucas, 2020). Such assumptions may often appear unrealistic, but they can be supplemented with statements about the degree to which conclusions are sensitive to violations of these assumptions. For example, alternative models with alternative sets of assumptions may be reported; and quantitative methods can be used to estimate to what extent conclusions are sensitive to unobserved confounding (e.g., Blackwell, 2014; Imai, Keele, & Yamamoto, 2010; Oster, 2019; VanderWeele, 2010).

Reviewers may feel tempted to judge manuscripts more harshly when assumptions are spelled out, rather than hidden away.[14] Thus, it is critical that *reviewers* are sufficiently

[14] This curse of transparency can also occur in other situations in which researchers aim for openness. For example, a preregistration may alert reviewers to discrepancies that would have gone unnoticed otherwise; open code may invite critical scrutiny where reviewers would have simply assumed that no errors occurred.

well-trained in causal inference to understand that a lack of explicit assumptions points to assumptions that the researchers are not aware of, or even to assumptions that were intentionally hidden away. There is no free lunch in causal inference.

With PROCESS-style models, the list of assumptions will be rather long—scrutinizing or even testing all of them will be too big a task for a single manuscript. Luckily, we need not tackle this task alone. If assumptions are taken seriously, this provides an opening for other researchers to join in. Transparent assumptions can be openly discussed in the community and examined in further studies, to corroborate or question the robustness of claims. Such criticism and probing of other people's work—be it conceptual or empirical—is once again a scientific contribution in its own right and should be valued accordingly.[15] Eventually, our collective understanding of the phenomenon may grow to a point at which we are comfortable making strong and specific assumptions. At this point, we may be able to do a conditional process analysis and have confidence in the resulting estimates.

It is possible that such a rigorous approach to causal inference might lead to the "disappearance" or at least shrinkage of effects that were previously deemed important; indeed, there is evidence that more rigorous designs lead to smaller causal effect estimates in the medical and social sciences (Branwen, 2014 maintains a list of studies on the topic). However, this pattern may partly be attributable to publication bias; and in principle, biases can also hide true causal effects or lead to their underestimation. Thus, it is an open question how less casual causal inference would affect our understanding of psychology.

Our vision is one in which psychological research is inherently transparent and collaborative, collectively striving towards greater robustness and culmination of knowledge. It is aligned with recent pushes towards greater transparency and rigor (e.g., Vazire, 2018), towards separating authorship from contributorship (Holcombe, 2019), and towards increased distributed collaboration (Moshontz et al., 2018). It may be an ambitious vision, but it is one in which any single research article can afford to be less ambitious in its scope. Instead of making sweeping complex causal claims, let's focus on getting one piece of the puzzle right at a time. Research is, after all, a process.

---

[15] Valuing such further probing would hopefully also help remedy psychology's "toothbrush problem," which was originally phrased for theories (Mischel, 2008), but has already been applied to models (Watkins, 1984), and could easily be expanded to, for example, experimental paradigms or latent constructs: no self-respecting person wants to use anyone else's.

References

Alvarez-Vargas, D., Braithwaite, D. W., Lortie-Forgues, H., Moore, M. M., Castro, M., Wan, S., Martin, E. A., & Bailey, D. H. (2020). *Hedges, Mottes, and Baileys: Causally Ambiguous Statistical Language can Increase Perceived Study Quality and Policy Relevance*. https://doi.org/10.31234/osf.io/nkf96

Avin, C., Shpitser, I., & Pearl, J. (2005). *Identifiability of Path-Specific Effects*. https://escholarship.org/uc/item/45x689gq

Bareinboim, E., & Pearl, J. (2012). Causal Inference by Surrogate Experiments: z-Identifiability. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1210.4842

Blackwell, M. (2014). A Selection Bias Approach to Sensitivity Analysis for Causal Effects. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*, *22*(2), 169–182.

Branwen, G. (2014). *How often does correlation=causality?* gwern.net. https://www.gwern.net/Correlation

Briley, D. A., Livengood, J., & Derringer, J. (2018). Behaviour genetic frameworks of causal reasoning for personality psychology. *European Journal of Personality*, *32*(3), 202–220.

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism?(don't expect an easy answer). *Journal of Personality and Social Psychology*, *98*(4), 550.

Chan, M., Hu, P., & K. F. Mak, M. (2020). Mediation Analysis and Warranted Inferences in Media and Communication Research: Examining Research Design in Communication Journals From 1996 to 2017. *Journalism & Mass Communication Quarterly*, 1077699020961519.

Chen, B., & Pearl, J. (2013). Regression and Causation: A Critical Examination of Six Econometrics Textbooks. In *Real-World Economics Review, Issue*.

https://papers.ssrn.com/abstract=2338705

Dunning, T. (2012). *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge University Press.

Elwert, F. (2013). Graphical Causal Models. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 245–273). Springer Netherlands.

Elwert, F., & Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, *40*, 31–53.

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests – An analysis of articles published in 2015. *Journal of Experimental Social Psychology*, *75*, 95–102.

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, *47*(6), 1231–1236.

Foster, E. M. (2010a). Causal inference and developmental psychology. *Developmental Psychology*, *46*(6), 1454–1480.

Foster, E. M. (2010b). The U-shaped relationship between complexity and usefulness: A commentary. *Developmental Psychology*, *46*(6), 1760–1766.

Götz, M., O'Boyle, E. H., Gonzalez-Mulé, E., Banks, G. C., & Bollmann, S. S. (2020). The "Goldilocks Zone": (Too) many confidence intervals in tests of mediation just exclude zero. *Psychological Bulletin*. https://doi.org/10.1037/bul0000315

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The Taboo Against Explicit Causal Inference in Nonexperimental Psychology. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *15*(5), 1243–1255.

Guest, O., & Martin, A. E. (2020). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691620970585

Hayes, A. F. (2017). *Introduction to Mediation, Moderation, and Conditional Process*

Analysis, Second Edition: A Regression-Based Approach. Guilford Publications.

Holcombe, A. O. (2019). Contributorship, Not Authorship: Use CRediT to Indicate Who Did What. *Publications*, *7*(3), 48.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies. *The American Political Science Review*, *105*(4), 765–789.

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 51–71.

Krieger, N., & Davey Smith, G. (2016). The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. *International Journal of Epidemiology*, *45*(6), 1787–1808.

Kuroki, M., & Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, *101*(2), 423–437.

Lewis, N. (2020, May 1). *How many (and whose) lives would you bet on your theory?* The Hardest Science. https://thehardestscience.com/2020/05/01/how-many-and-whose-lives-would-you-bet-on-your-theory/

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the Mediation, Confounding and Suppression Effect. *Prevention Science: The Official Journal of the Society for Prevention Research*, *1*(4), 173–181.

MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social*

Psychology, Inc, *19*(1), 30–43.

Manski, C. F. (2009). *Identification for Prediction and Decision*. Harvard University Press.

Mischel, W. (2008, December 1). *The Toothbrush Problem*. psychologicalscience.org. https://www.psychologicalscience.org/observer/the-toothbrush-problem

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and Causal Inference*. Cambridge University Press.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., … Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing Psychology through a Distributed Collaborative Network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229.

Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics: A Publication of the American Statistical Association*, *37*(2), 187–204.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Rees, T., & Freeman, P. (2009). Social Support Moderates the Relationship Between Stressors and Task Performance Through Self-Efficacy. *Journal of Social and Clinical Psychology*, *28*(2), 244–263.

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.

Rohrer, J. M., & Arslan, R. C. (2020). Precise answers to vague questions: Issues with

interactions. In *PsyArXiv*. https://doi.org/10.31234/osf.io/7fm2j

Rohrer, J. M., & Lucas, R. E. (2020). *Causal Effects of Well-Being on Health: It's Complicated*. https://doi.org/10.31234/osf.io/wgbe4

Rozin, P. (2001). Social Psychology and Science: Some Lessons From Solomon Asch. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *5*(1), 2–14.

Simonsohn, U. (2019). *[80] Interaction Effects Need Interaction Controls*. http://datacolada.org/80

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

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384.

Smaldino, P. E., & O'Connor, C. (2020). *Interdisciplinarity Can Aid the Spread of Better Methods Between Scientific Communities*. https://osf.io/cm5v3/download

Textor, J., Hardt, J., & Knüppel, S. (2011). DAGitty: a graphical tool for analyzing causal diagrams. *Epidemiology* , *22*(5), 745.

Thoemmes, F. (2015). Reversing Arrows in Mediation Models Does Not Distinguish Plausible Models. *Basic and Applied Social Psychology*, *37*(4), 226–234.

VanderWeele, T. J. (2009). On the distinction between interaction and effect modification. *Epidemiology* , *20*(6), 863–871.

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology* , *21*(4), 540–551.

Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *13*(4), 411–417.

Watkins, M. J. (1984). Models as toothbrushes. *The Behavioral and Brain Sciences*, *7*(1),

86–86.

Westfall, J., & Yarkoni, T. (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. *PloS One*, *11*(3), e0152719.

Wood, R. E., Goodman, J. S., Beckmann, N., & Cook, A. (2008). Mediation Testing in Management Research: A Review and Proposals. *Organizational Research Methods*, *11*(2), 270–295.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *12*(6), 1100–1122.

Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology*, *40*(3), 424–431.