Draftable Comparison Export

This document is an exported comparison with limited functionality, generated by Draftable Desktop. To access full functionality, use Draftable's powerful comparison viewer in any of our products.

Left document:	PCIRR Stage 1 Manuscript.pdf
Right document:	PCIRR Stage 1 Manuscript Revision.pdf

What is this document?

This is a comparison of two documents. The two documents are interleaved such that the left document is displayed on even pages and the right document is displayed on odd pages.

Is there a specific way I should view this file?

This document is intended to be viewed in Two Page Continuous mode (or sometimes called 'Two Page Scrolling'). It should open in this mode by default when using Adobe Acrobat and most popular PDF readers.

If the document opens in a different view, you can often change this in the settings. In Adobe Acrobat, go to **View > Page Display > Two Page Scrolling**.

Why are there blank pages?

Blank pages are inserted to keep both documents as aligned as much as possible.

How do I read the changes?

Text deleted from the left document and, hence, not in right document is highlighted red. Text added to the right document and, hence, not in left document is highlighted green.

Tip for printing

When printing this document, we recommend printing double-sided and include this first page. This will result in the matching text being displayed on different pages and easily readable, much like a book.

For more information

Draftable offers powerful document comparison solutions for all use-cases. To view our products, please visit our website: draftable.com.

How Does Model Specification Impact Statistical Power and Type I Error Rate in Moderated Mediation Analysis? A Registered Report

Jessica L. Fossum[1,2], Amanda K. Montoya[2], and Samantha F. Anderson[3]

[1]Department of Psychology, Seattle Pacific University

[2]Department of Psychology, University of California, Los Angeles

[3]Department of Psychology, Arizona State University

Author Note

Jessica L. Fossum https://orcid.org/0000-0002-4907-0160

Amanda K. Montoya https://orcid.org/0000-0001-9316-8184

Samantha F. Anderson https://orcid.org/0000-0003-4451-5295

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Jessica L. Fossum, School of Psychology, Family, and Community, Seattle Pacific University, 3307 Third Ave. W., Suite 107, Seattle, WA 98119-1922.

Email: fossumj@spu.edu

How Does Model (Mis)Specification Impact Statistical Power, Type I Error Rate, and Parameter Bias in Moderated Mediation? A Registered Report

Jessica L. Fossum[1], Amanda K. Montoya[2], and Samantha F. Anderson[3]

[1]Seattle Pacific University

[2]University of California - Los Angeles

[3]Arizona State University

Author Note

Jessica L. Fossum https://orcid.org/0000-0002-4907-0160

Amanda K. Montoya https://orcid.org/0000-0001-9316-8184

Samantha F. Anderson https://orcid.org/0000-0003-4451-5295

Correspondence concerning this article should be addressed to Jessica L. Fossum, Department of Psychology, Seattle Pacific University, 3307 3rd Ave W. Suite 107, Seattle, WA 98119. E-mail: fossumj@spu.edu.

How does model specification impact statistical power and type I error rate in moderated mediation analysis? A registered report

Psychological researchers are often interested in explaining how and when effects occur, such as how discrimination affects internalizing symptoms through peer victimization, and when those relationships differ based on age (Ramos et al., 2021). Mediation analysis provides a way of examining effects whether a proposed mediator variable (e.g., peer victimization) serves as a mechanism by which one variable effects another (e.g., discrimination affects internalizing). Moderation analysis provides a way of examining when or for whom effects occur. For example, age could moderate the effect of discrimination on peer victimization, meaning that this effect may be stronger among younger or older children. These procedures can be used together in a moderated mediation analysis, exploring when or for whom specific processes occur. In moderated mediation, moderators are allowed to moderate any of the paths in a mediation model. Moderated mediation analyses have become more common, with WebofScience counting 2,602 published articles using the analysis in 2020, 3,499 in 2021, and 3,815 in 2022.

As moderated mediation becomes increasingly popular, it is important for researchers to consider the many choices involved in the specification of these models and the implications of those choices on statistical properties, such as type I error and power. In this paper we focus on model specification (where moderation is allowed to occur in the mediation model) and it's implications for sample size planning and power. Any number of paths in a mediation model can be moderated (Preacher, Rucker, & Hayes, 2007), so a choice must be made about which paths in the mediation are moderated. Each additional moderated path introduces an additional interaction into the model, which can be very difficult to detect with limited sample sizes. While prior research suggests that specification of the order of variables in a mediation model should come from theory (Fiedler, Schott, & Meiser, 2011; Fiedler, Harris, & Schott, 2018), there have been limited explorations of how to specify moderation in these models (Rohrer, Hünermund, Arslan, &

Abstract

Moderated mediation models are commonly used in psychological research and other academic fields to model when and how effects occur. Researchers must choose which paths in the mediation model are moderated when specifying this type of model. While the ultimate goal is to specify the model correctly, researchers may struggle to determine whether to err on the side of including too many moderated paths (maximalist approach) or including too few moderated paths (minimalist approach). This registered report examines how the specification of moderation impacts statistical power, type I error rate, and parameter bias for the index of moderated mediation. In a systematic review, we found that six model specifications account for 85% of published moderated mediation analyses and the median sample size was 285. We ran a Monte Carlo simulation study to examine the impacts of model specification on power and type I error rate, and results were analyzed using multilevel logistic regression. In reference to the data-generating process, the data analysis model can either be correctly specified, over-specified, under-specified, or completely misspecified. Over-specified models were hypothesized to have lower statistical power to detect a significant index of moderated mediation compared to correctly specified models, and relatively low parameter bias. Under-specified models were hypothesized to have lower statistical power than correctly specified models, but unacceptably high parameter bias. Completely misspecified models were hypothesized to have inflated type I error rates and unacceptable parameter bias. Implications of results on study planning (specification and sample size) for moderated mediation will be discussed.

Keywords: moderated mediation, statistical power, type I error rate, model misspecification

Elson, 2022), or the effect on statistical power and type I error rate.

Statistical power is the likelihood of detecting an effect if that effect truly exists in the population (Neyman & Pearson, 1993). Because statistical power depends on sample size, the goal in sample size planning is to find the optimal balance between maximizing power and minimizing wasted resources (Maxwell & Kelley, 2011). Planning for adequate statistical power is an important part of study design to be able to detect effects (Aberson, 2019a). Low power has been cited as a common source of problems in scientific literature (Ioannidis, 2005), particularly with respect to the replicability crisis (Anderson & Maxwell, 2017; Earp & Trafimow, 2015).

Prior research suggests that there is reason to believe that mediation and moderation analysis may suffer from low power, typically due to insufficient sample size. For example, Fritz and MacKinnon (2007) found that the median sample size used for mediation analysis was 187, but the minimum sample size required to detect a small effect common in psychology was 558. Götz, O'Boyle, Gonzalez-Mulé, Banks, and Bollmann (2021) conducted a large scale review of mediation analyses in psychology journals and found evidence that too many mediation analysis results were just barely significant, suggesting either foul play (e.g., p-hacking), issues with low power, or both. Charlton, Montoya, Price, and Hilgard (2021) found a similar result in marketing. Similarly, moderation analyses are believed to typically be underpowered, to the point where researchers have considered raising the α level for the benefit of higher power to detect a significant interaction (Marshall, 2007). Moderated effects can be difficult to detect because they are usually very small. For example, a 30-year review by Aguinis, Beaty, Boik, and Pierce (2005) found the average effect size to be $f^2 = .002$ while only 72% of the reviewed analyses had power of .8 to detect even a small effect.

To our knowledge, no prior studies have examined whether current moderated mediation analyses are well powered; however, given the research on mediation and moderation in concert, this suggests there is reason to be concerned about power in

How Does Model (Mis)Specification Impact Statistical Power, Type I Error Rate, and Parameter Bias in Moderated Mediation? A Registered Report

Psychological researchers are often interested in explaining how and when effects occur. For example, Cognitive Remediation Therapy (CRT) has been demonstrated to improve cognitive function, including planning, among individuals with schizophrenia (Wykes et al., 2012), and higher levels of cognitive functioning have been shown to improve effectiveness at work (Wykes et al., 2007). This suggests that cognitive function may be a mechanism by which CRT improves work effectiveness (Wykes & Spaulding, 2011). Mediation analysis quantifies the degree to which a proposed mediator variable (e.g., cognitive function) acts as an intermediary through which one variable (e.g. CRT) affects another (e.g., work effectiveness). Moderation analysis provides a way of examining when or for whom effects occur. For example, improvements in planning are expected to improve work effectiveness, but only for individuals with good memory (Wykes et al., 2012). These procedures can be used together in a moderated mediation analysis, exploring when or for whom specific processes occur. In these models, any of the paths in a mediation can be moderated (Preacher et al., 2007).

Researchers must choose which paths in the mediation are moderated, a process called model specification. Each additional moderated path introduces an additional interaction into the model, which can impact statistical power. Prior research emphasizes the importance of theory in specification of the order of variables in a mediation model (Fiedler et al., 2011, 2018). Still, there have been limited explorations of how to specify moderation in these models (Rohrer et al., 2022) or the effect of model (mis)specification on statistical power, type I error rate, and parameter bias.

Low power has been cited as a common source of problems in the scientific literature (Ioannidis, 2005), particularly concerning the replicability crisis (Anderson & Maxwell, 2017; Earp & Trafimow, 2015). Prior research suggests a combination of small effect sizes and insufficient sample sizes leads to low power for mediation and moderation moderated mediation analyses. Coupled with the methods increasing popularity, this topic is very important to examine. In addition, moderated mediation analyses present a unique problem with respect to model specification: there are many paths which could be moderated. This choice of which paths should be moderated may play an important role in determining power. Prior research in moderation analysis suggests that detecting more interactions and higher order interactions requires larger sample sizes (McClelland & Judd, 1993). However this issue has not been explored in the context of moderated mediation models.

In addition, misspecifications of these models may also increase type I errors (detecting effects when they do not exist). There are two potential philosophies of model specification: maximalism and minimalism. A maximalist perspective would suggest that all paths in the model should be moderated, as this would avoid missing any effects that might exist. A minimalist approach would suggest that the fewest possible paths should be moderated in order to maximize statistical power. In this paper we explore the impact of these philosophies on power and types I error.

In this registered report, we investigate how model specification impacts power and type I error rate in moderated mediation models. Our goal is to provide recommendations for researchers with respect to how they specify their models and plan their sample sizes. We conducted a systematic review to see what sample sizes were common in moderated mediation models and what model specifications tended to be most popular (maximalist vs. minimalist approach). Next we conducted a simulation study which examines how different model specification decisions impact type I error and power. Based on the findings of this simulation we will summarize the results, and provide a discussion of the implications for researchers with respect to model specification and sample size planning. We will also provide sample size recommendations for researchers using certain moderated mediation models included in this study. Finally, we will conclude with a discussion of these findings in hopes of aiding researchers in the decisions required for a moderated mediation analysis.

analysis. For example, Fritz and MacKinnon (2007) found that the minimum sample size required to detect a mediated effect when both paths involved in the indirect effect are small to medium (an effect size common in psychology) was 558, but also that the median sample size used for mediation analysis was only 187. Götz et al. (2021) and Charlton et al. (2021) conducted large-scale reviews of mediation analyses in psychology and marketing journals, respectively, and found evidence that too many mediation analysis results were just barely significant, suggesting either *p*-hacking, low power, or both. Moderation analyses also tend to be underpowered (Marshall, 2007). A 30-year review by Aguinis et al. (2005) found the average effect size to be very small ($f^2 = .002$) while only 72% of the reviewed analyses had power of .8 to detect an effect of $f^2 = .02$ (an order of magnitude larger). Prior research in moderation analysis suggests that detecting more and higher-order interactions requires larger sample sizes (McClelland & Judd, 1993). However, this issue has not been explored in moderated mediation models.

In this paper, we contrast two potential philosophies of model specification: maximalism and minimalism. A maximalist perspective would suggest that all paths in the model should be moderated, as this would avoid missing any effects that might exist. While the maximalist approach has not been discussed in the context of moderated mediation previously, it has been applied in the context of factor analysis (Barr et al., 2013) and multilevel modeling (Brysbaert, 2007; Matuschek et al., 2017). However, maximalist approaches may result in low statistical power (Matuschek et al., 2017). Maximalist approaches should also result in low parameter bias because including extraneous predictors should not result in bias (Robins et al., 1994). By contrast, a minimalist approach would suggest that the fewest possible paths should be moderated to maximize statistical power. If however, truly moderated paths are omitted, this could result in parameter bias and type I errors. Rimpler et al. (2024) found that omitting an interaction effect in linear regression drastically biased simple effects. Ultimately, the goal of model specification is to correctly specify the model. However, it is not always possible to know whether a model is correctly

The remainder of this introduction is organized as follows: We begin with an introduction to moderated mediation analysis. We describe estimation and inference for moderated mediation models, specifically the index of moderated mediation (IMM). Next, we summarize the current literature on sample size planning for moderated mediation analysis. Finally, we discuss our systematic review exploring current practices in moderated mediation analysis and propose our simulation study examining the impact of model specification on power and type I error rate.

Introduction to Moderated Mediation

Mediation occurs when a predictor variable X affects an outcome Y through a mediator variable M. The effect of X on Y when controlling for M is called the direct effect, and the product of the effect of X on M and the effect of M on Y controlling for X is the indirect effect. The indirect effect is the effect of interest in mediation analysis. Moderation can occur on any of these paths, where the effect of one variable on another depends on the value of a moderator variable, W. When paths in a mediation which make up the indirect effect are moderated, it is a moderated mediation model (Edwards & Lambert, 2007).

Many mediation, moderation, and moderated mediation models can be estimated with commonly used macro, PROCESS (Hayes, 2022). PROCESS allows any paths to be moderated, and in any combination, which provides researchers flexibility but also allows for possible misspecification in where the moderation occurs. This study focuses on simple mediation models (a single mediator) with one or more paths moderated. Moderated mediation with one mediator variable and one outcome variable requires two equations to describe the model: one equation for M and another equation for Y. There are two possible equations for M, depending on whether moderation occurs on the X to M path. Without moderation, the equation used for M is Equation 1

$$M_i = a_0 + aX_i + \varepsilon_{M_i} \tag{1}$$

specified, and researchers may need to consider whether to lean toward a maximalist or minimalist approach. In this registered report, we aim to provide guidance to researchers in this regard, demonstrating the impact of these two philosophies on power, type I error, and parameter bias in the context of moderated mediation.

It is important to consider model specification during the study planning phase, as the model specification will also impact sample size planning. One goal of this registered report is to identify if more complex models will require larger sample sizes to achieve similar levels of power, and so the relative costs of selecting a more general model could be corrected by planning to collect a larger sample size. This study provides guidance for understanding how much sample sizes should vary depending on model specification.

The remainder of this introduction is organized as follows: We begin with an introduction to moderated mediation analysis, including estimation and inference for the index of moderated mediation. Next, we summarize the current literature on sample size planning for mediation, moderation, and moderated mediation analysis. Finally, we outline our simulation study examining the impact of model specification on power, type I error rate, and parameter bias.

Introduction to Moderated Mediation

Mediation occurs when a predictor variable X affects an outcome Y through a mediator variable M. The effect of X on Y when controlling for M is called the direct effect, and the product of the effect of X on M and the effect of M on Y controlling for X is the indirect effect, which is the effect of interest in mediation analysis. Moderation can occur on any of these three paths, where the effect of one variable on another depends on the value of a moderator variable, W. When paths that make up the indirect effect are moderated, it is a moderated mediation model (Edwards & Lambert, 2007).

This study focuses on simple mediation models (a single mediator) with one or more paths moderated by a single moderator. These models are estimated using two linear regression equations: one for M and one for Y. There are two possible equations for M, where a_0 is the intercept, and a is the effect of X on M, and ε_{M_i} is the residual. With moderation on the path between X and M, called first-stage moderated mediation (Hayes, 2015), Equation 2 is used which includes an interaction between X and W. In Equation 2, a_0 remains the intercept and a_1 is the effect of X on M when $W = 0$. The coefficient a_2 is the effect of W on M when $X = 0$, and a_3 is the degree to which the effect of X on M depends on W.

$$M_i = a_0 + a_1 X_i + a_2 W_i + a_3 X_i W_i + \varepsilon_{M_i} \tag{2}$$

The second equation needed for moderated mediation is for Y. While some of the notation in the following equations is the same, the values will not necessarily be the equal. There are four options for this equation: First, it could include no interactions at all (Equation 3).

$$Y_i = c_0' + c' X_i + b M_i + \varepsilon_{Y_i} \tag{3}$$

Second, it could include an interaction between M and W (Equation 4). When moderation occurs only on the path between M and Y, it is called second-stage moderation (Hayes, 2015).

$$Y_i = c_0' + c_1' X_i + b_1 M_i + b_2 W_i + b_3 M_i W_i + \varepsilon_{Y_i} \tag{4}$$

Third, when moderation occurs on just the X to Y path but not the M to Y path, the equation for Y includes only one interaction as well, but it is the interaction between X and W (Equation 5).

$$Y_i = c_0' + c_1' X_i + c_2' W_i + c_3' X_i W_i + b M_i + \varepsilon_{Y_i} \tag{5}$$

depending on whether (Eq 1) or not (Eq 2) moderation occurs on the X to M path:

$$M_i = a_0 + a_1 X_i + a_2 W_i + a_3 X_i W_i + e_{M_i} \tag{1}$$

$$M_i = a_0 + a_1 X_i + e_{M_i} \tag{2}$$

The equation for Y can have the X to Y path moderated (Eq. 3), the M to Y path moderated (Eq. 4), both moderated (Eq. 5), or neither moderated (Eq. 6).

$$Y_i = c'_0 + c'_1 X_i + c'_2 W_i + c'_3 X_i W_i + b M_i + e_{Y_i} \tag{3}$$

$$Y_i = c'_0 + c'_1 X_i + c'_2 W_i + b_1 M_i + b_2 M_i W_i + e_{Y_i} \tag{4}$$

$$Y_i = c'_0 + c'_1 X_i + c'_2 W_i + c'_3 X_i W_i + b_1 M_i + b_2 M_i W_i + e_{Y_i} \tag{5}$$

$$Y_i = c'_0 + c'_1 X_i + b_1 M_i + e_{Y_i} \tag{6}$$

Pairing together the equations for M and the equations for Y results in eight possible models. This study focuses on six of these, as displayed in Figure 1. Two combinations are not used in this study: the model where no paths are moderated (Eq. 2 & 6) and the model where only the direct effect is moderated (Eq. 2 & 3) thus not a moderated mediation. Figure 1 displays each model using a conceptual diagram. In this study, we use the model numbering system from the PROCESS macro (Hayes, 2022). We conducted a systematic review of 411 articles to understand which models are most commonly used in practice, and six models emerged (Models 7, 8, 14, 15, 58, and 59; see Appendix A for more details on the systematic review). The equation numbers for both M and Y specifying each of the six moderated mediation models used in this simulation study are displayed in Figure 1.

When the indirect effect is moderated, the conditional indirect effect quantifies the indirect effect at a specific value of the moderator. Mathematically, the effect of X on M is

Fourth, if moderation occurs on both the X to Y path and the M to Y path, the equation for Y includes both interactions, and is represented by Equation 6.

$$Y_i = c'_0 + c'_1 X_i + c'_2 W_i + c'_3 X_i W_i + bM_i + b_3 M_i W_i + \varepsilon_{Y_i} \tag{6}$$

Combining the equations for M (1 and 2) and the equations for Y (3 - 6) in different ways creates eight models. This study focuses on six of these, as displayed in the following figures. Two combinations are not used in this study: No model uses Equation 1 for M and Equation 3 for Y because this results in a simple mediation model without any moderation, and no model uses Equation 1 for M and Equation 5 for Y because then only the direct effect would be moderated, and thus not a moderated mediation.

Figure 1 displays each model using a conceptual diagram (left) and a statistical diagram (right). In this paper we use the model numbering system from the PROCESS macro, as this is the most commonly used tool for moderated mediation analysis and these models are often referenced using these numbers in empirical research (Hayes, 2022). For example, Ramos et al. (2021) used Model 8: The mediation of interest was discrimination affecting internalizing symptoms through peer victimization, with age as the moderator. Because theory suggested that only the path between discrimination and peer victimization, and the path between discrimination and internalizing symptoms should be moderated, the researchers chose this model.

Index of Moderated Mediation

When moderation is present on either the X to M or M to Y path, the indirect effect varies across values of W. For example, the indirect effect of discrimination on internalizing symptoms through peer victimization differs depending on the age of the person. When the indirect effect is moderated, there is no longer a single indirect effect, rather conditional indirect effects, conditional on the value of the moderator. The

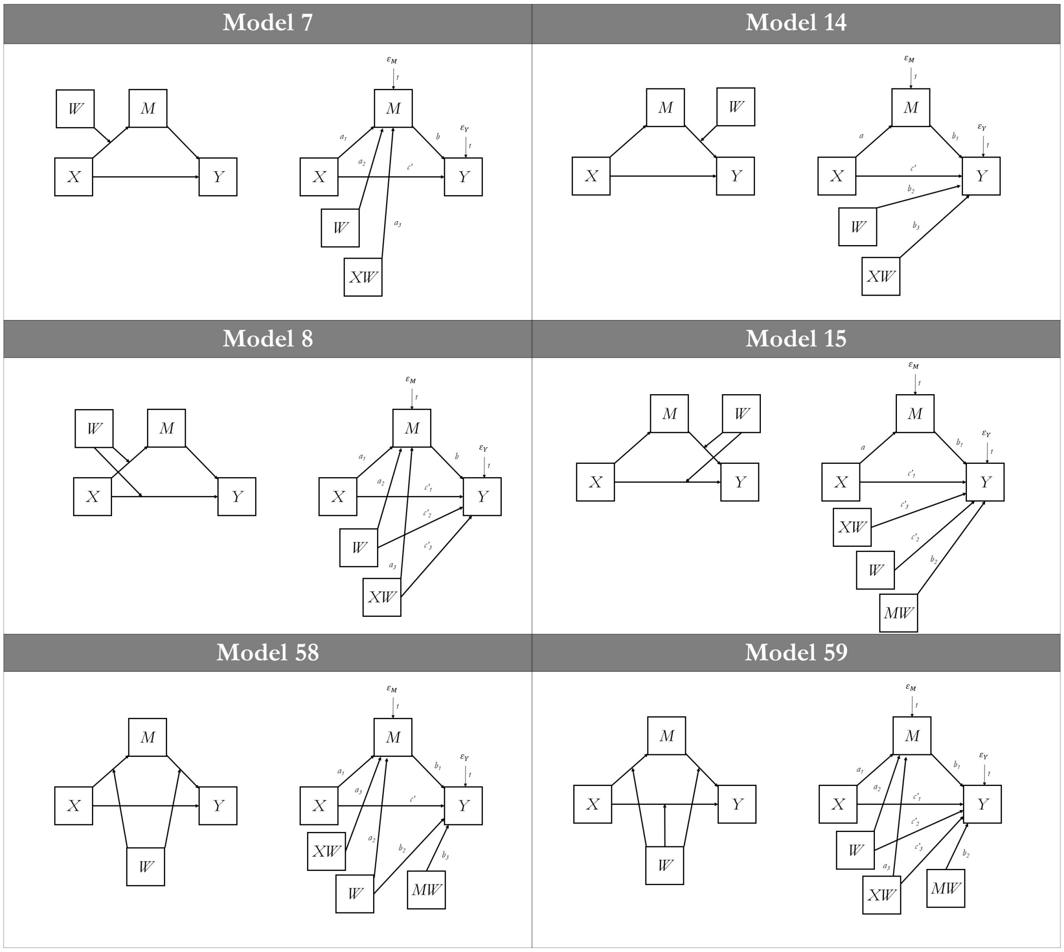

Figure 1. Moderated Mediation conceptual diagrams (left) and statistical diagrams (right) within each panel. Model 7 uses Equation 2 for M and Equation 3 for Y. Model 8 uses Equation 2 for M and Equation 5 for Y. Model 14 uses Equation 1 for M and Equation 4 for Y. Model 15 uses Equation 1 for M and Equation 6 for Y. Model 58 uses Equation 2 for M and Equation 4 for Y. Model 59 uses Equation 2 for M and Equation 6 for Y.

conditional indirect effect quantifies the indirect effect for individual with a specific value of the moderator.

The index of moderated mediation quantifies the degree to which the conditional indirect effect differs for individuals that differ by one unit on the moderator. A test on this index of moderated mediation can be used to evaluate the question “Is the mediation moderated?” (Hayes, 2015). If this index is zero at the population level, this means that the indirect effect is constant across values of the moderator, but if it is non-zero, the indirect effect depends on the value of the moderator (i.e., the mediation is moderated).

multiplied by the effect of M on Y to calculate the conditional indirect effect. For example if the effect of X on M is moderated by W, it is defined by $a_1 + a_3W$, otherwise it is a_1. If the effect of M on Y is moderated by W, it is defined by $b_1 + b_3W$, otherwise it is b_1. So, for example, if only the path from X to M is moderated the conditional indirect effect would be $(a_1 + a_3W)b_1$. Similar calculations can be used for any combination of moderated paths.

The index of moderated mediation quantifies the degree to which the conditional indirect effect depends on the value of the moderator. A test on this index can be used to evaluate the question “Is the mediation moderated?" (Hayes, 2015). If this index is zero at the population level, this means that the indirect effect is constant across the values of the moderator, but if it is non-zero, the indirect effect depends on the value of the moderator (i.e., the mediation is moderated).

The index of moderated mediation is only defined in cases where the conditional indirect effect is a linear function of the moderator with one exception (Hayes, 2015): When the moderator is dichotomous, the index is defined for any model because the index can be calculated as the difference between the two conditional indirect effects (evaluated at each value of the moderator) (Fairchild & MacKinnon, 2009). Figure 1 gives the index of moderated mediation for the six models described in this section. Inference can be conducted on the index of moderated mediation using a percentile bootstrap confidence interval (CI), which is a recommended method because it balances type I error and power (Coutts, 2023; Yzerbyt et al., 2018).

Sample Size Planning for Moderated Mediation

There are many factors that have been shown to affect statistical power in mediation and moderated regression separately (Aguinis, 1995; O’Rourke & MacKinnon, 2014), including effect size and sample size (Cohen, 1988), and correctly specifying the model (Dupont & Plummer, 1998; Rimpler et al., 2024). Previous research in both mediation analysis (Fairchild & McDaniel, 2017; Fritz & MacKinnon, 2007; Götz et al., 2021) and moderation analysis (Aguinis et al., 2005; Marshall, 2007) suggest that these

The equations for the conditional indirect effects and the index of moderated mediation are unique to the model being estimates (the combination of the equation for M and the equation for Y). The index of moderated mediation is only defined in cases where the conditional indirect effect is a linear function of the moderator, with one exception (Hayes, 2015). When the moderator is dichotomous, the index is defined for any model, because the index can be calculated as the difference between the two conditional indirect effects (evaluated at each value of the moderator) (Fairchild & MacKinnon, 2009). Table 1 gives equations for the six moderated mediation models described in this section, along with the index of moderated mediation.

Inference is conducted on the index of moderated mediation using a percentile bootstrap confidence interval (CI). Bootstrapping is the recommended method for conducting inference because it is commonly used in mediation analysis already, and has been shown to perform better in handling the product of two normally distributed variables than other methods in simulation studies while not inflating the type I error rate (Coutts, 2023; Yzerbyt, Muller, Batailler, & Judd, 2018).

Table 1
Defining the Index of Moderated Mediation

Model	Equation for M	Equation for Y	Index of Moderated Mediation
7	2	3	a_3b
8	2	5	a_3b
14	1	4	ab_3
15	1	6	ab_3
58	2	4	$(a_1 + a_3)(b_1 + b_3) - a_1b_1$
59	2	6	$(a_1 + a_3)(b_1 + b_3) - a_1b_1$

Note. Index of Moderated Mediation for Models 58 and 59 is only defined when the moderator is dichotomous. Equations for the index of moderated mediation assume dichotomous moderated is with values 0 and 1.

Sample Size Planning for Moderated Mediation

There are many factors that have been shown in simulation studies and analytically to affect statistical power in mediation and moderated regression separately (O'Rourke & MacKinnon, 2014; Aguinis, 1995), including effect size and sample size (Cohen, 1988), dichotomous vs. continuous predictor variables (McClelland & Judd, 1993), and correctly specifying the model (Dupont & Plummer, 1998). Previous research in both mediation analysis (Fairchild & McDaniel, 2017; Fritz & MacKinnon, 2007; Götz et al., 2021) and moderation analysis (Marshall, 2007; Aguinis et al., 2005) suggest that these analyses tend to be underpowered in psychology research .

Statistical power to detect the index of moderated mediation is difficult to approximate given the complexity of these models (Bakker, Hartgerink, Wicherts, & van der Maas, 2016). While there are a variety of packages and tools available to sample size planning in mediation and moderation separately (Zhang & Yuan, 2018; Schoemann, Boulton, & Short, 2017; Zhang & Wang, 2013; Kenny, 2017), there is only one tool we know of which conducts power analysis for any moderated mediation models. Power analysis for moderated mediation Models 7 and 14 are available in the R package pwr2ppl (Aberson, 2019b). In this package, the researcher must specify many parameters to calculate power for the index of moderated mediation. For example, to determine power to detect the index analyses tend to be underpowered in psychology research. Our systematic review (see Appendix A) found that the median sample size used for moderated mediation was 285. This median sample size is larger than those found in mediation (Fritz & MacKinnon, 2007), but still not large enough to detect even medium effects in mediation only. More complex models (e.g., moderated mediation) likely require larger sample sizes than less complex models. However, if researchers do not select their model a priori and plan their sample size accordingly, we may see similar sample sizes used across different models. In the systematic review, the most complex model we examined (Model 59) had the highest median sample size of 363 (but not the highest of all models). Otherwise, there were no other clear connections between sample size and model complexity. This suggests that researchers may not be accounting for model complexity in their sample size planning.

Statistical power to detect the index of moderated mediation is difficult to approximate (Bakker et al., 2016). While there are a variety of packages and tools available to do sample size planning in mediation and moderation separately (Kenny, 2017; Schoemann et al., 2017; Zhang & Wang, 2013; Zhang & Yuan, 2018), there is only one tool we know of that conducts power analysis for the index of moderated mediation. Power analysis for the index of moderated mediation for Models 7 and 14 is available in the R package pwr2ppl (Aberson, 2019). Currently, for models other than 7 and 14, there are no tools available to conduct power analysis for the index of moderated mediation. WebPower calculates power or the conditional indirect effect and for the moderation on a specific path (Zhang & Yuan, 2018), but not the index of moderated mediation, which is the parameter of primary interest. Statistical power analysis for moderated mediation is complex but still an important step in study planning. This study aims to provide guidance about the impact of model specification on power and thus how the selection of a model should impact sample size planning.

of moderated mediation for Model 7 the correlation between each pair of variables is required. This includes the correlation between all the variables and the interaction (X, M, Y, W, and XW), meaning at least 10 parameters that must be specified. Appropriately specifying these values instead of relying on "rules of thumb" effect sizes can be a challenge (Maxwell, 2000). Currently, for models other than 7 and 14, there are no tools available to assist with this process, meaning that researcher may need to create their own Monte Carlo simulation to evaluate appropriate sample sizes for their models. Statistical power for moderated mediation is complex but still an important step in study planning. This study aims to provide information about sample size adequacy at different effect size levels which could be used by researchers using any of the six models explored in this study.

Model Misspecification in Moderated Mediation

One factor that could impact statistical power and type I error rate is whether or not models are correctly specified (Dupont & Plummer, 1998). Correct specification of a moderated mediation model means that the order of the X, M, and Y variable is correct and that the correct paths in the mediation model are moderated in the analysis. For the purposes of this study, we assume that the order of the mediation variables is correct, and focus on specification of moderation. If the analysis model has too many or too few moderated paths, it is a mispecified model. Some researchers may choose to always moderate all the paths (maximalist), whereas others may try to minimize the number of moderated paths (minimalist). Each of these approaches is likely to impact power and type I error in different ways.

To understand how a data analysis would perform if the model is misspecified, it is helpful to distinguish the data generating process (DGP) from the model used for the data analysis. The former represents the truth in the population (and regression equations are used to generate these data in a simulation study such as this one). The latter is the model corresponding to the set of regression equations fitted with the data, which may differ from the (unknown to researchers) DGP. Based on this distinction, we refer to cases where the

Model Misspecification in Moderated Mediation

Model specification is an important factor that affects type I error rate, power, and parameter bias (Dupont & Plummer, 1998; Rimpler et al., 2024). In the context of this study, we use two pieces of information to determine if a model is misspecified: the data-generating process (DGP) and the data analysis model. The former represents the truth in the population. The latter is the model corresponding to the set of regression equations fitted with the data, which may differ from the DGP. Based on this distinction, we refer to cases where the data analysis model and the DGP do not match as model misspecification. Correct specification of a moderated mediation model means that the order of the X, M, and Y variables and the paths that are moderated are the same in the DGP and analysis model. For the purposes of this study, we assume that the order of the variables is always correct, and focus on specification of moderation. If the analysis model has too many, too few, or the incorrect paths moderated, it is a misspecified model. Some researchers may choose a maximalist approach which would always moderate all the paths, whereas others may choose a minimalist approach which would try to minimize the number of moderated paths. Both approaches can result in model misspecification, but the relative cost of each type of misspecification may differ.

We differentiate model misspecification for moderated mediation into three possible types which can result from maximalist or minimalist approaches. First, a maximalist approach can result in over-specification: All paths that are moderated in the DGP are moderated in the analysis model, plus at least one additional path is allowed to be moderated in the analysis model. For example, when the DGP is Model 7, X to M path moderated, using Model 8 for data analysis, X to M path and X to Y path moderated, is an over-specified model. Introducing extraneous interactions in the model can introduce excessive collinearity (e.g. between XW and MW in a model for Y) and reduce degrees of freedom, each of which may negatively impact power. This is a potential risk of the maximalist approach to model specification.

data analysis model and the DGP do not match as model misspecification.

We differentiate model misspecification for moderated mediation into three possible types. First, there is over-specification: All paths that are moderated in the DGP are moderated in the analysis model, plus at least one additional path is moderated in the analysis model. For example, when the DGP is Model 7, X to M path moderated, using Model 8 for data analysis, X to M path and X to Y path moderated, is an over-specified model. This has implications for statistical power. Introducing extraneous predictors in the model can introduce excessive collinearity, especially with the interactions, and reduce degrees of freedom, each of which may negatively impact power. This is a risk of the maximalist approach to model specification.

Second, there is under-specification: At least one path included in the indirect effect is moderated in both the DGP and data analysis model, but the data analysis model does not include all the moderated paths from the DGP. For example, if the DGP is Model 8 and the analysis model is Model 7, the model is under-specified because the analysis model has omitted the moderated direct effect. The data analysis model could also include additional moderated paths not included in the DGP. For example, if Model 58 is the DGP and Model 8 is used for data analysis, it is under-specified. It fits the criteria for a under-specification because at least one path involved in the indirect effect (here, the X to M path) is moderated in both models, but data analysis with Model 8 does not include the moderation on the M to Y path from the DGP. The data analysis model also moderates the direct effect, which is not moderated in the DGP, but this model still has the ability to detect a significant index of moderated mediation. Under-specification omits important elements of the DGP, which could bias parameters, but may also add unnecessary parameters and could have a complex impact on statistical power. This is a risk of the minimalist approach to model misspecification.

Third, there is complete misspecification, where none of the paths included in the indirect effect are correctly moderated and the index of moderated mediation for the data

Second, a minimalist approach can result in under-specification: At least one path included in the indirect effect is moderated in both the DGP and data analysis model, but the data analysis model does not include all the moderated paths from the DGP. For example, if the DGP is Model 8 and the analysis model is Model 7, the model is under-specified because the analysis model has omitted the moderated direct effect. The data analysis model could also include additional moderated paths not included in the DGP. For example, if Model 58 is the DGP and Model 8 is used for data analysis, we consider this under-specified because Model 8 does not include the moderation on the M to Y path from the DGP, but Model 8 also moderates the direct effect, which is not moderated in the DGP. Under-specification omits important elements of the DGP, which could bias parameters and lead to incorrect conclusions about which paths are moderated (Yzerbyt et al., 2018). This is a potential risk of the minimalist approach to model misspecification.

Minimalist approaches can also lead to complete misspecification, where the DGP includes moderation on a path that is not moderated in the data analysis model, and the data analysis model includes moderation of a path that is not moderated in the DGP. In this case, the index of moderated mediation calculated with the data analysis model should be 0 based on the DGP. For example, when the DGP is Model 7 with the X to M path moderated, using Model 14 (with only the M to Y path moderated) for the data analysis would be a complete misspecification. The index of moderated mediation from Model 14 is a_1b_3, which should be 0 based on the DGP. Moderation on the direct effect is not involved in determining complete misspecification because that path is not used for the index of moderated mediation. Incorrectly specifying where the moderation occurs in the model may lead the estimates of the paths to be biased and incorrect conclusions about which paths are moderated (Yzerbyt et al., 2018).

analysis model should be 0 according to the DGP. For example, when the DGP is Model 7 with the X to M path moderated, using Model 14 for the data analysis would be a complete misspecification because the X to M path is not moderated but the M to Y path is moderated. The index of moderated mediation from Model 14 is ab_3, which should be 0 base don the DGP. Moderation on the direct effect is not involved in determining complete misspecification because that path is not used for the index of moderated mediation. By incorrectly specifying where the moderation occurs in the model, researchers may get biased estimates of the paths, coming to incorrect conclusions about which paths are moderated.

Systematic Review

We performed a large-scale systematic literature review to better understand current practices in moderated mediation analyses. We aimed to explore two questions: 1) Which moderated mediation models are most commonly used?, 2) What sample sizes are typical for moderated mediation models. Papers were chosen to be included in the systematic review through a search on WebofScience for papers published in the year 2018 including keywords "moderated mediation," "mediated moderation," and "conditional process analysis." We identified and coded 411 unique moderated mediation analyses. For our study, we chose the six most commonly used moderated mediation models based on these papers, accounting for 86% of published models from the systematic review. Table 2 shows this percentage separated out by model. We collected other information about the analyses as well, such as types of variables that are most commonly used. Sample sizes ranged from 29 to 456,849, with a mean of 2472 and a median of 285. Median sample sizes separated out by model are given in Table 2. These results directly inform the parameters chosen for the proposed simulation study. This was done in an effort to make the results as useful and applicable as possible for researchers using moderated mediation. All of the papers included in this systematic review plus additional papers from more recent years are available in a searchable database:
`https://www.jlfossum.com/moderated-mediation-article-database`.

Table 2

Systematic Review Models and Sample Sizes

	7	8	14	15	58	59
Use Frequency	31%	13%	18%	3%	6%	14%
Median Sample Size	261	331	288	255	276	363

Note. Each column represents one of the six models used in this registered report

Current Study

This study examined the effect of model specification (over-, under-, or correctly specified) on statistical power and type I error rate in commonly used moderated mediation models. These effects were examined across a variety of realistic conditions: sample sizes, effect size of the interaction, and both dichotomous and continuous moderators and *X* focal predictor variables. As described earlier, model misspecification is when the data analysis model does not match the "true" underlying DGP. Mainly, we explored the effect of an over-specified or under-specified (vs. correctly specified) model on statistical power. Power was only assessed for over-specified models because we hypothesized including in additional interactions could reduce statistical power, and under-specified models because the index of moderated mediation of an under-specified model would still be detecting that the mediation was moderated. We also explored the effect of a completely misspecified (vs. correctly specified) model on type I error rate. Type I error rate was only assessed in completely misspecified models because excluding the "true" interaction means the null with regards to the index of moderated mediation is always false.

Research Question 1 examined what factors impact statistical power of the index of moderated mediation. We hypothesize that statistical power of the index of moderated mediation will be higher for correctly specified models compared to over-specified models (H1a). We also hypothesize that power will be higher for models with fewer moderated paths (H1b). If the DGP is Model 7, for example, Models 8, 58, and 59 are over-specified. Models 8 and 58 have two moderated paths but Model 59 has three. we hypothesize that

Current Study

This simulation study examined the effect of maximalist and minimalist approaches to model specification (correctly, over-, under-, or completely misspecified) on statistical power, type I error rate, and parameter bias in commonly used moderated mediation models. Table 1 gives which data analysis models are considered an over-specification, under-specification, or complete misspecification based on the DGP. When comparing over- and under-specified models to correctly-specified models, we focused on statistical power, given that in both cases, a positive test of the index of moderated mediation would be detecting true moderated mediation in the population even though the model is misspecified. When examining completely misspecified models, we focused on the type I error rate, given that for these models the index of moderated mediation in the analysis model is zero at the population level. For all types of models, we examine parameter bias, as model misspecification may also result in biased parameters, which can provide insight into patterns of type I error and power.

Research Question 1 examines the consequences of the maximalist approach: specifically, how over-specification impacts the statistical power of the index of moderated mediation and parameter bias. We hypothesized that the statistical power of the index of moderated mediation would be lower for over-specified models compared to correctly specified models (H1a). We also hypothesized that, within the set of over-specified models, power would be lower for models with more moderated paths (H1b). Finally, we hypothesized that parameter bias for over-specified models would be acceptable (<10%) in each condition (H1c).

Research Question 2 examines the consequences of the minimalist approach: specifically, how under-specification impacts the statistical power of the index of moderated mediation and parameter bias. We hypothesized that the statistical power of the index of moderated mediation would be lower for under-specified models compared to correctly specified models (H2a). We also hypothesized that parameter bias would be unacceptable for cases like this fewer moderated paths (Models 8 and 58) will have higher power than more moderated paths (Model 59).

Related to RQ1, we examined these factors affecting statistical power for the index of moderated mediation for under-specified models. We treated these models as exploratory, and do not have specific hypotheses for statistical power in these models. We will examine the effect of under-specification compared to correct specification (H1.1a), the number of moderated paths in the model (H1.1b).

Research Question 2 examined what factors impact type I error rate of the index of moderated mediation. Type I error rate was calculated when the model is completely misspecified. For complete misspecification, the index of moderated mediation of the data analysis model should be 0 given the DGP. So if the index of moderated mediation is statistically significant, this would be a type I error. We treated these models as exploratory, though we hypothesize that type I error rate would be too high in completely misspecified models (H2a) and type I error rate will increase as the number of incorrectly moderated paths increases as well (H2b).

This study tested six common arrangements of variables in moderated mediation analyses, and compared the statistical power and type I error rate from the different arrangements across effect sizes, sample sizes, and variable types common in the current moderated mediation literature. Conclusions from this study will inform the degree to which model specification and complexity impact statistical power and type I error rates in moderated mediation models. We will also provide general sample size planning recommendations for researchers using these particular moderated mediation models.

(>10%) for under-specified models (H2b).

Research Question 3 examined another consequence of the minimalist approach: how complete misspecification impacts the type I error rate for a test on the index of moderated mediation. We hypothesized that the type I error rate would be too high (liberal) in completely misspecified models (H3a). Additionally, we hypothesized that raw bias would be unacceptably high (H3b).

In total, we tested six common moderated mediation specifications, and we tested the above hypotheses across effect sizes, sample sizes, and variable types common in the current literature. Conclusions from this study inform the degree to which model specification and number of moderated paths impact statistical power, type I error rates, and parameter bias in moderated mediation models. We use this information to provide guidance for study planning with moderated mediation; in particular, how model specification should impact sample size planning.

Method

The goal of any model specification approach is to correctly specify the model; however, researchers may find themselves unsure about whether to allow certain paths in a moderated mediation to be moderated. For example, a researcher may hypothesize that the path from X to M is moderated and the path from M to Y is not, but have no clear hypothesis about the direct effect. Should that researcher select Model 7 (no moderated direct effect) or Model 8 (moderated direct effect)? These decisions map onto maximalist and minimalist approaches to model specification, both of which can result in model misspecification. The goal of this simulation study was to understand how model misspecification affects statistical power, type I error rate, and parameter bias in moderated mediation models.

We generated data using each one of the six DGPs, and then fit the data using all six data analysis models, one of which was correctly specified. Models 58 and 59 were not used for generation and analysis when the moderator was continuous. We recorded whether

Table 3

Analysis Model Specification based on DGP

DGP	Over-specified	Under-specified	Completely Misspecified
7	8, 58, 59		14, 15
8	59	7, 58	14, 15
14	15, 58, 59		7, 8
15	59	14, 58, 7, 8	
58	59	7, 8, 14, 15	
59		7, 8, 14, 15, 58	

Note. Moderated mediation DGP models (first column) and which *analysis* models are over-specified, under-specified, or completely misspecified for that DGP.

Method

The goal of this simulation study was to understand how model misspecification affects statistical power and type I error rate in moderated mediation models. There are two important attributes for each analysis: the DGP, (the true population moderated mediation model used to generate the data) and the data analysis model (the model we used to analyze the data).

We generated data using one of six DGPs, then fit the data using six data analysis models, one of which is correctly specified. We only used Models 58 and 59 for generation and analysis when the moderator was dichotomous. We recorded statistical power and/or type I error rate for each case, depending on if the analysis model is correctly specified (power), over-specified (power), under-specified (power), or completely misspecified (type I error rate).

Simulation Conditions

We used a Monte Carlo simulation with an incomplete 6 (Between: Generating Model) x 9 (Between: Sample Size) x 6 (Between: Effect Size) x 2 (Between: Normal or Dichotomous X) x 2(Between: Normal or Dichotomous W) x 6 (Within: Analysis Model) factorial design. The design is incomplete because Model 58 and 59 were only used as generating and/or analysis models when W was dichotomous. Each sample was analyzed the confidence interval for the index of moderated mediation excluded zero, which reflects statistical power (correctly, over-, and under-specified models) or type I error rate (completely misspecified models). We recorded parameter bias for the index of moderated mediation for all analysis models. Effects were examined across a variety of realistic conditions: sample sizes, the effect size of the interaction term(s) in the model, and both dichotomous and continuous W and X variables.

Simulation Conditions

We used a Monte Carlo simulation with an incomplete 6 (Between: Generating Model) x 9 (Between: Sample Size) x 3 (Between: Effect Size) x 2 (Between: Normal or Dichotomous X) x 2 (Between: Normal or Dichotomous W) x 6 (Within: Analysis Model) factorial design. Table 2 lists each condition and the levels used. The design is incomplete because Models 58 and 59 were only used to generate and analyze data when W was dichotomous because the index of moderated mediation is undefined in these models when W is continuous.

Simulation Procedure

We used GAUSS 21 on a Windows server for data generation, generating 5000 samples of data in each condition. We used the 10th and 90th percentiles of the sample sizes seen in our systematic review (Appendix A) as the maximum and minimum sample sizes examined in the simulation. Thus, we considered the following sample sizes: 100, 150, 200, 250, 300, 400, 500, 750, and 1000 as those corresponded to the deciles (when rounded). Four variables were generated: the predictor X, the mediator M, the outcome Y, and the moderator W. In all cases, X and the moderator W were independent. Data for each effect size combination and sample size were generated in each of the four (continuous W) or six (dichotomous W) different moderated mediation model configurations. Data were generated under these six conditions (see Figure 1: Model 7, Model 8, Model 14, Model 15, Model 58 (dichotomous W only), and Model 59 (dichotomous W only). We focused on observed variable systems, and since ordinary least squares (OLS) regression provides the with all four (continuous W) or six (dichotomous W) analysis models aligned with the generating models defined in Table 1. Effect size on the interaction term and sample size were varied. We generated data with continuous and dichotomous X and W variables. The specifics of these conditions are described in greater detail below.

Simulation Procedure

We used GAUSS 21 on a Windows server for data generation, generating 5000 samples of data in each condition. Based on quantiles from the systematic review, we used sample sizes of 100, 150, 200, 250, 300, 400, 500, 750, and 1000.

Four variables were generated: the primary predictor X, the mediator M, the primary outcome Y, and the moderator W. In all cases, X and the moderator W were independent. Data for each effect size combination and sample size were generated in each of the four (continuous W) or six (dichotomous W) different moderated mediation model configurations. Data were generated under these six conditions: W moderating only the X to M path (Model 7), the X to M path and X to Y path (Model 8), only the M to Y path (Model 14), the M to Y path and X to Y path (Model 15), the X to M path and M to Y path (Model 58, dichotomous W only), or all paths (Model 59, dichotomous W only). We focus on observed variable systems, and since OLS provides the same estimates as maximum likelihood in this case but is computationally less complex (Hayes, Montoya, & Rockwood, 2017), we use OLS throughout to estimate coefficients.

We set the variance explained by the X to M path and the M to Y path at 7% each as a commonly seen effect size in psychological research (Fritz & MacKinnon, 2007), with each interaction accounting for an additional 1%, 3%, or 5% of explained variance (McClelland & Judd, 1993). When multiple interactions were included in the model, they were all set to be the same size. Additionally, when W is included as an interaction, it also has a main effect set to explain 7% of the variance in the outcome (e.g., a_2, c'_2, or b_2). Path coefficients were calculated correspondingly by taking the square root of these R^2 effect sizes. For example, the X to M path explaining 7% of the variance has path coefficient same coefficient estimates as maximum likelihood in this case but is computationally less complex (Hayes et al., 2017), we used OLS regression to estimate coefficients.

The process for generating X, W, M, and Y was as follows. First, X and W were independently generated, either drawn from a standard normal distribution or dichotomous coded -1 and 1 with equal allocation to keep the variance at 1. From there, depending on the moderated mediation model chosen as the DGP, Equations 1 or 2 were used to first generate M, then use M in addition to other varied parameters to generate Y using Equations 3 - 6. Residuals for both models were generated from a normal distribution with mean 0 and the standard deviation set such that the standard deviation of the outcome (M or Y) is always 1 (i.e., standardized). For example, we used the path coefficients and adjusted the standard deviation of the residuals to be $\sqrt{1-(a_1^2+a_2^2+a_3^2)}$, where $a_1 = .26$, $a_2 = .26$, and $a_3 = .10$, .17, and .22.

The variance of the product term was equal to one in expectation, based on how we generated the predictor variables (X and W) to always have a variance of one and a mean of zero, relying on the assumption of independence. [1]

We set the variance explained by the X to M path (a_1) and the M to Y path (b_1) at 7% each as a commonly seen effect size in psychological research (Fritz & MacKinnon, 2007), with each interaction accounting for an additional 1%, 3%, or 5% of explained variance (McClelland & Judd, 1993). When multiple interactions were included in the model, they were all set to be the same size. Additionally, when W was included in an interaction, it also had a coefficient set to explain 7% of the variance in the outcome (e.g., a_2, c_2', or b_2). Path coefficients were calculated correspondingly by taking the square root of these R^2 effect sizes. For example, the X to M path explaining 7% of the variance has path

[1] We relied on the following equation to generate a product term with a variance of 1: $Var(XW) = Var(X)Var(W) + Var(X)(E(W))^2 + Var(W)(E(X))^2$ which applies if X and W are independent. We generated both W and X to have $E(X) = E(W) = 0$ and $Var(X) = Var(W) = 1$. This sets the variance of the product term to be 1 in expectation but is not fixed to be 1 in any given sample due to sampling variability.

$\sqrt{.07} = .26$ when X and M are standardized. Residuals were generated from a normal distribution centered at 0, with the standard deviation incorporating the path coefficients to ensure the proportion of explained variance remains as expected and the standard deviation of the outcome is always 1 (i.e., standardized). The standard deviation of the residuals was adjusted to be $\sqrt{1 - (a_1^2 + a_2^2 + a_3^2)}$, where $a_1 = .26$, $a_2 = .26$, and $a_3 = .10$, .17, or .22 (or -.10, -.17, -.22 to account for negative interaction effects).

To keep the variance equal to one in the dichotomous case, as was set in the continuous case when generating values from a standard normal distribution, the two categories were coded as -1 and 1. For this same reason, equal allocation was used in the dichotomous conditions. Interaction terms such as XW also had a variance of one. All variables involved in the interactions were centered as part of data generation, so that the lower order coefficients are conditioned on the mean of the other variable.

Data analysis models were then fit to each sample of generated data. Models were analyzed using the percentile bootstrap confidence interval set at 95% with 1000 bootstraps (Efron & Tibshirani, 1994).

Performance Metrics

There are two outcomes of interest in this study: statistical power and type I error rate for the index of moderated mediation. Both are calculated as the proportion of the 5000 generated samples that have a statistically significant result in each condition. Whether the result is power or type I error depends on the specification of the analysis model: correctly specified (power), over-specified (power), underspecified (power), completely misspecified (type I error). Table 3 gives which data analysis models would be considered an over-specification, under-specification, or complete misspecification from the DGP. We calculated rejection rate for the index of moderated mediation for Models 7, 8, 14, and 15 with both dichotomous and continuous W, and for Models 58 and 59 with dichotomous W (Fairchild & MacKinnon, 2009).

Power was calculated when the model is correctly specified, over-specified, or coefficient $\sqrt{.07} = .26$ when X and M are standardized.

Once data generation was complete, data analysis models were fit to each sample of generated data. Each of the 5000 samples was analyzed with all four (continuous W) or six (dichotomous W) analysis models. Inference for the index of moderated mediation was conducted using the percentile bootstrap confidence interval set at 95% with 1000 bootstraps (Efron & Tibshirani, 1994). The decision to reject the null hypothesis was based on the confidence interval recorded for each model for each sample excluding zero.

Performance Metrics

There were three outcomes of interest in this study: statistical power, type I error rate, and parameter bias for the index of moderated mediation. The first two are rejection rates calculated as the proportion of the 5000 generated samples within each condition where the null hypothesis is rejected (confidence interval excludes zero), which indicates the type I error rate when the true index is zero and power otherwise.

Power was calculated when the model is correctly specified, over-specified, or under-specified. Correctly specified models provide a baseline power level that can be used to compare to the over- and under-specified models. Rejection rates from over-specified models indicate power because while additional parameters not in the DGP are included in the data analysis model, a significant index of moderated mediation would still appropriately detect a true effect. Similarly, power was determined for under-specified models because these models should still have a significant index of moderated mediation based on their DGP.

Type I error rate was calculated for completely misspecified models. A significant index of moderated mediation would have to arise from an interaction that is 0 in the population. Because there is no comparison group for type I error, and previous simulations on moderated mediation analysis have found that type I error rates often differ from 0.05 for correctly specified models (Coutts, 2023; Yzerbyt et al., 2018), we use the liberal criterion from Bradley (1978) (.025 to .075) to classify type I error rates as overly under-specified. Correctly specified models should accurately detect effects, providing a baseline power level that can be used to compare to the misspecified models. Rejection rates from over-specified models indicate power, because while additional parameters are included in the DGP that are not in the data analysis model, a significant index of moderated mediation would still be detecting an effect that truly exists. Power was also determined for under-specified models because these models should still have a significant index of moderated mediation based on their DGP. For both over-specified and under-specified models, a path with an interaction term from the DGP must be included in the data analysis model, so a significant effect is detecting an effect that truly exists from the DGP.

Type I error rate was calculated the same as power, but for models where the data analysis model was completely misspecified. We include only completely misspecified models because these are the cases where a type I error is possible. A significant index of moderated mediation would have to be from an interaction in the data analysis model that is nonexistent in the DGP, meaning it is 0 in the population. Because there is no comparison group for type I error, and previous simulations in mediation analysis have found that type I error rates are often differ from 0.05 for correctly specified model, we use the criteria from Bradley (1978) and Serlin (2000) to classify type I error rates as overly conservative or liberal.

Analysis Plan

To test our hypotheses about model specification on power and type I error rate, we will use multilevel logistic regression with random intercepts only to predict rejection. Rejection is a binary 0/1 indicator from the simulation where 0 indicates the confidence interval includes 0, and 1 indicates 0 was excluded from the confidence interval. To test the significance of the main effect of the factors and all possible two-way through six-way interactions we will use an approach aligned with type II sums of squares because it tests each main effect in the model considering the other main effects in the model, but does not

account for interactions when testing main effects. For the model with the interactions, only effects with an odds ratio greater than 1.68 will be considered meaningful, because that corresponds to a small effect (Chen, Cohen, & Chen, 2010). We chose this effect size metric instead of statistical significance due to the large amount of data likely favoring statistical significance. For significance tests we plan to use a threshold of $p < .001$ for significant effects.

To test Hypothesis 1, we will only include correctly specified and over-specified models. The multilevel logistic regression will have 6 main effects: model specification (over vs. correct), generating model (dummy coded), sample size (sequentially coded), effect size (sequentially coded), type of X (continuous vs. dichotomous), and number of moderated paths in the analysis model (sequentially coded). We fit two separate models: one for continuous W and one for dichotomous W, since we had an incomplete design where Model 58 and 59 were only used as generating and analysis models when W was dichotomous. In our exploratory analysis we will examine a model which includes all possible interactions, separating effect size into two predictors (magnitude and sign). We also plan to make tables and figures to provide the raw rejection rates in each condition, and visualize patterns across conditions.

Hypothesis 1a would be supported if we find a significant coefficient for model specification (over vs correct) such that power is lower when models are over-specified. We would expect to see this in both the model with continuous W and dichotomous W. However, if it is non-significant in both models, we would conclude that over-specification does not negatively impact power. If it is significant in the hypothesized direction for just one model, we would interpret this as partial support for Hypothesis 1a. If it is significant and in the hypothesized direction for both models, we will interpret this as complete support for Hypothesis 1a. If any of the results are significant and in the opposite direction as hypothesized, we will interpret these results appropriately and explore the underlying cause of these unintuitive findings. We will explore whether there are factors that interact conservative or liberal.

Parameter bias was calculated using relative bias ($\frac{estimate - parameter}{parameter}$ averaged across all replications), where values <10% are acceptable, except in completely misspecified models. Since completely misspecified models preclude calculating relative bias, we used raw bias for these cases ($estimate - parameter$, averaged across all replications). A raw bias of .00286 was considered unacceptable, corresponding to the 10% relative bias value for the smallest effect size condition evaluated in this study. We calculated all of these quantities for Models 7, 8, 14, and 15 with both dichotomous and continuous W, and for Models 58 and 59 with dichotomous W across all the conditions.

Analysis Plan

We now describe how we tested our hypotheses about the consequences of maximalist and minimalist approaches to model specification. When our analysis involved significance testing, we set $\alpha = .001$. We also report 99.9% confidence intervals and odds ratios to contextualize the results further.

H1a-1c focused on over-specified models. To test H1a (lower power for over-specified models), we used only cases with correctly or over-specified models. We used a multilevel logistic regression model with random intercepts for the data analysis model (within-subjects factor since each generated sample of data is analyzed using all six data analysis models) to predict rejection. The model had six main effects: model specification (over vs. correct), generating model (dummy coded with Model 7 as the reference category), sample size (sequentially coded), effect size (sequentially coded), type of X (continuous vs. dichotomous), and number of moderated paths in the analysis model (sequentially coded). We fit two separate models: one for continuous W and one for dichotomous W since we had an incomplete design where Models 58 and 59 were only used as generating and analysis models when W was dichotomous (see Table 2 for a list of conditions). H1a would be supported if we find a significant coefficient for model specification (over- vs correct) such that power is lower when models are over-specified for with over vs. correct specification to determine power. Hypothesis 1.1a will be tested in the same manner but comparing underspecified to correctly specified models. We have no directional hypotheses for this test. We believe that in some cases underspecification may increase power and in other cases decrease power. We will use the exploratory interactions to identify such cases.

In Hypothesis 1b, we hypothesize that for over-specified models, power will be lower when these models include more moderated paths. This will be tested using a multilevel logistic regression model including all 6 main effects and an interaction between model specification (with overspecified models coded as zero) and number of moderated paths (sequentially coded). The hypothesis will be supported would be supported by a significant effect of number of moderated paths in the analysis model (tested using a Wald test for the set of coefficients). We hypothesize that more moderated paths would lead to lower power in overspecified model, and so we hypothesize that all the coefficients will be negative. We would expect to see this in both the model with continuous W and dichotomous W. However, if the test is non-significant in both models, we would conclude that number of moderated paths does not impact power. If the test is significant and in the hypothesized direction for just one model, we would interpret this as partial support for Hypothesis 1b. If the test is significant and in the hypothesized direction for both models, we will interpret this as complete support for Hypothesis 1b. If any of the results are significant and in the opposite direction as hypothesized, we will interpret these results appropriately and explore the underlying cause of these unintuitive findings. We will explore whether there are other factors that interact with number of moderated paths to determine power. Hypothesis 1.1b will be tested in the same manner but comparing underspecified to correctly specified models. We have no directional hypotheses for this test. We believe that in some cases more paths may increase power and in other cases decrease power. We will use the exploratory interactions to identify such cases.

To test Hypothesis 2, we only included completely misspecified analysis models. To both continuous and dichotomous W. To test H1b (lower power for over-specified models with more moderated paths) we used only cases with over-specified models and an adapted version of the model from H1a, which removes the first main effect (model specification). H1b would be fully supported if all four coefficients for the number of moderated paths are significant, such that power is lower when there are two compared to one, and three compared to two moderated paths for both continuous and dichotomous W. If only some of the coefficients are significant in the predicted direction, H1b would be partially supported. To test H1c (acceptable parameter bias for over-specified models), we simply interpreted the parameter bias resulting from these models. H1c would be fully supported if few (<10%) of the conditions result in a relative bias value of over 10%. Partial support would be if between 10%-20% of the conditions resulted in a relative bias value of over 10%. If we see relative bias over 10% in over 20% of conditions, H1c is not supported, and we would interpret this as a particularly high risk for a maximalist approach.

H2a-2b focused on under-specified models. To test H2a (lower power for under-specified models), we used the same multilevel logistic regression model as in H1a, adapting the first main effect (model specification) to compare under-specified to correctly specified models. Again, we fit two separate models: one for continuous W and one for dichotomous W. H2a would be supported if we find a significant coefficient for model specification (under- vs correct) such that power is lower when models are under-specified for both continuous and dichotomous W. H2b (unacceptable parameter bias for under-specified models) was tested similarly to H1c. H2b would be fully supported if many (>20%) of the conditions result in a relative bias value of over 10%. Partial support would be if between 10%-20% of the conditions resulted in a relative bias value of over 10%. If H2b is fully or partially supported, we will examine patterns among unacceptable bias values. If we see relative bias over 10% in over 20% of conditions, H2b is not supported, we would see this as a particularly high risk for a minimalist approach.

H3a-3b focused on completely misspecified models. To test H3a (inflated type I examine Hypothesis 2a, we will evaluate how many conditions have type I errors above the criteria set by Bradley (1978) and Serlin (2000). If the majority of conditions show elevated type I error rates, Hypothesis 2a will be supported. In addition, we will explore what factors predict type I error rate using a multilevel logistic regression with 5 main effects and all possible interaction: generating model (dummy coded), sample size (sequentially coded), effect size (sequentially coded), type of X (continuous vs. dichotomous), and number of moderated paths in the analysis model (sequentially coded). We fit two separate models: one for continuous W and one for dichotomous W. Hypothesis 2b will be tested using a multilevel logistic regression with only the main effects above, and a Wald test for number of moderated paths. Hypothesis 2b would be supported if the test is significant (in both models) and the coefficients are in the hypothesized direction. If the effect is only significant in one model, or some paths are in the opposite direction, we will interpret these results accordingly. We will also use the model with the interactions to explore factors which might interact with number of paths to predict type I error.

error rate for completely misspecified models), we interpreted the type I error rates resulting from the completely misspecified models. H3a would be supported if a non-negligible proportion of conditions (>10%) result in a type I error rate > .075. To test H3b (unacceptable parameter bias for completely misspecified models), we interpreted the raw bias resulting from the different simulation conditions. H3b would be fully supported if >20% of conditions result in a raw bias value above .00286. Partial support would be if between 10%-20% of the conditions resulted in a raw bias value above .00286. If H3b is fully or partially supported, we will examine patterns among unacceptable bias values, and if the proportion of unacceptable values exceeds 50%, we would see this as a particularly high risk for a minimalist approach.

Tables and figures with type I error, power, and parameter bias for each appropriate condition will be presented. Due to the complexity of the design, we may generate tables or figures for a subset of conditions (e.g., only dichotomous moderators) for clarity of presentation, but we will provide the corresponding plot for the remaining conditions (e.g., continuous moderators) in an appendix for completeness of reporting. Table B1 provides an example of a table for power, Table B2 provides an example of a table for parameter bias, Table B3 provides an example of a table for type I Error, Figure B1 provides an example of a figure for power, and Figure B2 provides an example of a figure for type I error rate.

Data Availability Statement

All data will be made available on the OSF page for this study. The GAUSS simulation code to generate the data, a .csv file of the simulation results, and the R analysis script will all be posted at https://osf.io/vgkdt/.

Stage 1 Registered Report

At the time of submission as a Stage 1 registered report, pilot data have been generated and analyzed as part of the first author's dissertation study. However, data for this study have not yet been generated and no analyses have been completed. Simulation code has already been written to generate data, and the script for data analysis has also

References

Aberson, C. L. (2019a). *Applied power analysis for the behavioral sciences* (2nd ed.). New York: Routledge.

Aberson, C. L. (2019b). *pwr2ppl:Power analysis for common designs. R package version 0.1.* Retrieved from `https://cran.r-project.org/web/packages/pwr2ppl/index.html`

Aguinis, H. (1995). Statistical power problems with moderated multiple regression in management research. *Journal of Management*, *21*(6), 1141–1158.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, *90*(1), 94-107.

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the ”replication crisis”: Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, *52*(3), 305–324. doi: 10.1080/00273171.2017.1289361

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069–1077.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Charlton, A., Montoya, A. K., Price, J., & Hilgard, J. (2021). Noise in the process: an assessment of the evidential value of mediation effects in marketing journals. doi: 10.31234/osf.io/ck2r5

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics - Simulation and Computation*, *39*(4), 860–864. doi: 10.1080/03610911003650383

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale:

already been written. Both are available on the OSF page for the study:
https://osf.io/vgkdt/.

References

Aberson, C. L. (2019). Pwr2ppl: Power analysis for common designs. R package version 0.1.

Aguinis, H. (1995). Statistical power problems with moderated multiple regression in management research. *Journal of Management*, *21*(6), 1141–1158.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, *90*(1), 94–107.

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, *52*(3), 305–324. https://doi.org/10.1080/00273171.2017.1289361

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069–1077.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/https://doi.org/10.1016/j.jml.2012.11.001

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Brysbaert, M. (2007). The language-as-fixed-effect fallacy: Some simple spss solutions to a complex problem. *Report for EU Marie Curie Research Training Network: Language and Brain*.

Charlton, A., Montoya, A. K., Price, J., & Hilgard, J. (2021). Noise in the process: An assessment of the evidential value of mediation effects in marketing journals. https://doi.org/10.31234/osf.io/ck2r5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd). NJ: Lawrence Erlbaum Associates.

NJ: Lawrence Erlbaum Associates.

Coutts, J. (2023). *Contrasting contrasts: An exploration of methods for comparing indirect effects in mediation models* (Unpublished doctoral dissertation). The Ohio State University.

Dupont, W. D., & Plummer, W. D. (1998). Power and sample size calculations for studies involving linear regression. *Controlled Clinical Trials*, *19*(6), 589–601. doi: 10.1016/s0197-2456(98)00037-3.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 1–11. doi: 10.3389/fpsyg.2015.00621

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods*, *12*(1), 1–22. doi: 10.1037/1082-989X.12.1.1

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap.* Boca Raton, FL: Chapman Hall/CRC.

Fairchild, A. J., & MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science*, *10*(2), 87–99. doi: 10.1007/s11121-008-0109-6

Fairchild, A. J., & McDaniel, H. L. (2017). Best (but oft-forgotten) practices: mediation analysis. *The American Journal of Clinical Nutrition*, *105*(6), 1259–1271. doi: https://doi.org/10.3945/ajcn.117.152546

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests – an analysis of articles published in 2015. *Journal of Experimental Psychology*, *75*. doi: 10.1016/j.jesp.2017.11.008

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

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated

Coutts, J. (2023). *Enhancing the specification, testing, and interpretation of conditional indirect effects* [Dissertation]. The Ohio State University.

Dupont, W. D., & Plummer, W. D. (1998). Power and sample size calculations for studies involving linear regression. *Controlled Clinical Trials*, *19*(6), 589–601. https://doi.org/10.1016/s0197-2456(98)00037-3

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 1–11. https://doi.org/10.3389/fpsyg.2015.00621

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods*, *12*(1), 1–22. https://doi.org/10.1037/1082-989X.12.1.1

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman; Hall/CRC.

Fairchild, A. J., & MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science*, *10*(2), 87–99. https://doi.org/10.1007/s11121-008-0109-6

Fairchild, A. J., & McDaniel, H. L. (2017). Best (but oft-forgotten) practices: Mediation analysis. *The American Journal of Clinical Nutrition*, *105*(6), 1259–1271. https://doi.org/https://doi.org/10.3945/ajcn.117.152546

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests – an analysis of articles published in 2015. *Journal of Experimental Psychology*, *75*. https://doi.org/10.1016/j.jesp.2017.11.008

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

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*(3), 233–239. https://doi.org/10.1111/j.1467-9280.2007.01882.x effect. *Psychological Science*, *18*(3), 233–239. doi: 10.1111/j.1467-9280.2007.01882.x

Götz, M., O'Boyle, E. H., Gonzalez-Mulé, E., Banks, G. C., & Bollmann, S. S. (2021). The "Goldilocks" zone: (Too) many confidence intervals in tests of mediation just exclude zero. *Psychological Bulletin*, *147*(1), 95–114. doi: 10.1037/bul0000315

Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate Behavioral Research*, *50*(1), 1 – 22. doi: 10.1080/00273171.2014.962683

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis* (3rd ed.). New York, NY: Guilford Press.

Hayes, A. F., Montoya, A. K., & Rockwood, N. J. (2017). The analysis of mechanisms and their contingencies: PROCESS versus structural equation modeling. *Australasian Marketing Journal*, *25*(1), 76–81. doi: 10.1016/j.ausmj.2017.02.001

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi: 10.1371/journal.pmed.0020124

Kenny, D. A. (2017). *An interactive tool for the estimation of power in tests of mediation [computer software].* Retrieved from https://davidakenny.shinyapps.io/MedPower/

Marshall, S. W. (2007). Power for tests of interaction: effect of raising the type i error rate. *Epidemiologic perspectives & innovations: EP+ I*, *4*(1), 4. doi: 10.1186/1742-5573-4-4

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods*, *5*(4), 434–458. doi: 10.1037/1082-989X.5.4.434

Maxwell, S. E., & Kelley, K. (2011). Ethics and sample size planning. In A. T. Panter & S. K. Sterba (Eds.), *Handbook of ethics in quantitative methodology* (pp. 159–184). Routledge.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*(2), 376–390. doi: 10.1037/0033-2909.114.2.376

Götz, M., O'Boyle, E. H., Gonzalez-Mulé, E., Banks, G. C., & Bollmann, S. S. (2021). The "Goldilocks" zone: (Too) many confidence intervals in tests of mediation just exclude zero. *Psychological Bulletin*, *147*(1), 95–114. https://doi.org/10.1037/bul0000315

Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate Behavioral Research*, *50*(1), 1–22. https://doi.org/10.1080/00273171.2014.962683

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis* (3rd ed.). Guilford Press.

Hayes, A. F., Montoya, A. K., & Rockwood, N. J. (2017). The analysis of mechanisms and their contingencies: PROCESS versus structural equation modeling. *Australasian Marketing Journal*, *25*(1), 76–81. https://doi.org/10.1016/j.ausmj.2017.02.001

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kenny, D. A. (2017). An interactive tool for the estimation of power in tests of mediation [computer software]. https://davidakenny.shinyapps.io/MedPower/

Marshall, S. W. (2007). Power for tests of interaction: Effect of raising the type I error rate. *Epidemiologic perspectives & innovations: EP+ I*, *4*(1), 4. https://doi.org/10.1186/1742-5573-4-4

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/https://doi.org/10.1016/j.jml.2017.01.001

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*(2), 376–390. https://doi.org/10.1037/0033-2909.114.2.376

O'Rourke, H. P., & MacKinnon, D. P. (2014). When the test of mediation is more powerful than the test of the total effect. *Behavior Research Methods*, *47*, 424–442. https://doi.org/10.3758/s13428-014-0481-z

Neyman, J., & Pearson, E. S. (1993). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, *29*(4), 492–510. doi: 10.1017/S030500410001152X

O'Rourke, H. P., & MacKinnon, D. P. (2014). When the test of mediation is more powerful than the test of the total effect. *Behavior Research Methods*, *47*, 424–442. doi: 10.3758/s13428-014-0481-z

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, *42*(1), 185–227. doi: 10.1080/00273170701341316

Ramos, G., Delgadillo, D., Fossum, J. L., Montoya, A. K., Thamrin, H., Rapp, A., … Chavira, D. (2021). Discrimination and internalizing disorders in rural latinx adolescents: An ecological model of etiology. *Children and Youth Services Review*, *130*, 106250. doi: 10.1016/j.childyouth.2021.106250

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). doi: 10.1177/25152459221095827

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science*, *8*(4), 379 — 386. doi: 10.1177/1948550617715068

Serlin, R. C. (2000). Testing for robustness in monte carlo studies. *Psychological Methods*, *5*(2), 230 – 240.

Yzerbyt, V. Y., Muller, D., Batailler, C., & Judd, C. M. (2018). New recommendations for testing indirect effects in mediational models: The need to report and test component paths. *Journal of Personality and Social Psychology: Attitudes and Social Cognition*, *115*(6), 929–943. doi: 10.1037/pspa0000132

Zhang, Z., & Wang, L. (2013). Methods for mediation analysis with missing data. *Psychometrika*, *78*(1), 154 — 184.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, *42*(1), 185–227. https://doi.org/10.1080/00273170701341316

Rimpler, A., Kiers, H., & van Ravenzwaaij, D. (2024). To interact or not to interact: Pros and cons of including interactions in linear regression models. *PsyArXiv*. https://doi.org/10.31234/osf.io/6utws

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866. https://doi.org/10.1080/01621459.1994.10476818

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). https://doi.org/10.1177/25152459221095827

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science*, *8*(4), 379–386. https://doi.org/10.1177/1948550617715068

Wykes, T., Newton, E., Landau, S., Rice, C., Thompson, N., & Frangou, S. (2007). Cognitive remediation therapy (CRT) for young early onset patients with schizophrenia: An exploratory randomized controlled trial. *Schizophrenia Research*, *94*(1-3), 221–230.

Wykes, T., & Spaulding, W. (2011). Thinking about the future cognitive remediation therapy - what works and could we do better? *Schizophrenia Bulletin*, *37*(Suppl. 2), S80–S90. https://doi.org/10.1093/schbul/sbr064

Wykes, T., Reeder, C., Huddy, V., Taylor, R., Wood, H., Ghirasim, N., Kontis, D., & Landau, S. (2012). Developing models of how cognitive improvements change functioning: Mediation, moderation and moderated mediation. *Schizophrenia Research*, *138*(1), 88–93. https://doi.org/10.1016/j.schres.2012.03.020

Zhang, Z., & Yuan, K. H. (2018). *Practical statistical power analysis using webpower and R*. Granger, IN: ISDSA Press.

Yzerbyt, V. Y., Muller, D., Batailler, C., & Judd, C. M. (2018). New recommendations for testing indirect effects in mediational models: The need to report and test component paths. *Journal of Personality and Social Psychology: Attitudes and Social Cognition*, *115*(6), 929–943. https://doi.org/10.1037/pspa0000132

Zhang, Z., & Wang, L. (2013). Methods for mediation analysis with missing data. *Psychometrika*, *78*(1), 154–184.

Zhang, Z., & Yuan, K. H. (2018). *Practical statistical power analysis using webpower and R*. ISDSA Press.

Appendix

Study Design Template

Table 1

Analysis Model Specification based on DGP

DGP	Over-specified	Under-specified	Completely Misspecified
7	8, 58, 59		14, 15
8	59	7, 58	14, 15
14	15, 58, 59		7, 8
15	59	14, 58	7, 8
58	59	7, 8, 14, 15	
59		7, 8, 14, 15, 58	

Note. Moderated mediation DGP models (first column) and which *analysis* models are over-specified, under-specified, or completely misspecified for that DGP. All model numbers are from the PROCESS model numbering system (Hayes, 2022).

Question	Hypothesis	Sampling plan	Analysis Plan	Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes
Research Question 1 examined what factors impact statistical power of the index of moderated mediation.	We hypothesize that statistical power of the index of moderated mediation will be higher for correctly specified models compared to over-specified models (H1a). We also hypothesize that power will be higher for models with fewer moderated paths (H1b). If the DGP is Model 7, for example, Models 8, 58, and 59 are over-specified. Models 8 and 58 have two moderated paths but Model 59 has three. we hypothesize that for cases like this fewer moderated paths (Models 8 and 58) will have higher power than more moderated paths (Model 59).	*Applies to all Research Questions:* We generated data using one of six DGPs, then fit the data using six data analysis models, one of which is correctly specified. We only used Models 58 and 59 for generation and analysis when the moderator was dichotomous. We recorded statistical power and/or type I error rate for each case, depending on if the analysis model is correctly specified (power), over-specified (power), under-specified (power), or completely misspecified (type I error rate). We used a Monte Carlo simulation with an incomplete 6 (Between: Generating Model) x	To test our hypotheses about model specification on power and type I error rate, we will use multilevel logistic regression with random intercepts only to predict rejection. Rejection is a binary 0/1 indicator from the simulation where 0 indicates the confidence interval includes 0, and 1 indicates 0 was excluded from the confidence interval. To test the significance of the main effect of the factors and all possible two-way through six-way interactions we will use an approach aligned with type II sums of squares because it tests each main effect in the model considering the other main effects in the model,	*Applies to all Research Questions:* Only effects with an odds ratio greater than 1.68 will be considered meaningful, because that corresponds to a small effect (Chen et al., 2010). We chose this effect size metric instead of statistical significance due to the large amount of data likely favoring statistical significance. For significance tests we plan to use a threshold of p < .001 for significant effects.	Hypothesis 1a would be supported if we find a significant coefficient for model specification (over vs correct) such that power is lower when models are over-specified. We would expect to see this in both the model with continuous W and dichotomous W. However, if it is non-significant in both models, we would conclude that over-specification does not negatively impact power. If it is significant in the hypothesized direction for just one model, we would interpret this as partial support for Hypothesis 1a. If it is significant and in the hypothesized direction for both models, we will interpret this as complete support for Hypothesis 1a. If any	

Table 2

Simulation Conditions

Design Factor		Levels
Generating Model (6)	Between	7 | 8 | 14 | 15 | 58 | 59
Sample Size (9)	Between	100 | 150 | 200 | 250 | 300 | 400 | 500 | 750 | 1000
Effect Size (3)	Between	1% | 3% | 5%
X Generation (2)	Between	Dichotomous | Continuous
W Generation (2)	Between	Dichotomous | Continuous
Analysis Model (6)	Within	7 | 8 | 14 | 15 | 58 | 59

Note. The number in the parentheses after each factor indicates the number of levels for that condition. Models 58 and 59 were only included when W generation was dichotomous.

		9 (Between: Sample Size) x 6 (Between: Effect Size) x 2 (Between: Normal or Dichotomous X) x 2(Between: Normal or Dichotomous W) x 6 (Within: Analysis Model) factorial design. The design is incomplete because Model 58 and 59 were only used as generating and/or analysis models when W was dichotomous. Each sample was analyzed with all four (continuous W) or six (dichotomous W) analysis models aligned with the generating models defined in Table 1. Effect size on the interaction term and sample size were varied. We generated data with continuous and dichotomous X and W variables. The specifics of these conditions are described in the main manuscript. We used GAUSS 21 on a Windows server for data generation,	but does not account for interactions when testing main effects. To test Hypothesis 1, we will only include correctly specified and over-specified models. The multilevel logistic regression will have 6 main effects: model specification (over vs. correct), generating model (dummy coded), sample size (sequentially coded), effect size (sequentially coded), type of X (continuous vs. dichotomous), and number of moderated paths in the analysis model (sequentially coded). We fit two separate models: one for continuous W and one for dichotomous W, since we had an incomplete design where Model 58 and 59 were only used as generating and analysis models when W was dichotomous. Hypothesis 1b will be tested using a		of the results are significant and in the opposite direction as hypothesized, we will interpret these results appropriately and explore the underlying cause of these unintuitive findings. We will explore whether there are factors that interact with over vs. correct specification to determine power. Hypothesis 1b will be supported would be supported by a significant effect of number of moderated paths in the analysis model (tested using a Wald test for the set of coefficients). We hypothesize that more moderated paths would lead to lower power in overspecified model, and so we hypothesize that all the coefficients will be negative. We would expect to see this in both the model with continuous W and dichotomous W. However, if the test	

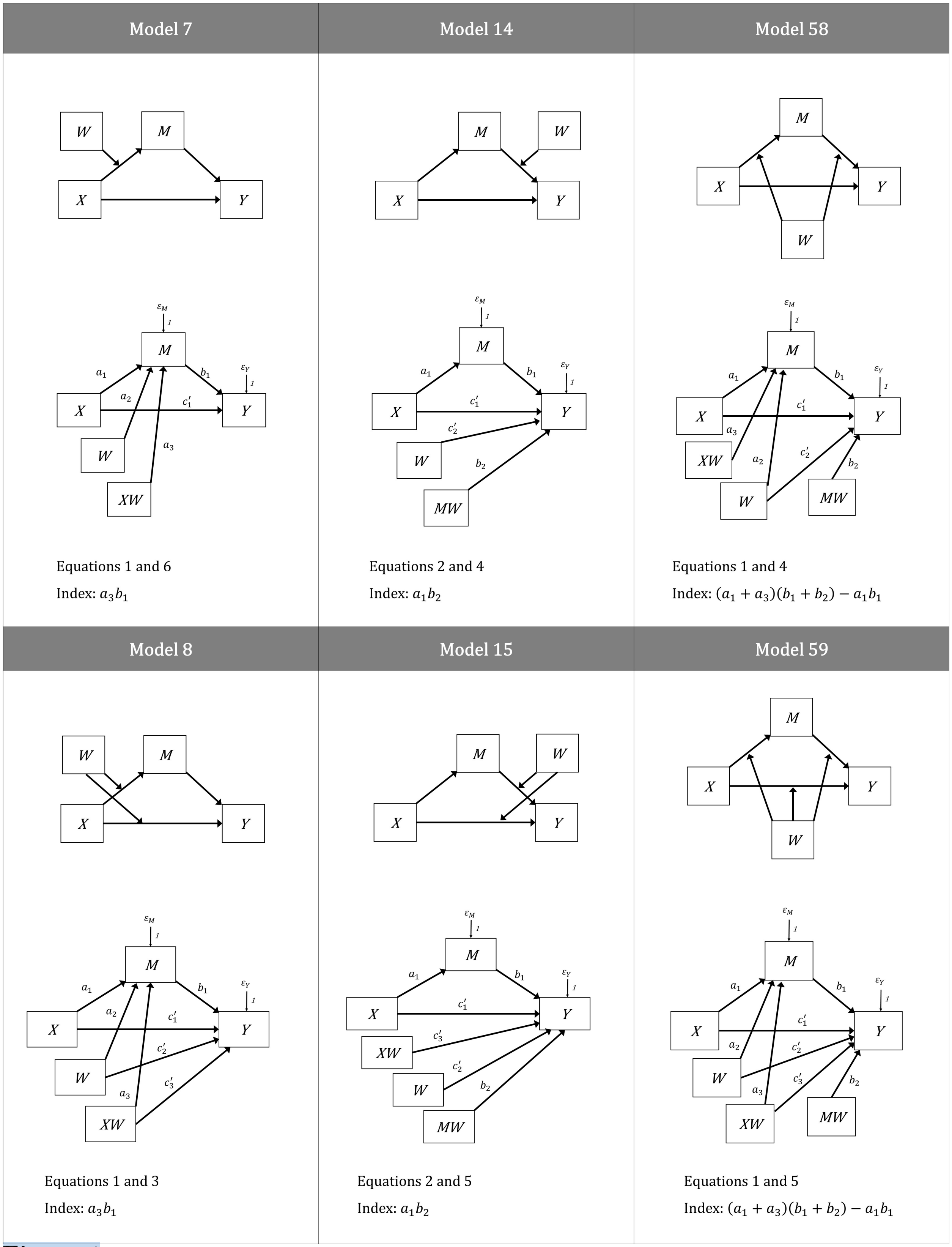

Figure 1

Moderated mediation conceptual diagrams (top diagram) and statistical diagrams (bottom diagram). Equations and indexes of moderated mediation (IMM) are also referenced. IMM for Models 58 and 59 is only defined when the moderator is dichotomous.

		generating 5000 samples of data in each condition. Based on quantiles from the systematic review, we used sample sizes of 100, 150, 200, 250, 300, 400, 500, 750, and 1000.	multilevel logistic regression model including all 6 main effects and an interaction between model specification (with overspecified models coded as zero) and number of moderated paths (sequentially coded).		is non-significant in both models, we would conclude that number of moderated paths does not impact power. If the test is significant and in the hypothesized direction for just one model, we would interpret this as partial support for Hypothesis 1b. If the test is significant and in the hypothesized direction for both models, we will interpret this as complete support for Hypothesis 1b. If any of the results are significant and in the opposite direction as hypothesized, we will interpret these results appropriately and explore the underlying cause of these unintuitive findings. We will explore whether there are other factors that interact with number of moderated paths to determine power.	
Related to RQ1, we examined these factors affecting statistical power for	We treated these models as exploratory, and do not have specific		In our exploratory analysis we will examine a model which includes all		Hypothesis 1.1a will be tested in the same manner as Hypothesis 1a but	

Appendix A

Systematic Review

We performed a large-scale systematic literature review to better understand current practices in moderated mediation analyses. We aimed to explore two questions: 1) Which moderated mediation models are most commonly used? and 2) What sample sizes are typical for moderated mediation analyses? Papers were chosen to be included in the systematic review through a search on WebofScience for papers published in the year 2018 including keywords “moderated mediation," “mediated moderation," and “conditional process analysis." We identified and coded 411 unique moderated mediation analyses. From this, we found that nine models were used most commonly (at least 10 examples of each were found in this review). From those nine models, six models were chosen as the focus for this registered report, and in total those six models accounted for 86% of published models from the systematic review. We limited the scope of this registered report to only include models with one moderator variable, which excludes Models 9 and 21, and Model 74 was excluded because the predictor variable is also used as the moderator variable. Table A1 shows the percentage of articles from this systematic review that used each particular model. Sample size results are summarized by model in the next row and in and Figure A1. In general, there does not seem to be an obvious pattern where researchers use larger sample sizes for more complex models. The highest median sample size among the models used for the main study was Model 59 where all three paths are moderated, but in the primary literature review the median sample size for Model 74 (where X moderates the path from M to Y) was higher. These results directly inform the parameters chosen for the proposed simulation study. This was done in an effort to make the results as useful and applicable as possible for researchers using moderated mediation. The data from the systematic review are available at https://osf.io/m5f3h. All of the papers included in this systematic review plus additional papers from more recent years are available in a searchable database: https://www.jlfossum.com/moderated-mediation-article-database.

the index of moderated mediation for under-specified models.	hypotheses for statistical power in these models. We will examine the effect of under-specification compared to correct specification (H1.1a), the number of moderated paths in the model (H1.1b).		possible interactions, separating effect size into two predictors (magnitude and sign). We also plan to make tables and figures to provide the raw rejection rates in each condition, and visualize patterns across conditions.		comparing underspecified to correctly specified models. We have no directional hypotheses for this test. We believe that in some cases underspecification may increase power and in other cases decrease power. We will use the exploratory interactions to identify such cases. Hypothesis 1.1b will be tested in the same manner but comparing underspecified to correctly specified models. We have no directional hypotheses for this test. We believe that in some cases more paths may increase power and in other cases decrease power. We will use the exploratory interactions to identify such cases.	
Research Question 2 examined what factors impact type I error rate of the index of moderated mediation.	For complete misspecification, the index of moderated mediation of the data analysis model should be 0 given		To test Hypothesis 2, we only included completely misspecified analysis models. To examine Hypothesis 2a, we		If the majority of conditions show elevated type I error rates, Hypothesis 2a will be supported.	

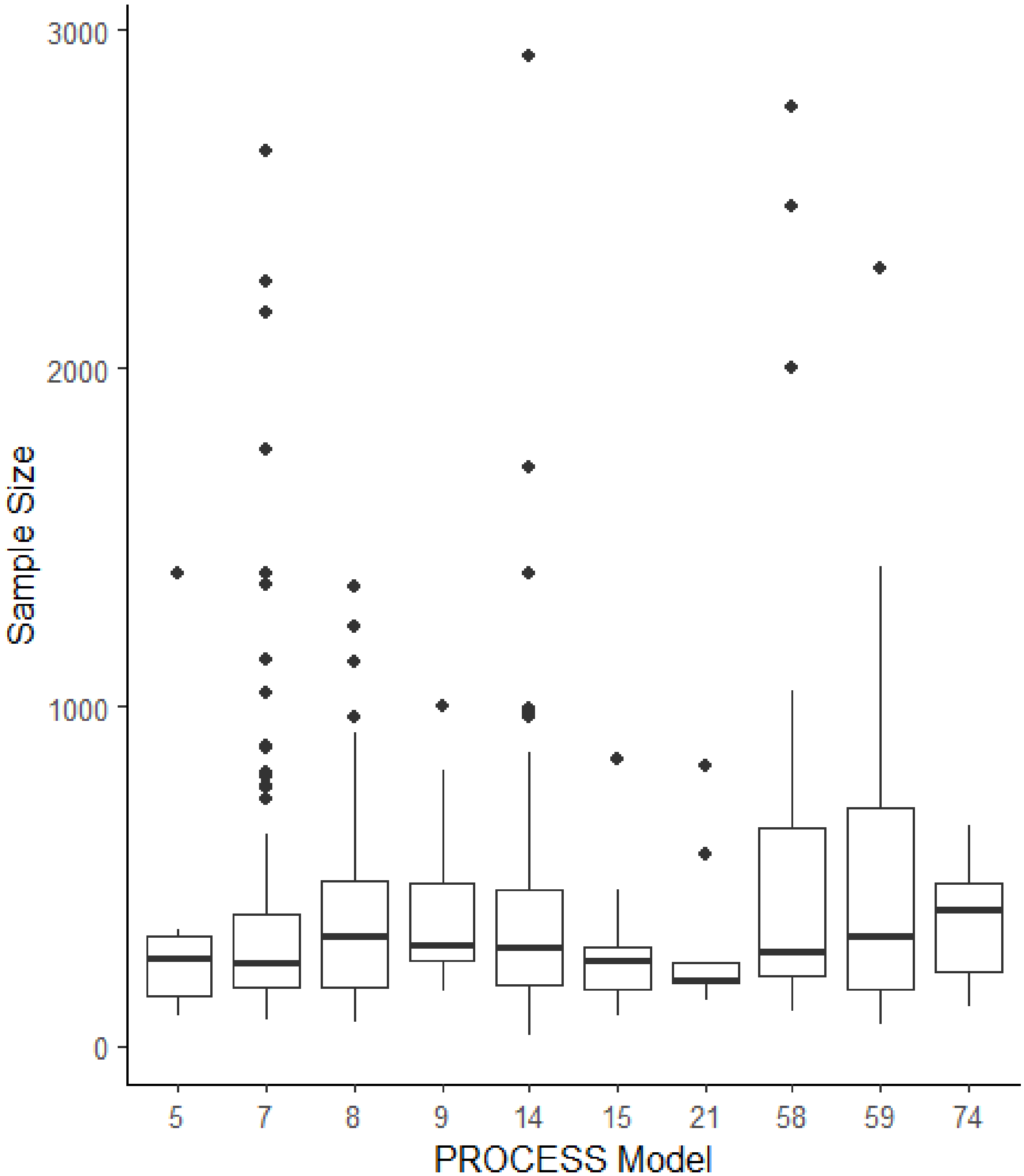

Figure A1

Box plots displaying the range of sample sizes reported in the articles included in the systematic review, separated out by PROCESS model. For clarity, outliers above 3,000 were excluded.

	the DGP. So if the index of moderated mediation is statistically significant, this would be a type I error. We treated these models as exploratory, though we hypothesize that type I error rate would be too high in completely misspecified models (H2a) and type I error rate will increase as the number of incorrectly moderated paths increases as well (H2b).		will evaluate how many conditions have type I errors above the criteria set by Bradley et al. (2008) and Serlin et al. (2000). In addition, we will explore what factors predict type I error rate using a multilevel logistic regression with 5 main effects and all possible interaction: generating model (dummy coded), sample size (sequentially coded), effect size (sequentially coded), type of X (continuous vs. dichotomous), and number of moderated paths in the analysis model (sequentially coded). We fit two separate models: one for continuous W and one for dichotomous W. Hypothesis 2b will be tested using a multilevel logistic regression with only the main effects above, and a Wald test for number of moderated paths.		Hypothesis 2b would be supported if the test is significant (in both models) and the coefficients are in the hypothesized direction. If the effect is only significant in one model, or some paths are in the opposite direction, we will interpret these results accordingly. We will also use the model with the interactions to explore factors which might interact with number of paths to predict type I error.	

Table A1

Systematic Review Models and Sample Sizes

Model	7	8	9	14	15	21	58	59	74
Use Frequency	31%	13%	3%	18%	3%	2%	6%	14%	2%
Median Sample Size	261	331	317	288	255	199	276	363	430

Note. Each column represents a PROCESS Model number.

Guidance Notes

- **Question**: articulate each research question being addressed in one sentence.
- **Hypothesis**: where applicable, a prediction arising from the research question, stated in terms of specific variables rather than concepts. Where the testability of one or more hypotheses depends on the verification of auxiliary assumptions (such as positive controls, tests of intervention fidelity, manipulation checks, or any other quality checks), any tests of such assumptions should be listed as hypotheses. Stage 1 proposals that do not seek to test hypotheses can ignore or delete this column.
- **Sampling plan**: For proposals using inferential statistics, the details of the statistical sampling plan for the specific hypothesis (e.g power analysis, Bayes Factor Design Analysis, ROPE etc). For proposals that do not use inferential statistics, include a description and justification of the sample size.
- **Analysis plan**: For hypothesis-driven studies, the specific test(s) that will confirm or disconfirm the hypothesis. For non-hypothesis-driven studies, the test(s) that will answer the research question.
- **Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis**: For hypothesis-driven studies that employ inferential statistics, an explanation of how the authors determined a relevant effect size for statistical power analysis, equivalence testing, Bayes factors, or other approach.
- **Interpretation given different outcomes**: A prospective interpretation of different potential outcomes, making clear which outcomes would confirm or disconfirm the hypothesis.
- **Theory that could be shown wrong by the outcomes**: Where the proposal is testing a theory, make clear what theory could be shown to be wrong, incomplete, or otherwise inadequate by the outcomes of the research.

Appendix B

Sample Tables and Figures

Table B1

Hypotheses 1a, 1b, and 2a

	Analysis Model					
DGP	Model 7	Model 8	Model 14	Model 15	Model 58	Model 59
7						
8						
14						
15						
58						
59						

Note. Table for the main manuscript showing statistical power (proportion of correctly rejected hypothesis tests for the index of moderated mediation) from the simulation. The columns represent the data analysis model, and the rows represent the DGP. All power is for continuous moderators and continuous X with a medium interaction effect size at sample size 300. Additional tables showing other conditions (all sample sizes and effect sizes in each table, separated by dichotomous moderators, and dichotomous X for total = 4 tables) will be provided in the supplemental material.

Table B2

Hypotheses 1c, 2b, and 3b

	Analysis Model					
DGP	Model 7	Model 8	Model 14	Model 15	Model 58	Model 59
7						
8						
14						
15						
58						
59						

Note. Table for the main manuscript showing raw parameter bias from the simulation. The columns represent the data analysis model, and the rows represent the DGP. All parameter bias is for continuous moderators and continuous X with a medium interaction effect size at sample size 300. Additional tables showing other conditions (all sample sizes in each table, separated by dichotomous moderators, and dichotomous X for total = 4 tables) will be provided in the supplemental material.

Table B3

Hypothesis 3a

		Small Effect Size				*Medium Effect Size*				*Large Effect Size*			
Sample Size	*DGP*	7	8	14	15	7	8	14	15	7	8	14	15
100	7												
	8												
	14												
	15												
150	7												
	8												
	14												
	15												
200	7												
	8												
	14												
	15												
250	7												
	8												
	14												
	15												
300	7												
	8												
	14												

	15
400	7
	8
	14
	15
500	7
	8
	14
	15
750	7
	8
	14
	15
1,000	7
	8
	14
	15

Note. Type I error rate by sample size. The columns represent the data analysis model, and the DGP is listed in the row. The three effect sizes are shown side-by-side. Type I error rates in the table are shown only for continuous X. One additional table (total = 2 tables) with dichotomous X will be provided in the supplemental material. Type I error rates outside criteria set by Bradley (1978) are in **bold**.

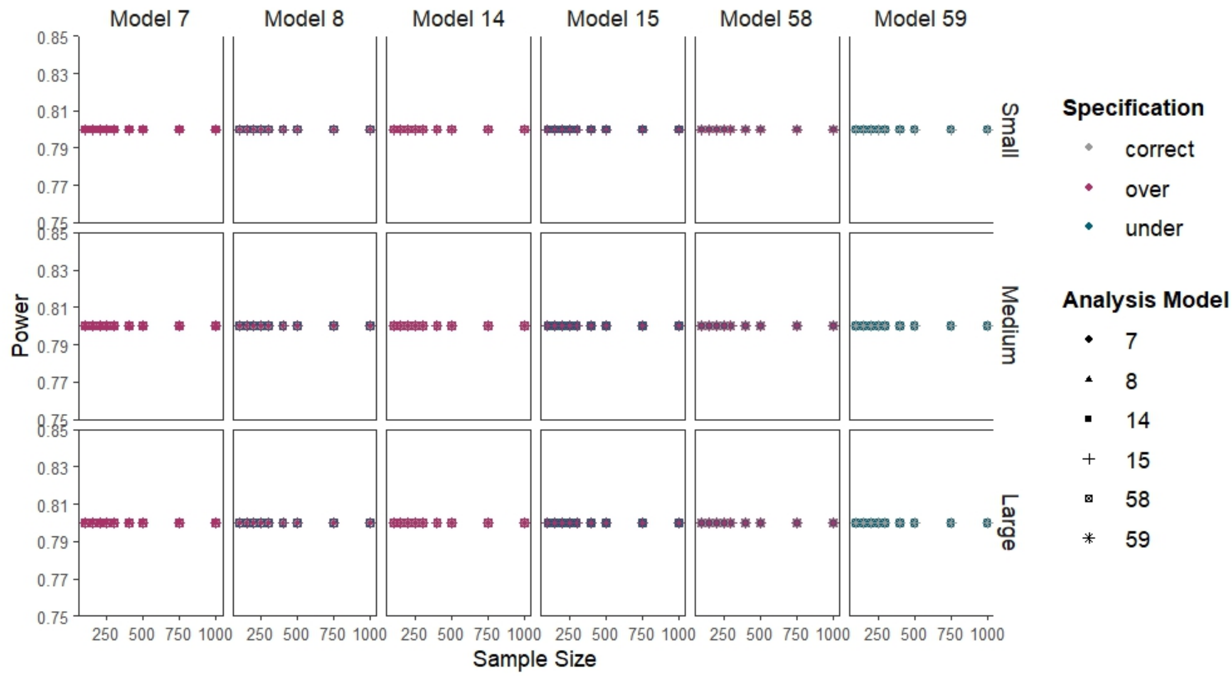

Figure B1

Example figure for statistical power. Power is arbitrarily set at .8 for each condition, but we are expecting power curves to be in the actual results. Additional figures showing dichotomous X and dichotomous W combinations (total = 4 figures) will be provided in the supplemental material.

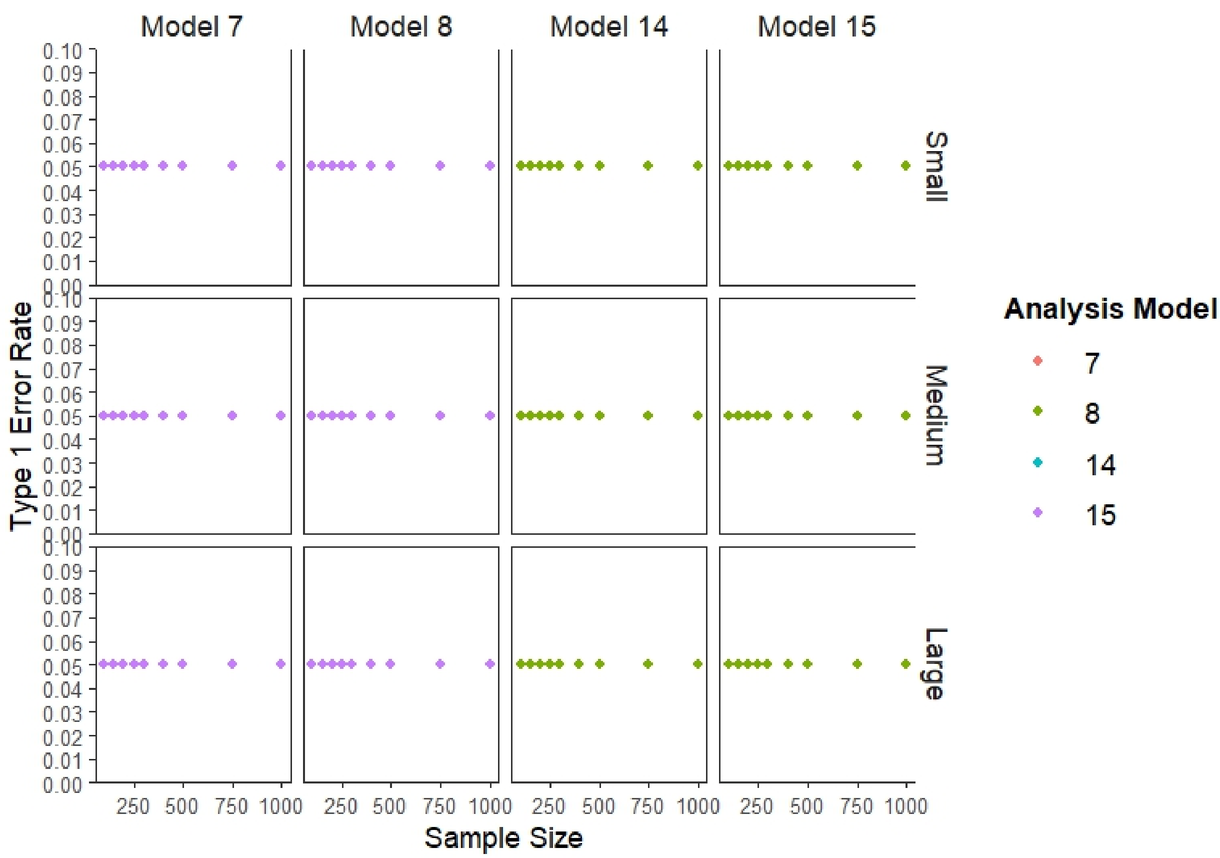

Figure B2

Example figure for Type I Error Rate. Type I error rate is arbitrarily set at .05 for each condition, but we are expecting actual results to vary. Additional figures showing dichotomous X (total = 2 figures) will be provided in the supplemental material.