UNIVERSITY OF CALIFORNIA

Los Angeles

Impacts of Model Specification on Statistical Power and Type I Error Rate in Moderated Mediation Analysis

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Psychology

by

Jessica Louise Fossum

2023

ABSTRACT OF THE DISSERTATION

Impacts of Model Specification on Statistical Power and Type I Error Rate in Moderated Mediation Analysis

by

Jessica Louise Fossum
Doctor of Philosophy in Psychology
University of California, Los Angeles, 2023
Professor Amanda K. Montoya, Chair

Moderated mediation models are used commonly in psychological research and other academic fields to model how and when effects occur. Researchers must choose which paths from the mediation model are moderated when specifying this type of model. This dissertation examines how model specification impacts statistical power and type I error rate for the index of moderated mediation. In a meta-analytic review, we found that six model specifications account for 85% of published moderated mediation analyses, so this dissertation focuses on those six models. When considering power and type I error rate, two attributes matter: the data analysis model, and the data generating process (DGP). In reference to the DGP, the data analysis model can either be correctly specified, over-specified, under-specified, or completely misspecified. A Monte Carlo simulation study was run to examine the impacts of model specification on power and type I error rate, and results were analyzed using multi-level logistic regression along with figures and tables. Over-specified models had lower statistical power to detect a significant index of moderated mediation compared to correctly specified models. Under-specified models had slightly higher power when modera-

tion on the direct effect was omitted, but otherwise, under-specified models had much lower power than correctly specified models. Parameter bias was also unacceptably high for most under-specified models. Completely misspecified models generally still had acceptable type I error rates, with a notable exception of inflated type I error rates where moderation was omitted from the direct effect. Overall, while many published moderated mediation models may not have large enough sample sizes for adequate statistical power, over-specifying or under-specifying models can lead to lower statistical power as well, while complete model misspecification risks an inflated type I error rate.

This dissertation of Jessica Louise Fossum is approved.

Samantha F. Anderson

Craig Kyle Enders

Katherine H. Karlsgodt

Amanda K. Montoya, Committee Chair

University of California, Los Angeles

2023

DEDICATION

This dissertation is dedicated to all the places I otherwise would not have seen.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

Data collection for the meta-analytic review in Chapter 2 was partially supported by the UCLA Graduate Research Mentorship Program. Chapter 3 is a version of Fossum, Montoya, and Anderson (in preparation).

Thank you first and foremost to my advisor, Amanda, for supporting me and challenging me to grow in my thinking throughout this entire process. Thank you to my committee members, Craig, Katie, and Samantha, for your support and the great conversations about research ideas. Thank you Tom for your mentorship this whole way. Community matters, and I'm so grateful for the quantitative area at UCLA. Tristan, I could not have asked for a better lab mate. Thank you for being such a great example and for making me laugh along the way. Valerie, thank you for always being there for me, from day trips to LA to layovers in Seattle. To the friends in other major areas, you made graduate school the kind of fun, collaborative environment I had always wanted.

Thank you to my family and friends, for being a constant source of encouragement. Mom, thank you for being the one who always understands me. Dad, thank you for always reminding me you're proud of me. Grandma, thank you for being an example of living a life of faith. You've all let me go on this adventure of graduate school, and I'm so grateful. To those that are gone too soon, thank you for the impact you made in my life. To all the friends in many states who have seen me through this process, know the FaceTime calls, trips to visit, encouraging messages, and space to process all meant the world to me.

Thank you to my advisors for being open to letting me explore teaching, and thank you Craig for letting me practice. I'm not sure I would have trusted such a novice TA to try teaching a lab section during those early days on Zoom, but I'm grateful for that experience. To the faculty at Pepperdine, thank you for the opportunity to teach a statistics course. I'll always remember how surreal it was to make materials for my first class from the library with a view of the Pacific Ocean. To everyone at SPU, thank you for investing in me so much as an undergraduate student, and for welcoming me back. I'm so grateful.

BIOGRAPHICAL SKETCH

EDUCATION

M.A. Psychology: *University of California, Los Angeles* 2020

B.A. Psychology: *Seattle Pacific University* 2019

B.S. Applied Mathematics: *Seattle Pacific University* 2019

ACADEMIC EMPLOYMENT

Instructor of Research Psychology, *Seattle Pacific University* 2022-

Adjunct Instructor, *Pepperdine University* 2022

PUBLICATIONS

Fossum, J. L. & Montoya, A. K. (2023). When to Use a Different Test for Power Analysis and Data Analysis for Mediation. *Advances in Methods and Practices in Psychology.*

Bilder, R., Widaman, K., Bauer, R., Drane, D., Loring, D., Umfleet, L., Reise, S., Vannier, L. C., Wahlstrom, D., **Fossum, J. L.**, Wong, E., Enriquez, K., Whelan, F., & Shih, S. (2022). Construct Identification in the Neuropsychological Battery: What Are We Measuring? *Neuropsychology.*

Fossum, J. L. & Montoya, A. K. (2022). Heterogeneity. In B. B. Frey (Ed.), *Encyclopedia of Research Design* (2nd ed.). Thousand Oaks, CA: SAGE.

Ramos, G., Delgadillo, D., **Fossum, J.**, Montoya, A. K., Thamrin, H., Rapp, A., Escovar, E., & Chavira, D. (2021). Discrimination and Internalizing Disorders in Rural Latinx Youth: An Ecological Model of Etiology. *Children and Youth Services Review*, *130*.

Montoya, A. K., Krenzer, W. D. L. & **Fossum, J. L.** (2021). Opening the Door to Registered Reports: Census of Journals Publishing Registered Reports. *Collabra: Psychology*, *7*(1).

SELECTED PRESENTATIONS

Fossum, J. L., Lafit, G., & Montoya, A. K. (2022, July). Power analysis methods for mediation and moderated mediation models. Presented at the International Psychometric Society Annual Meeting, Bologna, Italy.

Yang, N., **Fossum, J. L.**, & Montoya, A. K. (2022, May). *Extending Mediation Analysis to Within-subjects Data With Dichotomous Outcomes.* Presented at UCLA Undergraduate Research Week, Los Angeles, CA.

Fossum, J. L. & Montoya, A. K. (2021, February). Comparison of power analysis methods for mediation. Presented at the Society for Personality and Social Psychology Annual Meeting, Virtual.

Impacts of Model Specification on Statistical Power and Type I Error Rate in Moderated Mediation Analysis

Psychological researchers are often interested in explaining when and how effects occur, such as how discrimination affects internalizing feelings of racism through peer victimization, and whether that relationship is different based on age (Ramos et al., 2021). Mediation analysis provides a way of examining effects that occur through a mediator variable, such as peer victimization, and moderation analysis provides a way of examining effects that differ based on the level of the moderator variable, such as age (Baron & Kenny, 1986). These can be used together in the same model in a moderated mediation analysis, exploring both how and when effects occur. In moderated mediation, variables in the mediation model are allowed to interact with a moderator variable. Moderated mediation analyses are used widely, with WebofScience counting 3,815 published articles in 2022, 3,449 in 2021, and 2,602 in 2020. There are many decisions researchers must make when planning a moderated mediation analysis, all of which should be made before analyzing the data (Gelman & Loken, 2013).

Two important decisions researchers must make before doing a moderated mediation analysis are how large of a sample to collect and which paths are moderated. These decisions may impact statistical power, which is the likelihood of detecting an effect if there is an effect in the population (Neyman & Pearson, 1993). Planning for adequate statistical power is an important part of study design for reliably finding effects, including moderated mediation. Many factors have been shown in simulation studies to affect statistical power, including effect size and sample size (J. Cohen, 1988), dichotomous vs. continuous predictor variables (McClelland & Judd, 1993), and model specification (Dupont & Plummer, 1998). Since statistical power is dependent on sample size, the goal in sample size planning is to find the optimal balance between having as many participants as possible yet does not waste research resources unnecessarily. Power for mediation analysis specifically has been studied in the recent literature, with the conclusion that these analyses tend to be underpowered in psychology research (Fairchild & McDaniel, 2017; Fritz & MacKinnon, 2007; Götz, O'Boyle, Gonzalez-Mulé, Banks, & Bollmann, 2021).

Work by Götz et al. (2021) investigated the potential bias towards publishing only statistically significant mediation analyses by examining 2,569 tests of mediation in published psychology literature. They recorded how many confidence intervals of the mediated effect lie in the *Goldilocks Zone*, where just the right point estimate and confidence interval width come together to form a significant result questionably close to zero. While finding a significant effect that truly exists that close to zero can and does happen, too many results in this zone can be indicative of questionable research practices (QRPs). According to John, Loewenstein, and Prelec (2012): "QRPs are the steroids of scientific competition, artificially enhancing performance and producing a kind of arms race in which researchers who strictly play by the rules are at a competitive disadvantage." Possible examples include dropping hypotheses due to a lack of support or even running many models and selectively reporting the results, then aligning hypotheses after the fact accordingly (Götz et al., 2021). Reporting only the statistically significant effects is also a QRP (Fiedler & Schwarz, 2016), though often it is done unintentionally (Gelman & Loken, 2013). These issues are not limited to mediation analysis but are also relevant for moderation analysis. In addition, effect sizes for interactions have been found to be very small in published literature (Aguinis, Beaty, Boik, & Pierce, 2005; Vize, Sharpe, Miller, Lynam, & Soto, 2022; Vize et al., 2023), compounding the problem and potentially making the more complicated test of moderated mediation even more underpowered compared to mediation.

Underpowered mediation and moderated mediation analyses are particularly concerning because most published mediation models are significant, even without adequate statistical power (Schoemann, Boulton, & Short, 2017). This is concerning in terms of type I errors, which happen when a test has a significant result but there is truly no effect in the population. It is suspected that published mediation analyses, and by extension moderated mediation analyses, have higher than average type I error rates in the published literature, given this large proportion of positive results found by Götz et al. (2021). The 161 tests of moderated mediation included in the Götz et al. (2021) study had a relative proximity of 74% compared to 68% for the 1,372 tests of mediation, meaning the confidence intervals around the indices of moderated mediation tended to be very close to zero. Very little additional research beyond the work done by Götz et al. (2021) has been done to extend these findings to moderated mediation analyses, which are more complex than mediation. The goal of this research is to explore how factors including sample size and model specification affect statistical power, type I error rate, and parameter estimation accuracy in moderated mediation models.

Chapter 1 introduces mediation, moderation, and moderated mediation, with specific attention paid to sample size planning and factors affecting statistical power and type I error rate. Chapter 2 describes a meta-analytic review exploring current practices in published moderated mediation literature for sample sizes used and how those current practices inform the following simulation study. Chapter 3 presents an applied example of a moderated mediation model using real data, showing how statistical power can be affected by over-specifying a model. Chapter 4 then describes the method and results from this simulation study looking at how various factors impact statistical power and type I error rate in six commonly used moderated mediation models. This dissertation concludes with a general discussion in Chapter 5 including overall conclusions about statistical power and type I error for moderated mediation, implications for researchers, limitations, and future directions.

1 Introduction

This introduction chapter explores sample size planning and factors affecting power and type I error rate in mediation and moderation analyses. I begin with an introduction of mediation analysis, then an introduction of moderation analysis, and finally an introduction of moderated mediation analysis. All three include equations, path diagrams, estimation, and inference. I also discuss factors that affect power, type I error rate, and discuss sample size planning. I then conclude the chapter by discussing the emerging literature related to sample size planning strategies for moderated mediation models.

1.1 Mediation Analysis

Mediation analysis uses a system of equations to model how a focal predictor variable (X) affects an outcome variable (Y) through a mediator variable (M). Mediation is inherently causal, so researchers must pick the order of variables in their analysis under the assumption that there are no confounding variables not accounted for in the model (Pearl, 2001). For the purposes of this study, the causal order of the variables in a model is assumed to be correctly specified and all additional assumptions are met (Rubin, 2005; Pearl, 1995).

There are three regression equations used to define a mediation model. Below, Equation 1.1 shows X predicting Y without the mediator present, then Equation 1.2 and Equation 1.3 are used in mediation to predict M and Y, respectively.

$$Y_i = c_0 + cX_i + \varepsilon_{Y_i} \tag{1.1}$$

$$M_i = a_0 + aX_i + \varepsilon_{M_i} \tag{1.2}$$

$$Y_i = c_0' + c'X_i + bM_i + \varepsilon_{Y_i'} \tag{1.3}$$

In each equation, X_i, M_i, and Y_i represent an individual i's score on the focal predictor variable, mediator, and the outcome, respectively. Equation 1.1 shows the total effect (c), with the regression coefficient for X predicting Y, and the intercept is c_0' and residual error $\varepsilon_{Y_i'}$. This is represented visually at the top of Figure 1.1. Equations 1.2 and 1.3 predict M and Y, respectively. In Equation 1.2, the regression coefficient for X predicting M is a, with intercept a_0 and residual error ε_{M_i}. In Equation 1.3, the regression coefficient for X predicting Y controlling for M is c' (called the direct effect), and the regression coefficient for M predicting Y controlling for X is b, with intercept c_0' and residual error $\varepsilon_{Y_i'}$ (see bottom of Figure 1.1).

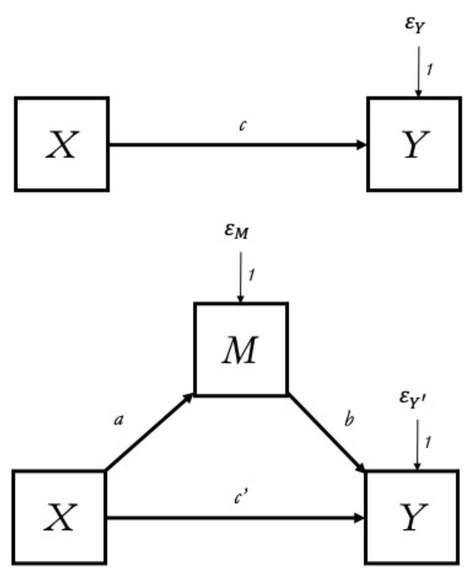

Figure 1.1: Path diagram for Simple Mediation.

The total effect, direct effect, and indirect effect of X on Y through M can be estimated using ordinary least squares regression. The indirect effect is what is tested in mediation analysis, since it describes the effect of X on Y through the mediator. The estimate for X predicting M from Equation 1.2 ($\hat{a}$) and M predicting Y controlling for X ($\hat{b}$) are multiplied together to form the indirect effect ($\widehat{ab}$). The indirect effect is how much Y changes as a result of the X variable manipulation through M, assuming the causal order of variables is correctly specified and that there are no omitted confounders (Pearl, 1995). Since the indirect effect is the product of two normally distributed coefficients, the sampling distribution of the indirect effect is not normally distributed (Lomnicki, 1967).

1.1.1 Inference for Mediation

While all the individual paths in mediation analysis are regression coefficients, the indirect effect is not so simple. The indirect effect is the product of two regression coefficients, so several methods have been created or adapted for making inferences about the indirect effect. Some methods, such as bootstrapping, produce confidence intervals around the indirect effect.

There are a few commonly used methods for testing the statistical significance of the indirect effect. The preferred method is the percentile bootstrap confidence interval (Preacher & Hayes, 2004, 2008; Shrout & Bolger, 2002), which uses bootstrapping as a resampling method to generate a distribution of indirect effect estimates (Bollen & Stine, 1990). For this method, bootstrapping is done thousands of times, each time recording the indirect effect from the resampled data. Constructing a confidence interval then involves arranging the bootstrapped indirect effect estimates in ascending order. Using the percentile bootstrap method, the limits of the confidence interval are established based on the chosen significance level, denoted as α. For example, with $\alpha = .05$, the values falling below the 2.5^{th} percentile and above the 97.5^{th} percentile are excluded from the interval. If the interval does not encompass zero, it indicates a statistically significant indirect effect at $\alpha = .05$.

One of the main reasons the percentile bootstrap confidence interval is the preferred inferential method for mediation analysis is that it has been shown to perform better than other methods which assume the indirect effect is normally distributed (Williams & MacKinnon, 2008; MacKinnon, Lockwood, & Williams, 2004; Biesanz, Falk, & Savalei, 2010). The bias-corrected bootstrap confidence interval (Stine, 1989) was once a popular method for inference on the indirect effect but has fallen from favor due to its inflated type I error rate (Chen & Fritz, 2021; Tibbe & Montoya, 2022).

Historically, the indirect effect was tested using the causal steps approach (Baron & Kenny, 1986). For this method, the null hypotheses for three effects must be rejected. First, there must be a statistically significant effect of X on Y (c). Next, there must be a statistically significant effect of X on M (a). Finally, there must be a statistically significant effect of M on Y, controlling for X (b). If, and only if, all three significant effects are found, then mediation can be claimed. This approach has been criticized because it requires the total effect (X on Y) to be statistically significant to claim mediation (Kenny, Kashy, & Bolger, 1998). This is not a valid criterion because it is still possible for M to be causally between X and Y even if X and Y are not associated (Hayes, 2018b). In other words, the size of the c path does not constrain either the a or b path involved in the indirect effect. Addressing this criticism, the joint significance test drops the requirement for a significant total effect to claim mediation and also does not require the direct effect to be significant. Using the joint significance test, only two separate significant paths are required: the effect of X on M and the effect of M on Y (Kenny et al., 1998). This avoids the risk of missing a mediated effect due to the significant total effect requirement.

Other methods for conducting inference on the indirect effect exist, including the Sobel test, Monte Carlo confidence interval, and two bootstrapping methods. The Sobel test uses equations provided in Sobel (1982) to calculate standard errors but relies heavily on the assumption that the indirect effect is normally distributed. Other distribution of the product methods exist as well (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). The Monte Carlo confidence interval uses sample estimates for a and b to create separate sampling distributions for each based on normal distributions, then combine them to create an estimated sampling distribution for $\widehat{ab}$ where the 2.5^{th} and 97.5^{th} percentiles (when $\alpha =$ .05) form a CI (Selig & Preacher, 2008). Like the percentile bootstrap confidence interval, if the interval does not contain zero, it indicates a statistically significant indirect effect at $\alpha = .05$. Though for a time the bias-corrected bootstrap confidence interval was the default method in the commonly used tool for mediation analysis, the PROCESS macro (Hayes, 2022), it now uses the percentile bootstrap confidence interval. In a small review of published mediation studies, PROCESS was used in over 77% of the regression-based analyses (Rasoolimanesh, Wang, Roldán, & Kunasekaran, 2021). The percentile bootstrap confidence interval is used in this dissertation simulation study because of its popularity and because it has been shown to perform well in simulation studies.

1.1.2 Sample Size Planning for Mediation

There are several tools available for power analysis and sample size planning for mediation. Statistical power for mediation depends on the desired α level, an effect size measure from either the standardized a path and b path (and possibly the direct effect c' path depending on the method of inference) or correlations among all three variables, inferential method, and the sample size. The joint significance test offers a closed-form solution (O'Rourke & MacKinnon, 2014), enabling straightforward power calculation and enhancing computational efficiency. One commonly used reference for sample size planning for mediation analysis is a table of required sample sizes to achieve 80% power for several different effect size combinations (Fritz & MacKinnon, 2007). This table can be a useful guide, but only offers recommendations for four effect sizes on both the a and b paths (J. Cohen, 1969). It also only offers sample size estimates for 80% power, with no indication of how much the required sample sizes might change at differing levels of power. Psychological research, and replication studies in particular, are moving towards requiring higher power (van Zwet & Goodman, 2022), highlighting the need for methods of determining required sample sizes beyond a published table. While Fritz and MacKinnon (2007) has been cited over 4,000 times, there are now tools that can aid in calculating statistical power or sample size calculations that allow for more nuanced effect size estimates (e.g., Aberson, 2019a; Zhang & Yuan, 2018; Kenny, 2017, etc.).

There are several different tools available for power analysis for mediation, and they differ in terms of what inferential method is used and how the tool is accessed. One computationally-efficient tool that uses the Monte Carlo confidence interval is MCpowerMed (Schoemann et al., 2017). WebPower also has options to use the Monte Carlo method and also includes options for if the assumption of normality is expected to be violated, but defaults to using the Sobel test (Zhang & Yuan, 2018). Other tools make use of different inferential methods: MedPower (Kenny, 2017) uses the joint significance test; The package 'pwr2ppl' (Aberson, 2019a) uses either the Sobel test or the joint significance test; and the package bmem (Zhang & Yuan, 2018) has the option for either the Sobel test or the recommended percentile bootstrap confidence interval. Tools using bootstrapping are the least computationally-efficient resampling method because of their repeated resampling of the data instead of sampling from a distribution like the Monte Carlo confidence interval. The fact that different inferential methods are used by different tools is something for researchers to be aware of as different inferential methods may yield different statistical power (Fritz & MacKinnon, 2007). The sample size estimates can be very similar, especially among the joint significance test, Monte Carlo confidence interval, and the percentile bootstrap confidence interval, but the other methods are not always interchangeable (Fossum & Montoya, 2023). Differences in results between inferential methods are exacerbated in cases where the assumptions of normality or heteroskedasticity of the residuals are violated (Biesanz et al., 2010; Fossum & Montoya, 2023).

These tools are accessible either through web applications or as R packages and differ in their output capabilities. If researchers feel most comfortable with a web application tool, MCpowerMed (Schoemann et al., 2017), MedPower (Kenny, 2017), or WebPower (Zhang & Yuan, 2018) are the best options. Both tools also have the option to compute a required sample size given statistical power. WebPower (Zhang & Yuan, 2018) is also available in an R package, which is the only way to access 'pwr2ppl' (Aberson, 2019a) and bmem (Zhang & Yuan, 2018). Neither 'pwr2ppl' (Aberson, 2019a) nor bmem (Zhang & Yuan, 2018) can output a required sample size given statistical power. Compared to tools for more complicated analyses such as moderated mediation analysis, the sophistication of some of these tools to output a sample size for a desired level of power is very advantageous and can be incredibly time-saving compared to manually iterating through possible sample sizes. The only tools with this capability are WebPower (Zhang & Wang, 2013) and MedPower (Kenny, 2017). Since it is easier from the software perspective to output statistical power for a given sample size, the fact that these tools can estimate a sample size for a desired level of power makes them particularly useful, especially compared to the more limited options for sample size planning in more complex analyses.

1.2 Moderation Analysis

In a moderation analysis, the effect of predictor (X) on the outcome (Y) differs depending on the level of a moderator (W) (Baron & Kenny, 1986). For example, a university having a football team or not (W) could moderate the relationship between university ranking (X) and school spirit (Y). Students might reasonably have more school spirit if they go to a highly-ranked school, but this relationship could be much more salient if the university has a football team. Like simple mediation analysis, this is a simple moderation analysis because it only has one moderator variable, but including more than one moderator is possible.

In a typical linear regression analysis that is not moderated, the effect on the outcome Y of both the predictor variables X and W can be expressed using Equation 1.4. In Equation 1.4, each participant i's predicted score on the outcome variable based on their score on X_i and W_i is represented by Y_i. Including multiple predictors in the same model allows for interpreting the effect of one predictor on the outcome holding the other variable(s) constant. This predicted outcome $\hat{Y}_i$ is a linear function of participant i's score on the predictor variable X and moderator variable W. Each predicted score also has a residual error ($\varepsilon_{Y_{i'}}$), which is the difference between each individual i's predicted score and their actual score on the outcome variable. The intercept $b_{0'}$ is the estimated value of Y when all the predictors in the model are set to zero. The estimate for X predicting Y controlling for W is $b_{1'}$, and the estimate for W predicting Y, controlling for X is $b_{2'}$.

$$Y_i = b_{0'} + b_{1'}X_i + b_{2'}W_i + \varepsilon_{Y_{i'}} \tag{1.4}$$

This dissertation limits the discussion of moderation to only include linear interaction between the predictor variable X and the moderator variable W. Moderation in multiple regression can be tested by including the product of the two predictors in the linear regression model (Aguinis, 1995). In these models, the effect of X on Y is dependent, or *conditional*, on the value of W. The interaction is the product of X and W, and it can be added into the regression equation as a separate term as seen in Equation 1.5. It will have a separate coefficient, b_3, which represents the estimated amount by which the effect of X on Y differs between two hypothetical participants who are exactly one unit apart on W. The outcome for participant i in this equation including the interaction term is Y_i, and the error term is ε_{Y_i}. The intercept is b_0 is still the predicted score when X and W are zero, though its numerical value may differ from $b_{0'}$ in the previous equation. In Equation 1.5, $b_{1'}$ from Equation 1.4 is replaced by $b_1 + b_3W_i$. Then, applying the product rule, $Y_i = b_0 + (b_1 + b_3W_i)X_i + b_2W_i + \varepsilon_{Y_i}$ becomes Equation 1.5 (J. Cohen, Cohen, West, & Aiken, 2003). In Equation 1.5, b_1 is the effect of X on Y, holding W constant at 0. Similarly, b_2 is the effect of W on Y, holding X constant at 0.

$$Y_i = b_0 + b_1X_i + b_2W_i + b_3X_iW_i + \varepsilon_{Y_i} \tag{1.5}$$

The moderated effect in Equation 1.5 is the coefficient from the product term, b_3.

1.2.1 Inference for Moderation

The coefficient on the interaction b_3 can be tested for statistical significance in the same way any coefficient in multiple linear regression is, and all the coefficients are estimated using any multiple linear regression program (Hayes, 2018a). The coefficient on the interaction term, b_3, is used to test for moderation, determining how the effect of X on Y differs depending on the level of the moderator, W. These coefficients are seen on the right side of Figure 1.2, which is the statistical diagram for moderation. The path diagram for this model is on the left of Figure 1.2.

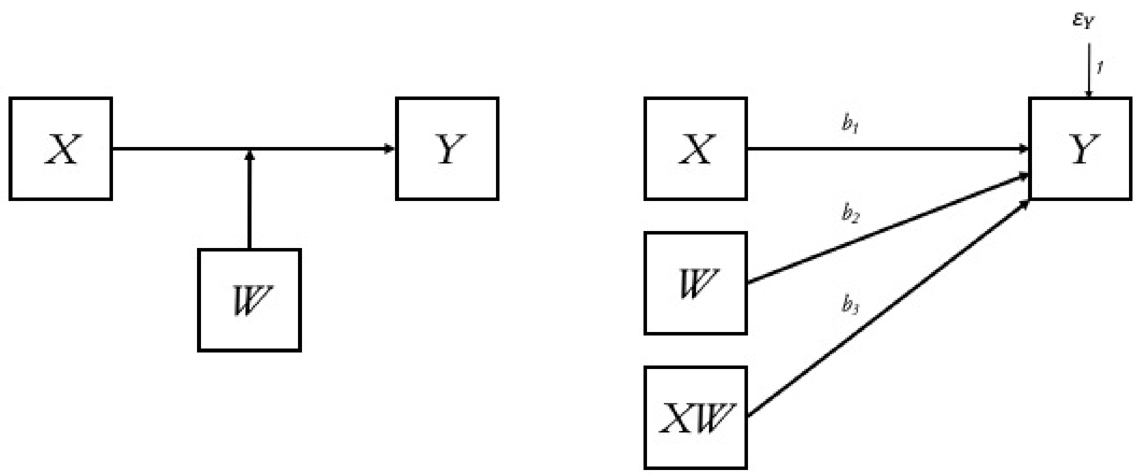

Figure 1.2: Path diagram (left) and statistical diagram (right) for Simple Moderation.

To conceptualize the difference between linear regression with and without moderation, it can be helpful to use real variables. Recalling the previous example, we can model the relationship between university ranking (X) and school spirit (Y), moderated by whether or not a university has a football team (W).

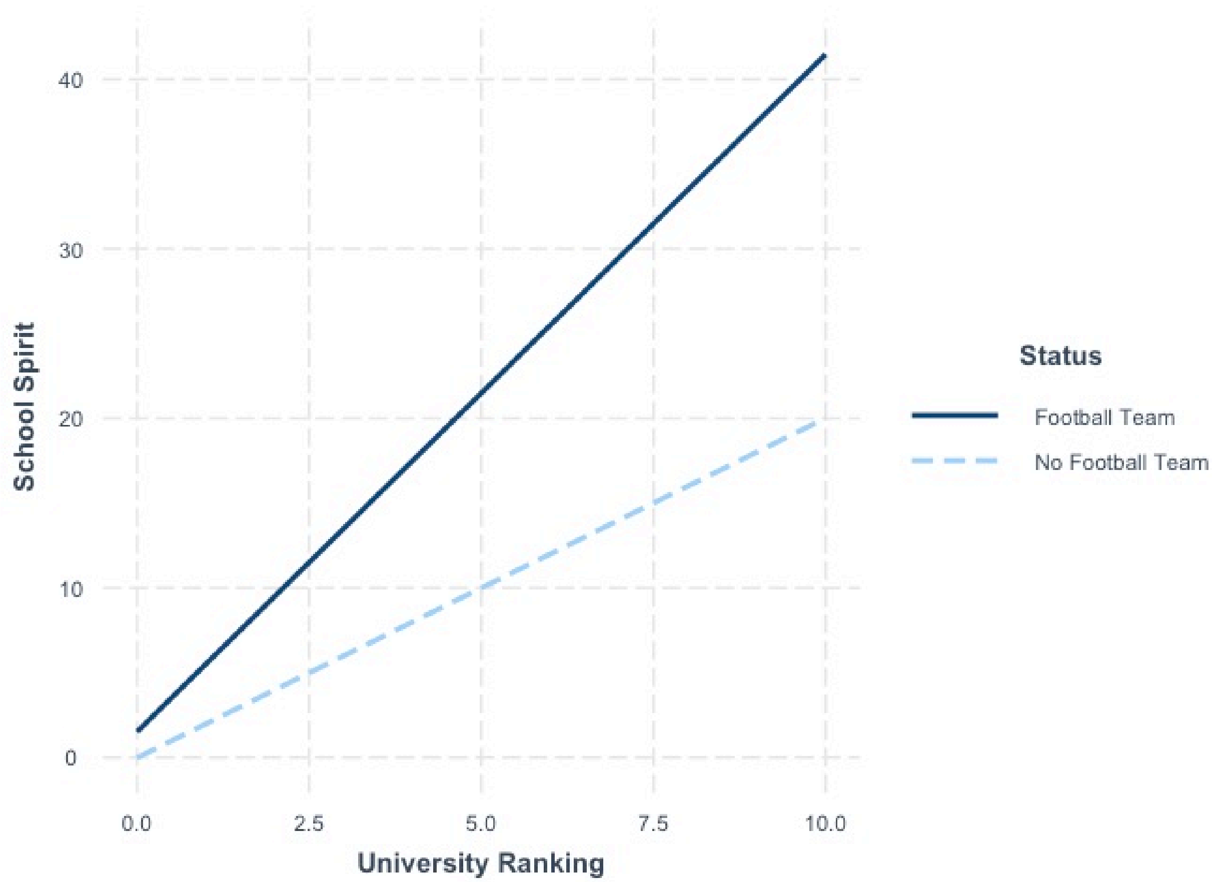

Figure 1.3: Simple slopes graph of the relationship between university ranking and school spirit moderated by football team status.

Once a significant interaction is found, it can be investigated further through conditional effects. The conditional effect in a simple moderation model is $b_1 + b_3W_i$, from Equation 1.5 when written in its equivalent form, $Y_i = b_0 + (b_{1_i} + b_3W_i)X_i + b_2W_i + \varepsilon_{Y_i}$. It is called the conditional effect because the effect of X on Y is conditional on the value of W. Continuing the example from the previous paragraph, Figure 1.3 shows this conditional effect using simple slopes. University ranking is on the x-axis in the graph, and school spirit is on the y-axis. The two lines both represent the relationship between university ranking and school spirit, separated by whether or not the university has a football team. Representing this example using values in Equation 1.6, the coefficient $\hat{b_3} = 2$ represents the difference between the two slopes.

$$Y_i = 2X_i + W_i + 2X_iW_i + \varepsilon_{Y_i} \tag{1.6}$$

If moderation was not present, b_3 would be zero, and the slopes of both lines in the graph would be the same. If the coefficient is significant using a hypothesis test, we can conclude that the relationship between university ranking and school spirit depends on having a football team, and this difference in slopes can be seen in Figure 1.3.

1.2.2 Sample Size Planning for Moderation

Moderation analyses should be carefully planned in psychological studies, including a clear description of exactly which variables interact and determining an adequate sample size to detect the interaction of interest (Rohrer & Aslan, 2021). In estimating power, the size of the effect is inversely related to sample size, where the smaller the effect, the larger the sample size needed to detect the effect. There are a few different measures of effect size for interactions. The most common measure is the proportion of explained variance in the outcome by the predictor(s), which is the change in R^2 between the full and reduced model. For moderation, the change in R^2 is the variance in the outcome uniquely explained by the interaction (full model including the interaction, reduced model including only the main effects), scaled by the total variance in the outcome. There are tools available to do power analysis using this information, including G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), Shiny app and R package 'InteractionPoweR' (Baranger, 2022), and 'pwr2ppl' (Aberson, 2019a) in R, described in more detail below. For more meaningful interpretations, researchers should not rely on Cohen's effect size benchmarks when choosing effect sizes and instead focus on finding estimates from the literature (Aguinis et al., 2005; Correll, Mellinger, McClelland, & Judd, 2021), though those tend to be affected by publication bias (Kühberger, Fritz, & Scherndl, 2014), or rely on another method such as specifying the smallest effect of interest

(Lakens, 2022), especially since most interactions tend to be even smaller than a small effect using benchmarks from J. Cohen (1969) (Aguinis & Stone-Romero, 1997).

Compared to mediation analysis, where software packages needed to be specifically designed to do power analysis for the indirect effect, any software package that can do power analysis for regression can be used for moderation. These all still require the researcher to specify the size of the interaction effect they are expecting along with their α level, and either their desired power to get a sample size or a sample size to see what the power would be (Faul et al., 2007). When an interaction is included in the model, there are several different ways to specify the effect size. Recently developed tool 'InteractionPoweR' can use Pearson's correlations among all the variables in the model as the effect size estimate. Alternatively, this population effect size can also be specified in terms of the change in R^2 between the model with the interaction and the model not including the interaction. 'InteractionPoweR' or G*Power can be used for this, or the 'pwr2ppl' R package has a special command called 'regintR2()' for moderation models (Aberson, 2019b). All require as input the R^2 estimate from the full model with the interaction term and the change in R^2 from the interaction, the number of predictors in the full model, and the sample size to calculate power analytically (Maxwell, 2000; J. Cohen, 1988). Both R^2 values are required to calculate the effect size as f^2, shown in Equation 1.7 for the impact of the full model including the interaction term (R^2_{full}) above and beyond the model without the interaction term, with the R^2 from the interaction term denoted as R^2_{int}.

$$f^2 = \frac{R^2_{int}}{1 - R^2_{full}} \tag{1.7}$$

G*Power is more flexible and can allow for multiple predictors to be compared as a set, so it also requires the number of predictors in the reduced model whereas 'pwr2ppl' and 'InteractionPoweR' assume the difference between the number of predictors in the full model compared to the reduced model is one. The output from both is statistical power. G*Power can advantageously output a sample size for a desired power. Less efficiently, 'pwr2ppl' and 'InteractionPoweR' include a method for giving a range of sample sizes to iterate through and output a list of power estimates. These tools are all for just moderation analysis, but moderation can also occur in combination with a mediation model.

1.3 Moderated Mediation Analysis

Moderation can occur on any path in a mediation model, meaning mediation and moderation can occur in the same model (Hayes, 2013). When the paths which make up the indirect effect are moderated, it is a moderated mediation model (Edwards & Lambert, 2007). The arrangement of variables in these types of models should always come from theory, since these decisions can make it possible to detect effects that are not real because the model was misspecified.

Many mediation, moderation, and moderated mediation models can be estimated with a commonly used macro available for SPSS, SAS, and R: 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 dissertation focuses on simple mediation models (a single mediator) with one or more paths moderated. Similar to simple mediation analysis with exactly one mediator variable, moderated mediation with exactly one mediator variable and one outcome variable requires two equations to describe the model. The first equation predicts M, and the second equation predicts Y. There are two possible equations for M, depending on if moderation occurs on the X to M path or not. Without moderation, the equation is the same as Equation 1.2. When moderation occurs only on this path, it is called first-stage moderation because only the path between X and M is moderated (Hayes, 2015). With W moderating the path between X and M, Equation 1.8 predicting M includes an interaction between X and W.

$$M_i = a_0 + a_1 X_i + a_2 W_i + a_3 X_i W_i + \varepsilon_{M_i} \tag{1.8}$$

The second equation needed for moderated mediation predicts Y. There are four possible options for this equation to describe the six moderated mediation models in this study. First, it could include no interactions at all if the model only includes first-stage moderation, and the equation is the same as Equation 1.3. Second, it could include an interaction between M and W. When moderation occurs only on this path, it is called second-stage moderation (Hayes, 2015). If moderation occurs only on this path and not the X to Y path, the equation for Y includes only one interaction and is represented by Equation 1.9. Third, When just the X to Y path is moderated but not the M to Y path, the equation for Y includes only one interaction as well between X and W, and is represented by Equation 1.10. 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 1.11.

$$Y_i = c'_0 + c'_1 X_i + b_1 M_i + b_2 W_i + b_3 M_i W_i + \varepsilon_{Y_i} \tag{1.9}$$

$$Y_i = c'_0 + c'_1 X_i + c'_2 W_i + c'_3 X_i W_i + b M_i + \varepsilon_{Y_i} \tag{1.10}$$

$$Y_i = c'_0 + c'_1 X_i + c'_2 W_i + c'_3 X_i W_i + b M_i + b_3 M_i W_i + \varepsilon_{Y_i} \tag{1.11}$$

Putting these equations together in different combinations creates the six most commonly used moderated mediation models, with the conceptual diagrams shown in Figure 1.4. For each model, the conceptual diagram shows how the variables appear theoretically and the moderated paths are marked with an arrow pointing directly from the moderator to the path. I 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). Figure 1.5 shows the statistical diagrams for each of the six models used in this dissertation, where each term in the equation is represented with an arrow pointing towards the outcome variable it is used to predict.

The equations used to define each model are also printed on the statistical diagrams in Figure 1.5, and also listed in Table 1.1. Pairing together the equations for M and the equations for Y results in eight possible models. This dissertation focuses on six of these, as displayed in the following figures. Two combinations are not used in this study: No model uses Equation 1.2 for M and Equation 1.3 for Y because this results in a simple mediation model without any moderation, and no model uses Equation 1.2 for M and Equation 1.10 for Y because then only the direct effect would be moderated, and thus not a moderated mediation.

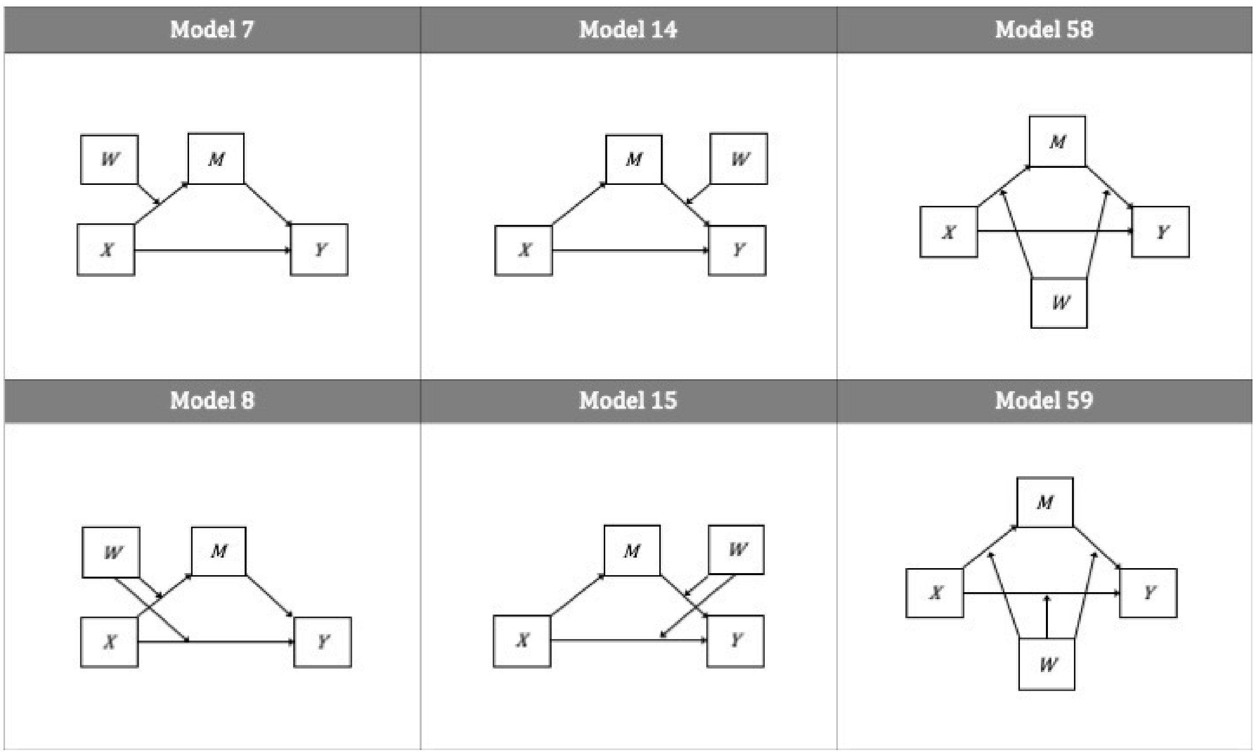

Figure 1.4: Moderated mediation conceptual diagrams.

1.3.1 Conditional Indirect Effects

Moderation means the effect of a predictor variable on the outcome variable differs at different levels of the moderator variable. In a simple mediation model, the effect of interest is the indirect effect, which is calculated by multiplying the a path by the b path. Extending this

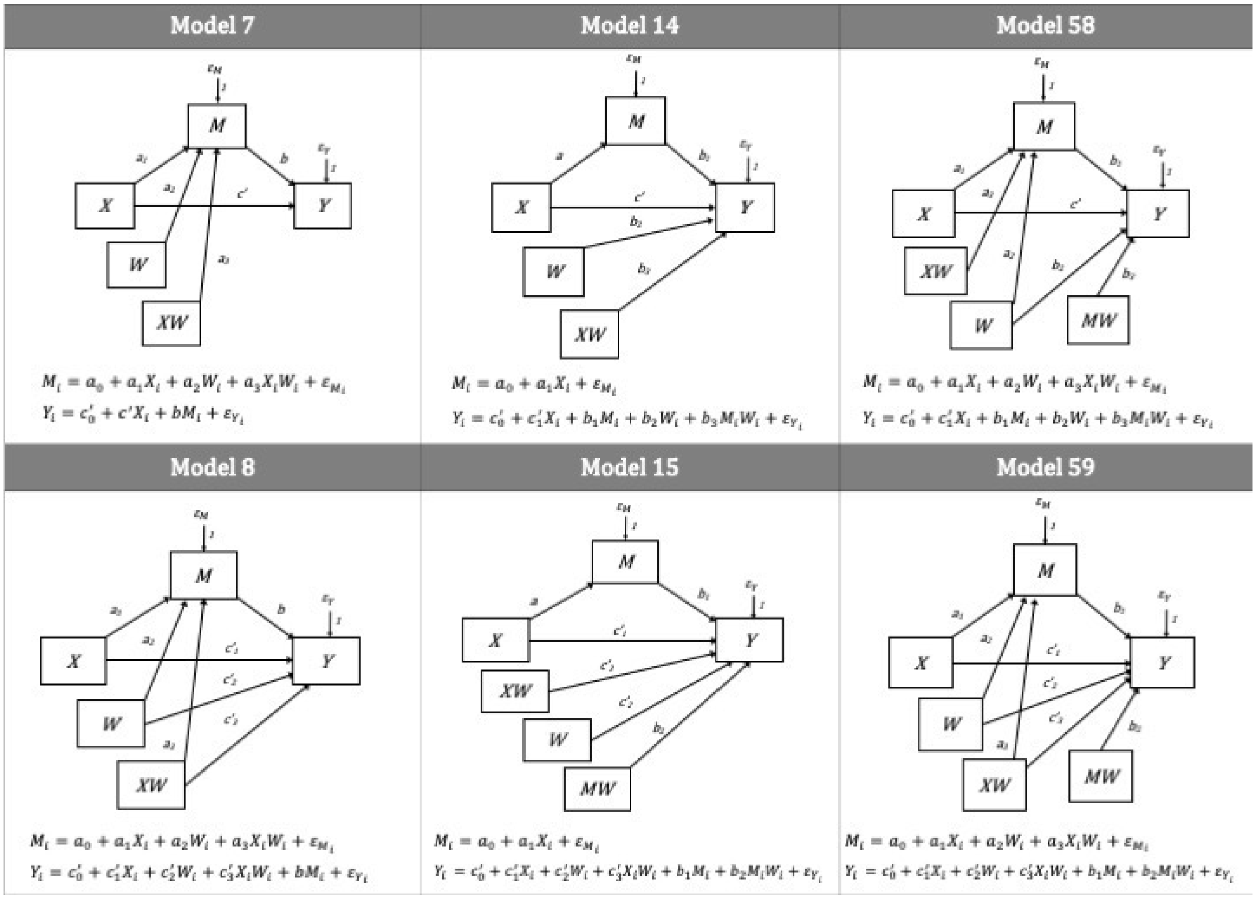

Figure 1.5: Moderated mediation statistical diagrams with corresponding equations.

to include the moderator depends on where the moderation occurs, and the indirect effect becomes the conditional indirect effect, conditional on the value of the moderator. Using Model 7 as an example, the conditional indirect effect is the combination of the effect of X on M, which is $a_1 + a_3W$, and the effect of M on Y, which is b, multiplied together to get $(a_1 + a_3W) \times b$. This is the indirect effect, conditional on a particular level of the moderator variable W. Extending this logic to Model 14, where instead the M to Y path is now moderated instead of the X to M path, the conditional indirect effect is $a \times (b_1 + b_3W)$, still conditional on a particular level of the moderator variable W. The conditional indirect effect for Model 8 is the same as for Model 7, and the conditional indirect effect for Model 15 is the same as for Model 14. The conditional indirect for both Model 58 and 59 is $(a_1 + a_3W) \times (b_1 + b_3W)$. Estimates will differ, however, depending on if the direct effect is moderated or not. Finding the conditional indirect effect for any given model means that the indirect effect of X on Y through M is dependent on setting a value for the moderator W. Table 1.1 gives the index of moderated mediation and conditional indirect effects for the six moderated mediation models described in this section.

1.3.2 Inference for Conditional Indirect Effects

Inference is conducted on the conditional indirect effect (and index of moderated mediation, described below) using a percentile bootstrap confidence interval. Bootstrapping is the recommended method for conducting inference because product terms are not normally distributed, and bootstrapping 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).

1.3.3 Index of Moderated Mediation

The index of moderated mediation is often used for testing the statistical significance of a moderated mediation model (Hayes, 2015), where the indirect effect is conditional on the value of the moderator (W). The difference of the conditional indirect effect when W is set two values, each one unit apart. For Model 7, the index of moderated mediation is calculated for chosen values of W in the equation $(a_1 + a_3W) \times b$. Multiplying this conditional indirect effect out, we have $a_1b + a_3bW$ as an example for Model 7. When 0 and 1 are chosen as the two values of W, we solve $(a_1b + a_3(1)b) - (a_1b + a_3(0)b)$, which equals a_3b. For Model 14, where the conditional indirect effect is $a \times (b_1 + b_3W)$, we get the difference between two values of W as $ab_1 + ab_3W$. Again setting 0 and 1 as two values of W, we solve we solve $(ab_1 + ab_3(1)) - (ab_1 + ab_3(0))$, which equals ab_3. If two different values of W were chosen, this effect would be the same, though scaled differently if the two values were anything other than one unit apart. When the moderator is dichotomous, this estimate for Models 58 and 59 is calculated as $(a_1 + a_3(1)) \times (b_1 + b_3(1)) - (a_1 + a_3(0)) \times (b_1 + b_3(0))$,

Model	Equation for M	Equation for Y	IMM	Conditional Indirect Effect
7	Equation 1.8	Equation 1.3	a_3b	$(a_1 + a_3W) \times b$
8	Equation 1.8	Equation 1.10	a_3b	$(a_1 + a_3W) \times b$
14	Equation 1.2	Equation 1.9	ab_3	$a \times (b_1 + b_3W)$
15	Equation 1.2	Equation 1.11	ab_3	$a \times (b_1 + b_3W)$
58	Equation 1.8	Equation 1.9		$(a_1 + a_3W) \times (b_1 + b_3W)$
59	Equation 1.8	Equation 1.11		$(a_1 + a_3W) \times (b_1 + b_3W)$

Table 1.1: Key equations used for the moderated mediation models in this study. IMM = Index of Moderated Mediation.

or $(a_1 + a_3) \times (b_1 + b_3) - a_1 \times b_1$ (Fairchild & MacKinnon, 2009). This calculated index of moderated mediation is the amount by which the indirect effect differs between levels of the moderator. With continuous moderators, the index of moderated mediation is not defined, but conditional indirect effects can be probed at any level of the moderator. Typically, -1, 0, and 1 standard deviation away from the mean or probing at particular percentiles to ensure the values exist in the data are chosen (Hayes, 2022).

1.3.4 Sample Size Planning for Moderated Mediation

Statistical power to detect the index of moderated mediation or a significant conditional indirect effect is difficult to estimate, especially because there is no closed-form solution for statistical power for these models. However, it is possible to use the joint significance test for testing moderated mediation (Coutts, 2023), which does have a closed-form solution defined for mediation (O'Rourke & MacKinnon, 2014). Unlike mediation and moderation, there are also not many resources for sample size planning for moderated mediation. Power analysis for moderated mediation Models 7 and 14 are documented and available in the R package 'pwr2ppl' (Aberson, 2019b), and Models 8 and 15 are possible using version 0.5.0 of 'pwr2ppl' as well. This R package handles power analysis for many complex analyses, and since moderated mediation is a more recent addition fewer models are available in its current version. Each moderated mediation model has many more parameters that must be specified compared to simple mediation or moderation models. To determine power for

Model 7, for example, the correlation between each pair of variables is required (including the correlation between variables and the interaction, where collinearity is not an issue if it is only between the variables and the product term), in addition to the sample size, type I error rate, and number of repetitions for it to simulate. With these complexities, there is no "one size fits all" function for power analysis for moderated mediation yet. Another thing to note is that output from the 'pwr2ppl' is again limited to power for an inputted sample size. This can be a barrier for researchers looking to determine a sample size quickly for a desired level of power since it will require testing different sample sizes until one is found that reaches adequate statistical power. Within the last year, WebPower has also added some moderated mediation capabilities as well and can do power analyses for Models 7, 8, 14, 15, and 58. However, power analyses using WebPower can be time-consuming, with Model 58 taking about ten minutes using the web application with all the default settings including 500 simulations to calculate power for a single sample size.

1.4 Current Study

Any number of paths in a mediation model can be moderated (Preacher, Rucker, & Hayes, 2007). It could be that the interaction happens on every path, a few select paths, or even on just one path in the model. When paths are moderated, there should be a justification for the theoretical framework of why an interaction occurs between those particular variables (Fiedler, Schott, & Meiser, 2011). The current research was motivated by a meta-analytic review finding that 14% of surveyed moderated mediation analyses moderated all the paths in the mediation model (See Chapter 2). This could be that theory in each case suggested that all paths should be moderated, or researchers include moderation on all the paths because the theory lacks specificity about where moderation occurs. If a model is over-specified in this way, then statistical power could be affected by the excess moderation in the model. Lavery, Acharya, Sivo, and Xu (2019) found that additional predictor variables in a model increased type II error rate holding all their other study parameters constant, which is relevant for moderated mediation models in particular. If the model is specified correctly, then researchers could have higher statistical power to detect significant moderated mediation.

This dissertation examined the impacts of model specification on statistical power and type I error rate. Several factors can influence statistical power for moderated mediation studies, such as sample size (Aberson, 2019a), amount of variance of the outcomes explained by the predictors in the model (J. Cohen, 1988), and dichotomous vs. continuous variables (McClelland & Judd, 1993). In this study, I also tested the effects of model specification on statistical power and type I error rate in moderated mediation models. The specific hypotheses are outlined below. Overall, this dissertation explored if the number or location of interactions can also affect statistical power and type I error rate in cases where the data analysis model may not match the true underlying data generating process (DGP).

There were two models for each analysis: the DGP, which is the true population underlying moderated mediation model used to generate the data, and the data analysis model, which is the model I fit and use to analyze the data. I refer to cases where the data analysis model and the data generating model do not match broadly as model misspecification. Within model misspecification, I differentiate between three possible types of misspecification that are possible for moderated mediation models. First, there is *over-specification*: allowing all of the paths moderated in the DGP to be moderated in the data analysis model, plus at least one additional moderated path. For example, if the DGP is Model 7 and the analysis model is Model 59, the model is over-specified because the analysis model is the DGP plus additional paths moderated. This is technically not a misspecification because the data generating model is a special case of the analysis model where specific path(s) are zero. Second, there is *complete misspecification*: The analysis model omits a moderated path from the indirect effect, and moderates a different path in the indirect effect. In other words, one path in the indirect effect is moderated in the DGP and the other path in the indirect effect is moderated in the data analysis model, and it is never the same path moderated. For example, if the DGP is Model 7 and the analysis model is Model 14, the model is completely misspecified because the analysis model has omitted one moderated path from the indirect effect and moderated the other one. Third, 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. Finally, *correct specification* refers to the DGP and data analysis model matching perfectly.

Research Question 1 examined what factors impact the statistical power of the index of moderated mediation for over-specified models compared to correctly specified models. To answer this research question, I only included over-specified models and correctly specified models. Statistical power was the outcome for over-specified models because detecting a significant index of moderated mediation in an over-specified model was still detecting the "true" underlying effect based on the DGP. I had three hypotheses related to statistical power for over-specified models compared to correctly specified models.

- H1a: I hypothesized that the statistical power of the index of moderated mediation will be lower for over-specified models compared to correctly specified models.

- H1b: I also hypothesized that power will be lower for over-specified data analysis models with additional moderated paths. For example, if the DGP is Model 7, then Models 8, 58, and 59 are all over-specified. Models 8 and 58 have two moderated paths but Model 59 has three. I hypothesized that for cases like this, for models with more moderated paths (Model 59), power will be lower than models with fewer moderated paths (Models 8 and 58).

Research Question 2 examined what factors impact the statistical power of the index of moderated mediation for under-specified models compared to correctly specified models. To answer this research question, I only included under-specified models and correctly specified models, because a significant index of moderated mediation detected by an under-specified model would still be detecting the “true” underlying effect based on the DGP. I had two hypotheses related to statistical power for under-specified models compared to correctly specified models.

- H2a: I hypothesized that the statistical power of the index of moderated mediation would be lower for under-specified models compared to correctly specified models.
- H2b: I also hypothesized that under-specified models would have higher parameter bias than correctly specified models.

Research Question 3 examined what factors impact the type I error rate of the index of moderated mediation in completely misspecified models. To answer this research question, I only included completely misspecified models because these are the only cases in moderated mediation where detecting a significant index of moderated mediation would be a type I error based on the DGP. The analysis model, if completely misspecified, should find the index of moderated mediation to be zero. I had two hypotheses related to type I error rate for completely misspecified models.

- H3a: I hypothesized that the type I error rate would be too high according to criteria set by Bradley (1978) and Serlin (2000) in completely misspecified models. The former criterion is more extreme than the latter, so it will be possible to differentiate between how many type I error rates were moderately too high and severely too high.
- H3b: I also hypothesized that the type I error rate would be higher in cases where the analysis model includes two moderated paths (i.e., Models 8 and 15) compared to one (i.e., Models 7 and 14).

This dissertation looked at six commonly used moderated mediation models and compares the statistical power and type I error from the different model specifications across effect sizes and sample sizes common in published moderated mediation studies. Analyses were performed for over-specified models, under-specified models, and completely misspecified models. The amount of explained variance, sample size, and type of predictor variables were manipulated. Conclusions from this study can inform the degree to which model specification impacts statistical power and type I error rates in moderated mediation models.

2 Meta-Analytic Review

Moderated mediation analysis has gained increasing attention as a statistical approach for exploring complex relationships between variables. This systematic, meta-analytic literature review includes a data collection aiming to provide a comprehensive overview of common methodological choices made in current practice. Additionally, the purpose of the review is to help understand if moderated mediation studies are well-powered. To accomplish these goals, we extracted several key pieces of information from articles that used moderated mediation analysis. These variables included sample size and where the moderation occurred in the mediation model, among other relevant pieces of information. The main objective of this meta-analytic review data collection was to explore current practices in sample size planning and models used for moderated mediation, and additionally, the results can now be applied to help inform a simulation study about factors that impact statistical power in moderated mediation.

Monte Carlo simulation is used widely in methodological research, allowing for the evaluation of the performance of different statistical techniques on simulated data. For this type of research, a crucial yet arbitrary step is specifying parameters in the simulation, which can have a huge impact on the final results. For example, a Monte Carlo simulation study was done to calculate sample sizes required for 80% power at various effect sizes for mediation analysis, reported in Fritz and MacKinnon (2007). If the effect sizes used had all been greater than .3, yet the average effect in the field of social psychology is around .21 (Schoemann et al., 2017), the Monte Carlo simulation study would not be particularly useful. Doing a meta-analytic review before a Monte Carlo simulation study can be useful if the goal of the simulation is to understand how methods perform in current practice. The simulation study in this dissertation uses information from published articles to inform two important parameter decisions, to make the results more applicable to a wide range of researchers. For example, we found that the six most common models (of which there are over 48 models in the PROCESS macro commonly used for these types of models) account for 85% of published moderated mediation models from this systematic meta-analytic review. This information helped inform the choice of which models to include in the simulation study for it to be most relevant according to current practice.

Since no particular PROCESS model is argued to be more important or accurate than another, I used the results of this meta-analytic review to narrow down the scope of the study and focus the simulation study on these six models, knowing it will be valuable to gain insight about these moderated mediation models used roughly 85% of the time. Additionally, information gathered through this meta-analytic review informed which sample sizes to include in the simulation. To select these important parameters that most accurately represent what is currently used in the field, we examined a set of published moderated mediation analyses for this information. We also wanted to collect this data to explore relationships between these variables.

We found it important to understand if more complex models used larger sample sizes on average, and an analysis of the results from this meta-analytic review allows for those comparisons. Sim, Kim, and Suh (2022) performed a similar literature review focusing on mediation, finding that the median sample size for a simple mediation analysis was 242, while the median sample size for the more complex design with two mediator variables in parallel was only 171. While this finding was based on only 50 published mediation models, it does demonstrate that larger sample sizes are not always used with more complex designs. When models include many interactions, which are typically very small (Aguinis et al., 2005), it can take a larger sample size to have sufficient statistical power (J. Cohen, 1988). The results from this meta-analytic review for moderated mediation show the average sample size used by model specification, to understand if more complex models tend to have higher sample sizes.

The simulation study for this dissertation focused on factors that affect statistical power in moderated mediation analysis. This meta-analytic review focuses on sample size and model specification. Parameters for future simulation studies can be informed by conducting a similar meta-analytic review to understand what values are commonly seen in a particular field. This chapter gives an overview of the article coding process from the moderated mediation meta-analytic review, including the scope of articles included and the procedure for coding the articles. In this chapter, I also give descriptive statistics from the variables coded in this meta-analytic search. Additionally, I describe where this information is available in a searchable online database, along with the future of the database and how articles can continue to be added. This is all to understand what sample sizes are commonly used in moderated mediation analyses. I then use this information such as common sample sizes and model specifications to directly inform parameter choices in the simulation study described in greater detail in the next chapter.

2.1 Method

To select the articles and gain a broad understanding of current practices for moderated mediation analyses, articles were selected using WebofScience. The subject area was not restricted, and keywords for the search were *moderated mediation*, *mediated moderation*, and *conditional process analysis* (Hayes, 2022). We used the last full calendar year at the time of initial article collection and found 411 analyses in published articles that met this criteria from 2018. We coded each article for several key variables, with two informing parameter values in this simulation study: sample size and moderated mediation model specification. We also recorded other variables that could potentially be related to statistical power, including the use of regression vs. structural equation modeling, the number of variables, and the number of analyses. There were three waves of data collection for this project. Wave 1 includes all articles that were part of this meta-analytic systematic review. Waves 2 and 3 were continuations of the original review completed by two different groups of research assistants, done to include even more recent articles and we also intentionally sought out a wider variety of models. Since Waves 2 and 3 used different coders and the selection of articles was not as objective as Wave 1, articles coded during Waves 2 and 3 were not included in this meta-analytic systematic review. However, information coded in these waves is available in the database described later in this chapter.

2.1.1 Scope of Articles Included

There were originally 518 articles resulting from the WebofScience search, then we excluded 163 articles that were either solely methodological such as Hayes (2018b), articles that used multilevel modeling, or different definitions of the terms such as *mediation* in a conflict resolution study with a third party *moderator*. The remaining 382 articles contained 411 moderated mediation analyses, each from a separate study. Only these 411 studies from Wave 1 were used to inform parameters in this dissertation simulation study.

2.1.2 Procedure for Coding Articles

Moderated mediation analyses from Wave 1 were all double-coded by research assistants in the QRClab. Coders were trained during a weekly research meeting on what each of the coded pieces of information was and where in a typical research article it could be found. Training documents were accessible to everyone throughout the coding process. Coders received a PDF from WebofScience with their assigned articles each week, then they found their articles and entered the relevant information into a spreadsheet. The four coders were systematically paired with a different coder each week (i.e., in Week 1 the pairs coding the same articles were AB and CD, in Week 2 the pairs coding the same articles were AC and BD, in Week 3 the pairs coding the same articles were AD and BC, then that cycle repeated). At the end of each week, I combined all the data into a spreadsheet to identify any discrepancies between coders on the same articles. We resolved these discrepancies together during the weekly meetings using the original article. Once agreement was reached by reviewing the article with the whole team and discussing it, the articles were considered coded and finalized.

Wave 2 of the project involved a research assistant receiving the same training and following the same procedure to code 74 additional articles. These are all articles that cited Hayes (2015), which was the paper introducing the index of moderated mediation, and we only included articles that actually ran a moderated mediation analysis on empirical data. These articles were only coded by one research assistant, but we met weekly to review all the articles. During this wave, the decision about which variables to include was revised, and the specific decisions about which variables to begin recording or coding differently are discussed in greater detail in the Coded Variables section below.

Wave 3 of this project was done to include more recent articles. This was used as a project for all research assistants in the lab if they had extra time after completing other lab activities each week. All members of the lab received training during the first meeting of the year, and articles were assigned to research assistants and were double-coded. Due to the lab having many projects going on during this time, only seven articles were added in this wave. Only articles that included a path diagram were assigned during this phase since those tended to be the most straightforward to code and these research assistants had the least direct supervision. I monitored a dedicated Slack channel for research assistants through which they could ask and answer questions. I then met with a senior research assistant who found all the articles that had been completed by both coders and after resolving discrepancies in that meeting by reviewing the articles together, they finalized the article information.

2.1.3 Coded Variables

Undergraduate research assistants extracted information from articles that included a moderated mediation analysis. The ten pieces of information were determined by Dr. Montoya and myself to better understand current practices in published articles for moderated mediation. Coded information included:

- Sample size
- PROCESS model number (Hayes, 2022)

- Number of moderator (W) variables
- Number of mediator (M) variables
- Number of antecedent (X) variables
- Number of outcome (Y) variables
- Number of covariates
- Number of moderated mediation analyses with different variable arrangements reported
- How the model was estimated (possible levels: regression or structural equation modeling)
- Primary and secondary research area (possible levels: psychology, business, etc.)

Notes and reference information were also recorded for each article, along with other exclusion criteria such as containing the words *moderated* and *mediation*, but not performing a moderated mediation analysis as defined in Hayes (2015). Multilevel models were excluded due to the complexity of the studies and because additional different information would need to be collected to determine statistical power. For example, Anand, Hu, Vidyarthi, and Liden (2018) used moderated mediation to evaluate employee performance in work groups, but we did not include it in the final analysis because the number of employees and the number of work groups added a level of complexity not captured by the single sample size piece of coded information and would have different statistical power.

From Wave 2 of article coding onward, we reevaluated our categories to optimize necessary information and also streamline the process in some cases. Covariates from this point on were a binary yes/no decision instead of including the number of covariates due to the difficulty in determining the exact number of covariates in a study. Additionally, a new category was introduced that might play a role in statistical power: type of X variable. Options included continuous, multi-categorical, and dichotomous. These categories remained the same for Wave 3 of coding.

The two main variables of interest used for this dissertation are sample size and PROCESS model number. The sample size is the number of participants in the study. If multiple studies were included in the same article, the sample size specifically referred to the study analyzed with a moderated mediation model. The final sample size retained for analyses was used, even if the total sample size was also reported. PROCESS model number was most typically determined by matching the figure in the manuscript with a figure in Appendix A of Hayes (2022), or by drawing out the model described in the text and matching it if no figure was provided in the manuscript. If a model was found that did not match any PROCESS model number, this field was left blank.

2.2 Results

Results from this meta-analytic review come from the 411 analyses included in Wave 1. Findings about sample sizes used are presented first, followed by findings about PROCESS model specifications. I also present other findings from the meta-analytic review, including the proportion of analyses done using regression vs. structural equation modeling, and which research areas were most common among the articles.

One key parameter for the simulation study was sample size because to be a useful guide for researchers the sample sizes should approximate what they might reasonably be able to collect. The sample sizes and moderated mediation model specifications were thus used to inform the sample sizes and models used in the simulation study. Sample sizes from the articles ranged from 29 to 456849 across all models. Figure 2.1 is a histogram displaying the sample sizes found in log form. To get representative sample sizes, I found the deciles of sample size were 120, 160, 201, 241, 285, 364, 459, 731, and 1307. With generous rounding, the final sample sizes used in this simulation study were 100, 150, 200, 250, 300, 400, 500, 750, and 1,000.

Another research question from this meta-analytic review was about how frequently each moderated mediation model was used. We investigated which models were most commonly used, and if there were any included in the PROCESS macro that were more rarely used. Of the 411 analyses coded in this search, six specifications were not identified as matching a PROCESS model, represented as N/A. Information collected through this meta-analytic review revealed the six most commonly used moderated mediation models: Model 7, Model 8, Model 14, Model 15, Model 58, and Model 59. Together, these six models accounted for 85% of the published moderated mediation analyses we coded. These are the six models included in the simulation study. Box plots displaying the sample sizes by model are provided in Figure 2.2. We may have included additional models unnecessarily without this meta-analytic review, not knowing that these six make up such a large proportion of published moderated mediation analyses.

By far, the most common model was Model 7, with moderation occurring on the X to M path. Table 2.1 gives the median sample size for each model specification, choosing to report the median instead of the mean due to a few very large outlier sample sizes. The table is in descending order based on model use frequency. Based on results reported in Table 2.1, there appears to sometimes be a pattern between sample size and model complexity, at least among the most commonly used models. Model 59 has a higher median sample size than any of the other six most commonly used models, but the median sample size used with Model 58 is between the median sample size for Model 7 and Model 14.

We were also interested in whether ordinary least squares regression or structural equation modeling was more common, and if their use differed regarding sample size. The PROCESS macro uses regression, so researchers using structural equation modeling are using other packages or tools to fit these models. Only 66 articles used structural equation modeling compared to 343 using regression, and two articles were unclear. The median sample size using regression was 266, and the median sample size using SEM was 481. Interestingly, regression was used for all three analyses with the highest sample sizes (all over 50,000). With

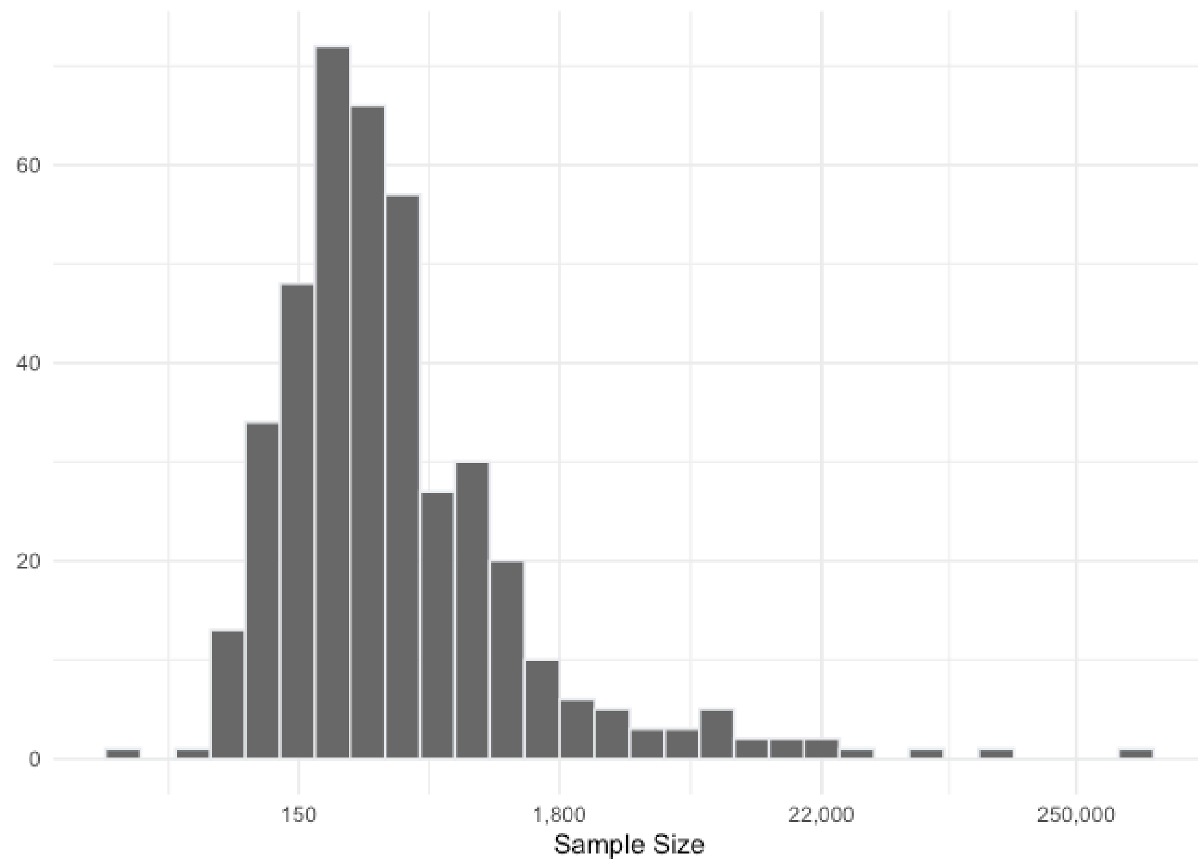

Figure 2.1: Histogram of the sample sizes recorded in the moderated mediation meta-analytic review. Due to the high outliers, this figure depicts log(Sample Size).

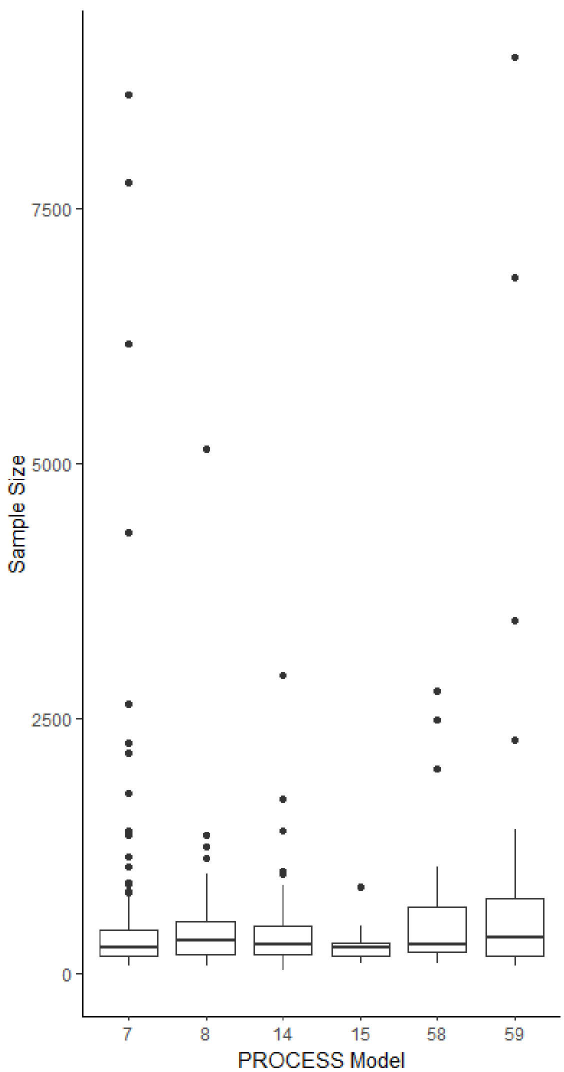

Figure 2.2: Box plots displaying the range of sample sizes reported in the articles included in the meta-analytic review, separated by PROCESS model. For clarity, outliers above 10,000 were excluded. Excluded outliers were sample sizes of 14,908 and 60,133 using Model 7, 456,849 using Model 8, 123,459 using Model 14, and 12,646 using Model 59.

Model Number	Median Sample Size	Frequency
7	261	31%
14	288	18%
59	363	14%
8	331	13%
58	276	6%
15	255	3%
9	317	3%
21	199	2%
74	430	2%
5	260	1%
11	174	1%
16	616	$< 1\%$
22	136	$< 1\%$
37	410	$< 1\%$
63	1307	$< 1\%$
72	131	$< 1\%$
75	1995	$< 1\%$
86	405	$< 1\%$
87	270	$< 1\%$
92	225	$< 1\%$
N/A	8320	$< 1\%$

Table 2.1: Median sample size separated by moderated mediation model and percent of analyses using that model specification. Rows are ordered by model use frequency, with the six models used in the simulation study in Chapter 4 in bold.

these high outliers, we can use quantile regression to see if the group difference is significant. The median sample size for analyses done using regression was 270, while the median for structural equation modeling was 491. This was a statistically significant difference, with the confidence interval using standard errors produced using a rank test being [79.10, 353.486] around the coefficient (Koenker, 1994). From this analysis, we concluded that analyses using structural equation modeling used higher sample sizes. Since structural equation modeling offers flexibility in model specification beyond the PROCESS models, this finding suggests that more complex models could use higher sample sizes.

We did not limit the research field in which we searched for articles. The majority of the analyses (136) were from psychology, with a median sample size of 287, followed by

76 analyses in business and economics with a median sample size of 255. The rest of the research areas had fewer than 25 published moderated mediation analyses. The six most common PROCESS model specifications remained the same when looking at the subset of psychology articles, but business and economics used Model 21 a total of five times whereas they only used Models 58 and 59 three times each. These findings about research area make the results of this study most relevant in psychology and partially to business and economics, due to their distributions of model specifications and the fact that those are the fields that publish the most moderated mediation analyses. However, these results should generalize to other fields since the focus was on the analysis and not the content.

2.3 Discussion

Doing a meta-analytic review and extracting relevant information from articles can help provide a clear picture of common practice in a particular field or for a certain technique or analysis, which can make methodological research most useful for applied researchers. In our meta-analytic review, we found the median sample size across all moderated mediation models was 285. Statistical power for any sample size depends on more factors than just a sample size, but to put this sample size into perspective, the sample size needed for a simple mediation model using the average effect size found from a meta-analysis of articles in social psychology on all paths is 290 (Schoemann et al., 2017). When considering a more complicated moderated mediation model specification, with more interactions, it is reasonable to assume it will need a larger sample size for adequate statistical power. Information about current practices such as information collected in this meta-analytic review can inform methodological research, including parameter choices for simulation studies to best represent and learn about scenarios commonly faced in applied research contexts.

This meta-analytic review also highlighted the need for including the complexity of the model in sample size planning for moderated mediation analyses. We saw that somewhat reflected in current practices, in that Model 59 with all paths moderated had a median sample size of 363, compared to most of the other models in the mid-200 range. However, including the specific model specification in a power analysis is the only way to guarantee sufficient statistical power for the actual planned analysis.

2.3.1 Recommendations for Reporting

Through the process of reading articles to pull information for this meta-analytic review project, we determined several helpful decisions authors made in their articles. These were helpful not only for our purposes in the meta-analytic review, but we found that these decisions also allowed readers to more easily understand the analyses that were implemented. First, the decision to include a path diagram as a figure in the manuscript made it far easier to determine which moderated mediation model was used. Additionally, describing the model in words in the data analytic plan provided useful context (e.g., age moderated the path between sleep and depression). Occasionally, the figure and data analytic plan would conflict, so authors should ensure consistency when reporting their models. When this happened, research assistants used their best judgment to decide which model was used, based on results presented in the tables. Articles should also clearly describe how data were analyzed, including what method of estimation was used, as coders found it challenging in many cases to determine if regression or structural equation modeling was used. Articles that reported their results in this way were more easily understood, and these decisions also have benefits for replication.

For reporting moderated mediation analyses, some key pieces of information to include:

1. Clearly labeled path diagram (conceptual and statistical)
2. Data analytic plan section in method, including description of model and any covariates
3. Regression or structural equation modeling (SEM) for data analysis

$\boxed{4.}$ Table(s) of regression or SEM results, including coefficients and confidence intervals

2.3.2 Database

Another major goal of this article coding project was to create a publicly available database so researchers can easily search for examples of published moderated mediation models. All of the information from the articles coded so far is available online in the database: `https://www.jlfossum.com/moderated-mediation-article-database`.

Articles using moderated mediation analysis have been added to the database in two ways: by research assistants in the lab or through a publicly available online form. Research assistants have entered the majority of the articles found in the database as part of their time in the lab. Articles coded by research assistants were double-coded for accuracy, and discrepancies were resolved before entry into the final database housed on Google Data Studio. The other method of article submission is through this online form: `https://forms.gle/wDC3Hw8jAqdEsaxp9`. The fields match the pieces of information the research assistants extracted from the other articles. Researchers can submit their own articles or good example articles they have read. Submissions are reviewed for accuracy and completeness before being made public on the database. Seventeen submissions have been entered using this form, and from those, only 15 provided sufficient information to be included in the database. Quarterly, all the articles in the database are downloaded into a .csv file and updated on UCLA's DataVerse, to keep an updated DOI for this project that has been cited in at least two conference presentations and one publication: `https://doi.org/10.25346/S6/7UTGQH`. For instances where we decided to collect an additional variable not included in the original set of variables, such as type of X variable, the database contains a blank space in this column. A future goal for this project is to go back through and fill that information in for those articles.

To try and get a wide variety of moderated mediation models represented, I specifically searched for more uncommon PROCESS model numbers to provide examples for researchers using these models at the end of Wave 2 of article collection. I did this by searching Google Scholar and WebOfScience for *moderated mediation* (or any of its synonyms) plus *PROCESS model [#]*. I was able to find at least one example of several of the uncommon models, and a research assistant and I inputted those into the database. However, PROCESS models 13, 20, 60, 61, 62, 64-71, 73, 84, 88, 89, and 90 remain elusive. Published work could still exist that uses these models, but if the specific PROCESS model number was not mentioned in the paper, it could not be discovered. Due to us not finding examples of these models in any of the articles coded in any of the three waves does suggest that these are the models that may be used more infrequently, and possibly even not used at all.

2.3.3 Future Directions

The searchable database of moderated mediation articles remains publicly available for researchers to use for any purpose. Articles are still being added by researchers through the online submission form. Anecdotally, I've received several Twitter messages and even an email expressing gratitude for this work. I've also gotten a request for an example using an uncommon model, and while I could not find an example of another paper using that model, I have hope that the authors will publish their paper and then submit it to the database for others to see. This database will remain available for researchers to use and contribute to through the submission form, and I have plans to continue systematically adding new, recently published articles in the coming years when I have an active research lab with interested research assistants.

All of the information publicly available in the database originated from the desire to understand current practices and sample sizes used for moderated mediation analyses. This project involved collecting information such as sample size and model specification from over 400 analyses, with articles still being added. Information from the original coded analyses found that the median sample size used was 285, leading to questions about the statistical power of these complex analyses. The range of sample sizes and most commonly used model specifications in published literature will inform the parameter values tested in a simulation study to learn more about factors that affect statistical power for moderated mediation. Future work could continue to code moderated mediation analyses to track trends over time, to see if certain models become more popular or if sample sizes vary. Additionally, analyses that were excluded previously could be included, to extend these findings to multilevel models popular in psychology, business, and educational research.

Using values from this meta-analytic review will help guide the results from the simulation study to be more informative for applied researchers planning to collect real data and publish a study that includes a moderated mediation analysis. The simulation focuses on the six most commonly used model specifications based on this review, accounting for 85% of published models. The sample sizes in the simulation were also informed by the sample sizes found in this review. The simulation study was based on these general findings, and motivated by problems that can occur during model specification. One such problem regarding statistical power is highlighted in the Applied Example in the next chapter.

3 Applied Example

Moderated mediation analysis requires specifying a model, which includes determining the role and order of each variable in the moderated mediation. Since the effect of the predictor variable on the mediator and the effect of the mediator on the outcome variable can depend on the level of other moderator variables, the direction or strength of the relationships may differ. This is important because theory is what drives the decision of the role of each variable (Fiedler et al., 2011), and nothing in the statistical model can determine if a variable is the predictor, mediator, or outcome (Fiedler, Harris, & Schott, 2018). Models can vary based on where different variables are in the model, but based on statistics alone it is equally likely for the variables to be in any order, such as the mediator being the outcome instead of the other way around (Pieters, 2017; Thoemmes & Ong, 2015). The output for the model does not say if the variables are correctly specified.

Over- or under-specifying moderation effects in a mediation model, or misspecifying the mediation model altogether, can lead to biased or incorrect estimates of the mediated effect and potentially obscure important findings (Fiedler et al., 2018). In this dissertation chapter, I describe an applied moderated mediation model, run two different moderated mediation analyses using the same variables and data set but two different model specifications, then examine the potential consequences of over-specifying a moderated mediation model. The data were collected as part of an ongoing study about guilt and shame, and are used here to illustrate how to test and interpret the moderated mediation effect properly. They are also used to demonstrate what could happen if the model is over-specified. Since this is an example using real data, it is impossible to know the true, correctly specified model in the population. I will assume, for this example, that the "true" model is a moderated mediation model with only one path moderated. This is mainly to demonstrate what could happen with over-specifying a model, which I will also do to this model by moderating additional paths. This will conclude with "missing" an effect in the over-specified model when that effect was present in the model with fewer moderated paths. Had the order been reversed by defining the "true" model as the one with moderation on all paths, then fitting the model with only one path moderated would be an example of an under-specification.

3.1 The Data

These data were collected in 2017 when I was an undergraduate research assistant at Seattle Pacific University. The Principal Investigator for this study was Dr. Thomas Carpenter, in the Moral Emotions Lab. Data were collected by recruiting participants from Amazon Mechanical Turk (mturk.com). Participants were asked to recall a time they were personally responsible for hurting or offending someone, write about that time, and then they filled out several personality measures and concluded with a demographic questionnaire. Only participants who consented into the study via the online consent form and provided a written response to a narrative prompt were kept in the final sample. Data quality was assessed by inspecting transgression narratives for non-serious participants. Seven did not provide valid narratives and were removed.

3.1.1 Participants

The final sample size included for analysis in this study was 390 participants. The participants (225 male, 165 female) identified as non-Hispanic White (74%), African American (11%), Hispanic/Latinx (10%), Asian (7%), and other (3%) and were located in 44 US States and the District of Columbia.

3.1.2 Measures

***X*: Age** Age was measured in years. Age ranged from 20 to 54, $M_{age} = 35.11$, $SD_{age} = 10.41$.

***M*: Repair** Repair, a facet of guilt-proneness, comes from the guilt- and shame-proneness scale measured developed and validated by T. R. Cohen, Wolf, Panter, and Insko (2011).

Participants read a scenario (e.g., "A friend tells you that you boast a great deal. What is the likelihood that you would stop spending time with that friend?") and then indicated the likelihood of the response on a 7-point Likert scale. Participants responded to four scenarios that had to do specifically with this repair facet of guilt-proneness.

***W*: Empathy** For the open-ended transgression narrative response, I created a codebook of nine different themes that could be found within each narrative. These themes included items such as expressions of remorse, any indications of intentionally harming the victim, or whether the narrative contained empathy. A team of undergraduate research assistants read through each narrative and coded based on each of these themes. These variables were operationalized in a codebook I created, and research assistants had access to that codebook and training materials throughout the coding process. All coded variables were binary, indicating whether or not the narrative contained that particular theme. Each narrative was triple-coded, and in the case of disagreement, the choice made by two of the three coders was chosen. This process resulted in interrater reliability as measured with Cohen's κ for dichotomous variables to above .70 for each theme, which is reasonably high (S. Sun, 2011).

***Y*: Disclosure** Disclosure was measured using the Linguistic Inquiry and Word Count. Linguistic Inquiry and Word Count (Tausczik & Pennebaker, 2010) software was used to retrieve the word count of each narrative response. Word count tends to be a useful indicator of narrative properties, as coding for positive and negative emotions using dictionaries tends to be uncorrelated or weakly correlated with self-report measures of those same emotions (J. Sun, Kaufman, & Smillie, 2018).

3.2 Theoretical Model

Conceptually, we were interested in several research questions with this data. Most research questions focused on predicting the outcome variable disclosure (Fossum & Carpenter, in prep). Specifically, how moral emotions such as shame-proneness and guilt-proneness

(T. R. Cohen et al., 2011), or any of the facets of either emotion such as repair, predict disclosure. There were several mediation hypotheses, including one about the repair facet of guilt-proneness acting as a mediator in a relationship predicting disclosure. This model uses age as an independent variable, based on a similar model from Carpenter, Isenberg, and McDonald (2019). In Carpenter et al. (2019), researchers found an indirect effect of age through repair, though they used a different outcome variable. Age was chosen as the predictor variable in the Carpenter et al. (2019) model because there was not a variable that predicts age, and guilt tends to decrease as people get older (Orth, Robins, & Soto, 2010). For the model used in this applied example, age was again the independent variable, predicting higher levels of disclosure through the mediator repair. We also assumed that the repair tendencies predict disclosure, and not the other way around, because the choice to disclose information, even to a researcher, needs to be motivated by some internal feeling (Tignor & Colvin, 2019).

For this applied example, the theoretical rationale was secondary to the analysis performing in a particular way to demonstrate a significant index of moderated mediation not being significant if the model is over-specified. The moderator chosen for this example was empathy, specifically because empathy as a moderator on the path between age and repair had a significant index of moderated mediation, and when empathy moderated all paths, there was no significant index of moderated mediation. Some rationale for why empathy was coded in the first place in the narratives is as follows. Empathy is a moral emotional process (Tangney, Stuewig, & Mashek, 2007), which could be detected in narrative responses, used here as the moderator because of research linking empathy to age (van Ggndy & Van Gundy, 2000). There also needs to be a justification for why the moderation is hypothesized occurs on each path, in addition to specifying the theoretical mediation model from theory (Fiedler et al., 2018). For example, the first model tested in this applied example has moderation on the path between age and repair. This is somewhat justified with the finding from Lindsey, Yun, and Hill (2007), where empathy was shown to moderate the relationship between anticipated guilt and motivation to help unknown others. For the other model in this applied example where all paths are moderated, justification for each of the three interactions would also be needed. This dissertation example is intended to show the potential impact of fitting a model with all paths moderated without clear theoretical justification for each path.

3.2.1 Data Analytic Plan

The theoretical model described above with empathy moderating the path between age and repair is Model 7 from Hayes' PROCESS Macro numbering system (Hayes, 2022). The independent variable is age, the mediator variable is repair, and the outcome variable was disclosure. This model, with empathy only moderating one path, is shown in Figure 3.1 on the left side. Then, PROCESS Model 59 was fit as an example of this model with all paths moderated by empathy. This model is shown on the right side of Figure 3.1. These analyses were done using the PROCESS macro in R with 5,000 bootstraps, and for reproducibility, the seed was set to 90263.

Model 59 is considered to be an over-specified version of Model 7. This would happen if the "true" DGP in the population is Model 7 with only one path moderated. Since this is real data instead of a simulation, that claim cannot be substantiated. Instead, this is just a demonstration of how a significant effect can be missed by using an over-specified model, and that being cautious to not miss an effect and moderating every path may not be the best option because it could potentially lead to missing a moderated effect.

3.3 Results

Model 7. A moderated mediation analysis was conducted using Hayes' PROCESS Model 7 to examine the relationships between the independent variable (age), the mediator (repair), the moderator (empathy), and the dependent variable (disclosure). The results showed a significant index of moderated mediation ($b = -0.22$, $SE_{boot} = 0.09$, $CI_{95\%} = [-0.42, -0.07]$). There was a significant effect of age on repair conditional on there being no empathy in

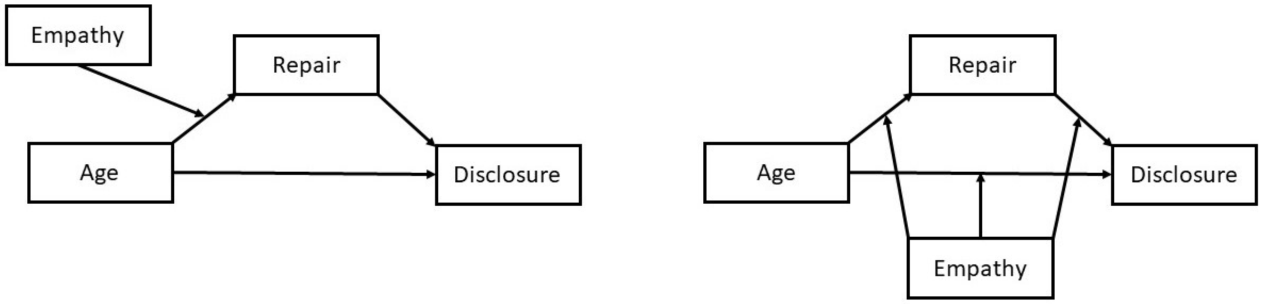

Figure 3.1: Conceptual diagrams of the moderated mediation models. The left is a conceptual diagram of Model 7 and the right is Model 59.

the narrative, $b = 0.16$, $p < .001$, $CI_{95\%} = [0.11, 0.22]$, and a significant effect of empathy on repair conditional on age being zero, $b = 3.57$, $p = .002$, $CI_{95\%} = [1.36, 5.79]$. Additionally, there was a significant interaction between age and empathy on repair, $b = -0.14$, $p = .01$, $CI_{95\%} = [-0.24, -0.03]$. There was also a significant effect of repair on disclosure, $b = 0.80$, $p < .001$, $CI_{95\%} = [0.40, 1.21]$, and age on disclosure (the direct effect), $b = 1.60$, $p < .001$, $CI_{95\%} = [0.74, 2.46]$. The conditional indirect effect of age on disclosure through repair was significant when empathy was not present in the narrative ($b = 0.16$, $SE_{boot} = 0.03$, $CI_{95\%} = [0.11, 0.22]$), but not when empathy was present ($b = 0.03$, $SE_{boot} = 0.04$, $CI_{95\%} = [-0.06, 0.12]$).

Model 59. A moderated mediation analysis was conducted using Hayes' PROCESS Model 59 to examine the relationships between the independent variable (age), the mediator (repair), the moderator (empathy), and the dependent variable (disclosure) where all paths in the mediation model are moderated by empathy. The results showed a nonsignificant index of moderated mediation ($b = -0.14$, $SE_{boot} = 0.12$, $CI_{95\%} = [-0.36, 0.11]$). There was, however, a significant effect of age on repair conditional on there being no empathy in the narrative, $b = 0.16$, $p < .001$, $CI_{95\%} = [0.11, 0.22]$, and a significant effect of empathy on repair conditional on age being zero, $b = 3.57$, $p = .002$, $CI_{95\%} = [1.36, 5.79]$. Additionally, there was a significant interaction between age and empathy on repair, $b = -0.14$, $p = .01$, $CI_{95\%} = [-0.24, -0.03]$. These results are the same as in Model 7 reported above, because the same variables are included in this first-stage moderated mediation. For Model 59, there was also a significant effect of repair on disclosure conditional on there being no empathy in the narrative, $b = 1.20$, $p = .01$, $CI_{95\%} = [0.26, 2.14]$ and age on disclosure conditional on there being no empathy in the narrative, $b = 0.96$, $p < .001$, $CI_{95\%} = [0.49, 1.42]$, but not a significant effect of empathy on disclosure conditional on age and repair being zero, $b = 10.76$, $p = .69$, $CI_{95\%} = [-41.45, 62.96]$. There was also not a significant interaction between age and empathy on disclosure, $b = -0.58$, $p = .20$, $CI_{95\%} = [-1.45, 0.30]$, or between repair and empathy on disclosure, $b = 0.91$, $p = 0.42$, $CI_{95\%} = [-1.31, 3.13]$. Based on these results, there is no evidence of moderated mediation, so no conditional indirect effects were examined.

3.4 Discussion

Overall, one moderated mediation model found support for the indirect effect of age on disclosure through repair being moderated by empathy, and the other model did not. The only difference between these models was how many paths in the mediation model were moderated. These results have important implications for the importance of not over-specifying a model without strong theoretical justification. This moderated relationship disappears when an additional interaction is included in the model, with the path between age and repair also moderated by empathy. No meaningful conclusions about moral emotions should be drawn from this applied example, as it is used only for demonstration, to show how a significant index of moderated mediation could be obscured by over-specifying a model.

This is an example of where specifying a model has the potential to be statistically significant with fewer paths moderated than when all paths are moderated. A researcher trying to "play it safe" by adding additional interactions into a mediation model, when those additional interactions are not supported by theory, might be inadvertently causing a Type II error and missing an effect that is truly present by over-specifying the model. Based on the literature and this example, I do not recommend fitting more models than the hypothesized models and only reporting the significant ones, because that is a QRP (Fiedler & Schwarz, 2016; John et al., 2012). With so many published moderated mediation models using Models 58 and 59, with two or three paths moderated, there is a risk of true effects being missed. This will be explored further in the following simulation study, investigating this finding further to see if over-specified models tend to have decreased power compared to correctly specified models.

Conclusions

The findings show that a significant index of moderated mediation could be present in one model but not in another, depending on where the moderation occurs. Over-specifying a model unnecessarily could potentially decrease power. PROCESS Model 7 and Model 59 were shown using the same moral emotions data and found a significant index of moderated mediation only in the model with one path moderated and not in the model with all paths moderated. Model 59 would be an over-specified version of Model 7 in this dissertation, but there is no way to know which model reflects the true underlying mechanisms in the population based on the statistical results. Due to this limitation, it is not recommended to add moderation on a path simply to avoid potentially missing a significant result. However, there is a risk then of under-specifying the model, and implications for this are explored further in the simulation study. This result illustrates the importance of specification for moderated mediation models.

4 Simulation Study

The goal of the simulation study in this dissertation was to understand how over-specifying, under-specifying, and completely misspecifying moderated mediation models affect statistical power and type I error rate. There were two core 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). Whether the data analysis model is an over-specification, under-specification, or complete misspecification of the DGP is listed in Table 4.1. Over-specified data analysis models include moderation on all the paths moderated in the DGP, plus at least one additional path. Under-specified data analysis models have at least one path included in the indirect effect moderated as in the DGP, but the data analysis model does not include all the moderated paths from the DGP. The data analysis model could include moderation on additional paths. Completely misspecified data analysis models include moderation on exactly one path in the indirect effect, but it is not the same one path moderated in the indirect effect as in the DGP. This simulation study examines the effect of model over- and under-specification on power and the effect of complete model misspecification on type I error rate.

Table 4.1: 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.

4.1 Method

This Monte Carlo simulation varied six factors to explore the effects of model specification on power and type I error rate. I generated data using each one of the six DGPs, then fit the data using all six data analysis models, one of which is correctly specified. 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). Simulated data analysis results were analyzed using multilevel logistic regression with random intercepts to predict rejection of the null hypothesis that the index of moderated mediation was 0, indicating either power or type I error rate, depending on the type of specification. Model over- and under-specification were hypothesized to have lower power than correct specification, and complete misspecification was hypothesized to have an inflated type I error rate.

4.1.1 Simulation Conditions

This dissertation used a Monte Carlo simulation with an incomplete 6 (Between: Generating Model) x 9 (Between: Sample Size) x 3 (Between: Effect Size Magnitude) x 2 (Between: Normal or Dichotomous X) x 2 (Between: Normal or Dichotomous W) x 6 (Within: Analysis Model) factorial design. Table 4.2 lists each condition and the levels used. The design is incomplete because Model 58 and 59 were only used as generating and/or analysis models when W was dichotomous because the index of moderated mediation is undefined when W is continuous.

4.1.2 Simulation Procedure

I used GAUSS 21 on a Windows server for data generation, generating 1,000 samples of data in each condition. Based on deciles (with rounding) from the meta-analytic review, using the 10^{th} and 90^{th} percentiles (inter-decile range) as the maximum and minimum, I chose nine sample sizes to include in this study to best represent typical sample sizes used in

Table 4.2: 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 | 1,000
Effect Size Magnitude (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 for data generation and data analysis when W was dichotomous.

moderated mediation studies. Sample sizes included were 100, 150, 200, 250, 300, 400, 500, 750, and 1,000. Four variables were generated in a piecewise manner: the primary predictor X, the moderator W, the mediator M, and the primary outcome Y. In all cases, X and the moderator W were independent and generated randomly from either a standard normal distribution (continuous cases) or assigned with equal allocation to each level (dichotomous cases).

Based on the DGP, Equations from Table 1.1 were used to generate M and Y. 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).

In each data generation equation, the coefficients were determined as follows. I 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. When multiple interactions were included in the model, they were all set to be the same size and direction[1]. Additionally, when W is included in 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

[1]This caused multi-interaction DGPs to have a higher overall effect size.

by taking the square root of these R^2 effect sizes. For example, the path coefficient is $\sqrt{.07} = .26$ when X and M are standardized. Residuals for both the model predicting M and predicting Y were generated from a normal distribution centered at 0, with the standard deviation incorporating the path coefficients included in the equation to ensure the proportion of explained variance remains as expected and the standard deviation of the outcome is always 1 (i.e., standardized)[2]. For example, the standard deviation of the residuals in the equation predicting M when the X to M path was moderated 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. I 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), OLS is used throughout to estimate coefficients. The variance of the product term was always equal to one when the predictor variables were independent, based on how the predictor variables were defined to always have a variance of one and a mean of zero. Variables X and W were generated independently to have a covariance of zero. When interacting variables both have a mean of zero, the variance of the product term is the product of both individual variances (Bohrnstedt & Goldberger, 1969). Since both variances are one, the variance of the product is also one. To keep the variance equal to one in the dichotomous case, the two categories were coded as -1 and 1 because the variance of that variable where the categories are equally allocated is one. For this reason, equal allocation was always used in the dichotomous conditions.

Once data generation was complete, data analysis models were fit to each sample of generated data. Each sample was analyzed with all four (continuous W) or six (dichotomous W) analysis models. Models were analyzed using the percentile bootstrap confidence interval set at 95% with 1,000 bootstraps, done for each of the 1,000 samples of data, with the decision of whether or not to reject the null hypothesis based on the confidence interval made for

[2]This is only true in cases where the predictor variables are independent. For these moderated mediation models, the variance of M was 1, but not Y. This could be part of the reason why results show greater effects of second-stage moderation than first-stage moderation.

each sample (Efron & Tibshirani, 1994).

4.1.3 Performance Metrics

There were two outcomes of interest in this study: statistical power and type I error rate for the index of moderated mediation. Both were calculated as the proportion of the 1,000 generated samples that had a statistically significant result in each condition, which is the rejection rate. Whether the resulting rejection rate is considered power or type I error depended on the specification of the analysis model: correctly specified (power), over-specified (power), under-specified (power), or completely misspecified (type I error). Table 4.1 shows which data analysis models would be considered an over-specification, under-specification, or complete misspecification from the DGP. I calculated the 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 was correctly specified, over-specified, or under-specified. Correctly specified models should accurately detect effects, providing a baseline power level that can be used to compare to the over- or under-specified 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 based on how we defined over- and under-specification, so a significant effect is detecting an effect that truly exists from the DGP.

Type I error rate was calculated in the same way as power, but for models where the data analysis model was completely misspecified. I only included completely misspecified models because these were 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 often differ from 0.05 for correctly specified models (Yzerbyt et al., 2018), the criteria from Bradley (1978) and Serlin (2000) were used to classify type I error rates as overly conservative or liberal (.025 to .075, and .035 to .065, respectively).

4.2 Hypotheses

Hypothesis 1: Over-specification and Power

- H1a: I hypothesized that the statistical power of the index of moderated mediation will be lower for over-specified models compared to correctly specified models.
- H1b: I also hypothesized that power will be lower for over-specified data analysis models with additional moderated paths. For example, if the DGP is Model 7, then Models 8, 58, and 59 are all over-specified. Models 8 and 58 have two moderated paths but Model 59 has three. I hypothesized that for cases like this, for models with more moderated paths (Model 59), power will be lower than models with fewer moderated paths (Models 8 and 58).

Hypothesis 2: Under-specification and Power

- H2a: I hypothesized that the statistical power of the index of moderated mediation would be lower for under-specified models compared to correctly specified models.
- H2b: I also hypothesized that under-specified models would have higher parameter bias than correctly specified models.

Hypothesis 3: Complete Misspecification and Type I Error Rate

- H3a: I hypothesized that the type I error rate would be too high according to criteria set by (Bradley, 1978) and (Serlin, 2000) in completely misspecified models.

- H3b: I also hypothesized that the type I error rate would be higher in cases where the analysis model includes two moderated paths (i.e., Models 8 and 15) compared to one (i.e., Models 7 and 14).

4.2.1 Analysis Plan

To test several of the hypotheses (specifically Hypotheses 1a, 1b, 2a, and 3b as these require inference), I used multilevel logistic regression with random intercepts for the within-subjects factor of data analysis model to predict rejection (either power or type I error rate). The rejection was a binary 0/1 indicator from the simulation where 0 indicates the confidence interval includes zero, and 1 indicates zero was excluded from the confidence interval. Rejection rate was the proportion of samples that were rejected. I fit two separate models for each hypothesis requiring inference that included Models 58 and 59 because it was an incomplete design: one model for continuous W and one model for dichotomous W (since Models 58 and 59 were only used as generating and analysis models when W was dichotomous). As such, Hypotheses 1a, 1b, and 2a will have two results reported in the results section: one from the continuous W model, and one from the dichotomous W model. For Hypothesis 3b, only one model was fit because Models 58 and 59 were completely excluded because those two models could never be completely misspecified.

For each coefficient, both statistical significance and effect size were determined. Statistical significance was determined by the p value on the coefficient from the logistic regression. Only coefficients with a p value less than .001 were considered to be statistically significant. Additionally, because it was predicted that many coefficients would be statistically significant, I also report the odds ratio as a measure of effect size. Wald chi-square tests were used to test sets of fixed effects when the factor had more than two categories. Parameter bias was also calculated for the index of moderated mediation as the estimated value minus the true population value set in the simulation. Parameter bias is reported for over-, under-, and completely misspecified models. Tests for specific hypotheses are described below.

To test Hypotheses 1a-b, I only included over-specified and correctly specified models. The multilevel logistic regression had six main effects: analysis model specification (over- vs. correct), sample size (sequentially coded), number of moderated paths in the analysis model (sequentially coded), generating model (dummy coded), magnitude of the effect size (sequentially coded), and type of X (continuous vs. dichotomous).

- H1a: This hypothesis was tested specifically by looking at the main effect of the *over-specified vs. correctly specified model indicator* variable.

- H1b: This hypothesis was tested specifically by looking at the main effect of the *number of moderated paths* variable and using a table.

To test Hypotheses 2a-b, I only included under-specified and correctly specified models. The multilevel logistic regression (only required for Hypothesis 2a) also had six main effects: analysis model specification (under- vs. correct), generating model (dummy coded), sample size (sequentially coded), magnitude of the effect size (sequentially coded), type of X (continuous vs. dichotomous), and number of moderated paths in the analysis model (sequentially coded). Two separate models were again fit, one for continuous W and one for dichotomous W, producing two results for each hypothesis.

- H2a: This hypothesis was tested specifically by looking at the main effect of the *under-specified vs. correctly specified model indicator* variable.

- H2b: This hypothesis was tested by comparing the parameter bias of the index of moderated mediation between under-specified models and correctly specified models. Using criteria from Forero, Maydeu-Olivares, and Gallardo-Pujol (2009), relative bias below 10% was considered acceptable. Values from 10% to 20% indicated substantial bias, whereas those above 20% were deemed unacceptable.

To test Hypotheses 3a-b, I only included completely misspecified models. The multilevel logistic regression (only required for Hypothesis 3b) had six main effects: the number of moderated paths in the analysis model (one vs. two), generating model (dummy coded), sample size (sequentially coded), the magnitude of the effect size (sequentially coded), type of X (continuous vs. dichotomous), and type of W (continuous vs. dichotomous). Instead of two separate models for continuous vs. dichotomous moderators like the models predicting power, the type of W variable was included as a main effect because Models 58 and 59 were never completely misspecified analysis models.

- H3a: This hypothesis was tested by comparing the type I error rate from each condition of complete misspecification to the criteria set by Bradley (1978) and Serlin (2000).
- H3b: This hypothesis was tested specifically by looking at the main effect of the *number of moderated paths in the analysis model* variable.

4.3 Results

The following results are presented by hypothesis because each hypothesis focused on a different type of misspecification. The goal was to understand which factors had the potential to increase or decrease statistical power, or increase or decrease type I error rate in cases of complete model misspecification. For Hypothesis 1, which focused on over-specified models compared to correctly specified models, the outcome was statistical power. For Hypothesis 2, which focused on under-specified models compared to correctly specified models, the outcome was also statistical power. For Hypothesis 3, which focused on completely misspecified models, the outcome was type I error rate. From the simulated data, each case represents a single data analysis model fit on a sample of generated data, and the data analysis model may or may not match the DGP. For each hypothesis, I fit multilevel logistic regression models to the raw simulation results predicting rejection from each simulated sample of data in each condition, to examine the effects of the manipulated factors in the simulation. In each model, the outcome was the rejection of the null hypothesis (null hypothesis: the index of moderated mediation is 0), indicating either statistical power or type I error rate depending on how the data analysis model was specified compared to the DGP. Each multilevel logistic regression model produced coefficients in log odds, and to test for statistical significance any p values less than .001 were considered statistically significant. To report an effect size, odds ratios were created by exponentiation of the coefficients in log odds. Several positive checks of the simulation results were included in all models and are presented here before the full simulation results.

Positive Checks

Over-specified Models As a positive check for the simulation data, sample size was as expected a significant predictor of statistical power with both continuous and dichotomous moderators, controlling for analysis model specification, number of moderated paths in the analysis model, DGP, magnitude of the effect size, and type of X variable. A Wald test for the set of fixed effects corresponding to sample size (sequentially coded) was significant for both the continuous moderator model and dichotomous moderator model, $\chi^2 = 711.18$, $p < .001$ and $\chi^2 = 89.79$, $p < .001$, respectively. All of the coefficients from the model had odds ratios above 1, indicating that the power to detect a significant indirect effect is always higher at higher sample sizes. As the power curves in the figures show, power increases most drastically with regards to sample size when sample sizes are lower, then tends to level off at higher sample sizes.

Additionally, the magnitude of the effect size was indeed a significant predictor of statistical power with both continuous and dichotomous moderators, controlling for analysis model specification, sample size, number of moderated paths in the analysis model, DGP, and type of X variable. A Wald test for the set of two fixed effects corresponding to effect size (sequentially coded for medium compared to small and large compared to medium) was significant for both the continuous moderator model and dichotomous moderator model, $\chi^2 = 799.32$, $p < .001$ and $\chi^2 = 781.26$, $p < .001$, respectively. All of the coefficients from the model had odds ratios above 1, indicating that power to detect a significant indirect effect is always higher at higher effect sizes. This is shown in Figure 4.1 and in Figure 4.2 by the rows, with larger effect sizes having higher power. This is clearest in the correctly specified models because only one data analysis model was used.

Under-specified Models Sample size was as expected a significant predictor of statistical power in both models in both the continuous and dichotomous moderator models for under-specified and correctly specified models, controlling for analysis model specification, number of moderated paths in the analysis model, DGP, magnitude of the effect size, and type of X variable. A Wald test for the set of fixed effects corresponding to sample size (sequentially coded) was significant for both the continuous moderator model and dichotomous moderator model, $\chi^2 = 770.65$, $p < .001$ and $\chi^2 = 1227.90$, $p < .001$, respectively. All of the coefficients from the model had odds ratios above 1, indicating that power to detect a significant indirect effect is always higher at higher sample sizes. As the power curves in the figures show, power still increases most drastically with regards to sample size when sample sizes are lower, then tends to level off at the higher sample sizes, though each coefficient on sample size remains statistically significant even at the highest sample sizes.

Additionally, the magnitude of the effect size was indeed a significant predictor of statistical power in both models, controlling for analysis model specification, sample size, number of moderated paths in the analysis model, DGP, and type of X variable. A Wald test for the set of two fixed effects corresponding to effect size (sequentially coded for medium compared to small and large compared to medium) was significant for both the continuous moderator model and dichotomous moderator model, $\chi^2 = 915.02$, $p < .001$ and $\chi^2 = 807.60$, $p < .001$, respectively. All of the coefficients from the model had odds ratios above 1, indicating that power to detect a significant indirect effect is always higher at higher effect sizes. This is shown in Figure 4.3 and in Figure 4.4 by the rows, with larger effect sizes having higher power. This remains clearest in the correctly specified models because only one data analysis model was used. The pattern still holds in Figure 4.4, but for Models 8 and 15 there are two possible under-specifications, so power is higher for higher effect sizes within each under-specified analysis model. For Models 58 and 59, where there are four or five possible under-specifications, respectively, effect size magnitude is more clearly separated regardless of the under-specified analysis model.

Completely Misspecified Models Sample size was as expected a significant predictor of type I error rate, controlling for the number of moderated paths in the analysis model, DGP, magnitude of the effect size, and type of X variable. A Wald test for the set of fixed effects corresponding to sample size (sequentially coded) was significant, $\chi^2 = 203.65$, $p < .001$. However, only the first two sequentially coded sample size comparisons were significantly different from the previous sample size, indicating that this is more of an issue with smaller sample sizes.

Additionally, the magnitude of the effect size was indeed a significant predictor of type I error rate, controlling for analysis model specification, sample size, number of moderated paths in the analysis model, DGP, and type of X variable. A Wald test for the set of two fixed effects corresponding to effect size (sequentially coded for medium compared to small and large compared to medium) was significant, $\chi^2 = 46.69$, $p < .001$. Both coefficients from the model had odds ratios above 1, indicating type I errors are more common at higher effect sizes. However from Figure 4.5 which separates out effect sizes by row, this statistical effect could be driven mainly by the inflated type I error rates of using Model 14 as the analysis model when the DGP was Model 8, where the difference is even more pronounced at larger effect sizes.

4.4 Statistical Power in Over-specified Models

The first hypothesis examined factors affecting statistical power in over-specified moderated mediation models compared to correct specification. The factors tested in this model included model specification, DGP, sample size, the magnitude of the effect size, type of X, and the number of moderated paths in the analysis model. I fit two separate multilevel logistic regression models, one for continuous W and one for dichotomous W. Table 4.3 shows that power for models with continuous moderators compared to dichotomous moderators is similar. There were 324,000 cases in the continuous model, and 810,000 cases in the dichotomous model. The results from both models were similar, with a few notable exceptions described in this section. All numerical results in this section will be presented first for the continuous moderator model, then for the dichotomous moderator model. Results are separated in Figure 4.1 and Figure 4.2 for continuous and dichotomous moderators, respectively. In both figures, the columns represent the DGP, and the rows represent the effect size of the interaction(s). Sample size is displayed on the horizontal axis, and the outcome of power is displayed on the vertical axis. The results are split up by the size of the model specification by color (grey for correct specification; pink for over-specification) and analysis model (shape).

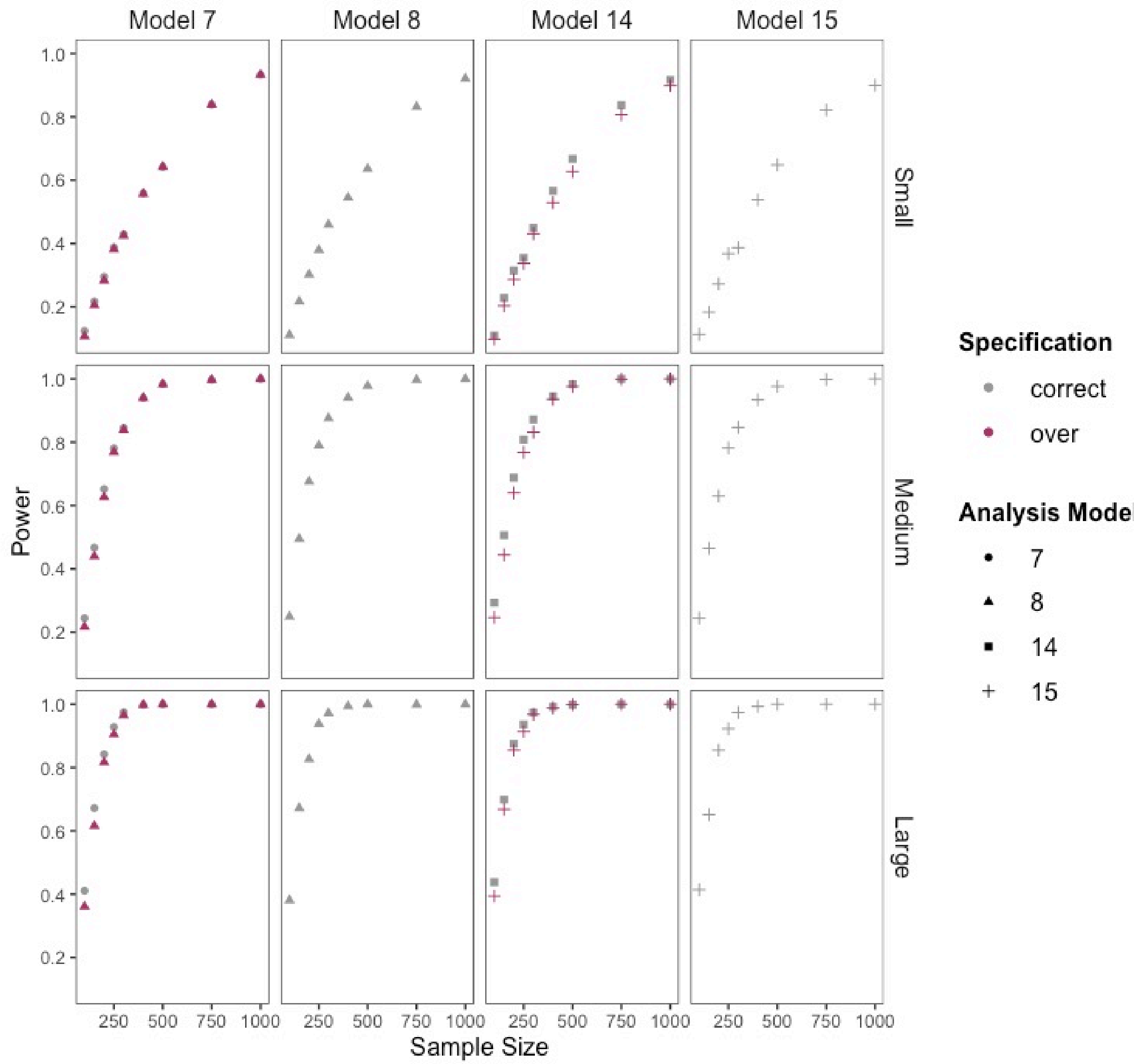

Figure 4.1: Figure displaying statistical power for moderated mediation models with a **continuous moderator**. The columns represent the DGP, and the rows represent effect size of the interaction(s). Sample size is displayed on the horizontal axis, and the outcome of power is displayed on the vertical axis. The results are split up by size of the model specification by color (grey for correct specification; pink for over-specification) and analysis model (shape). For clarity, the figure is shown for continuous X only. Additional figures are provided in the Appendix with dichotomous X.

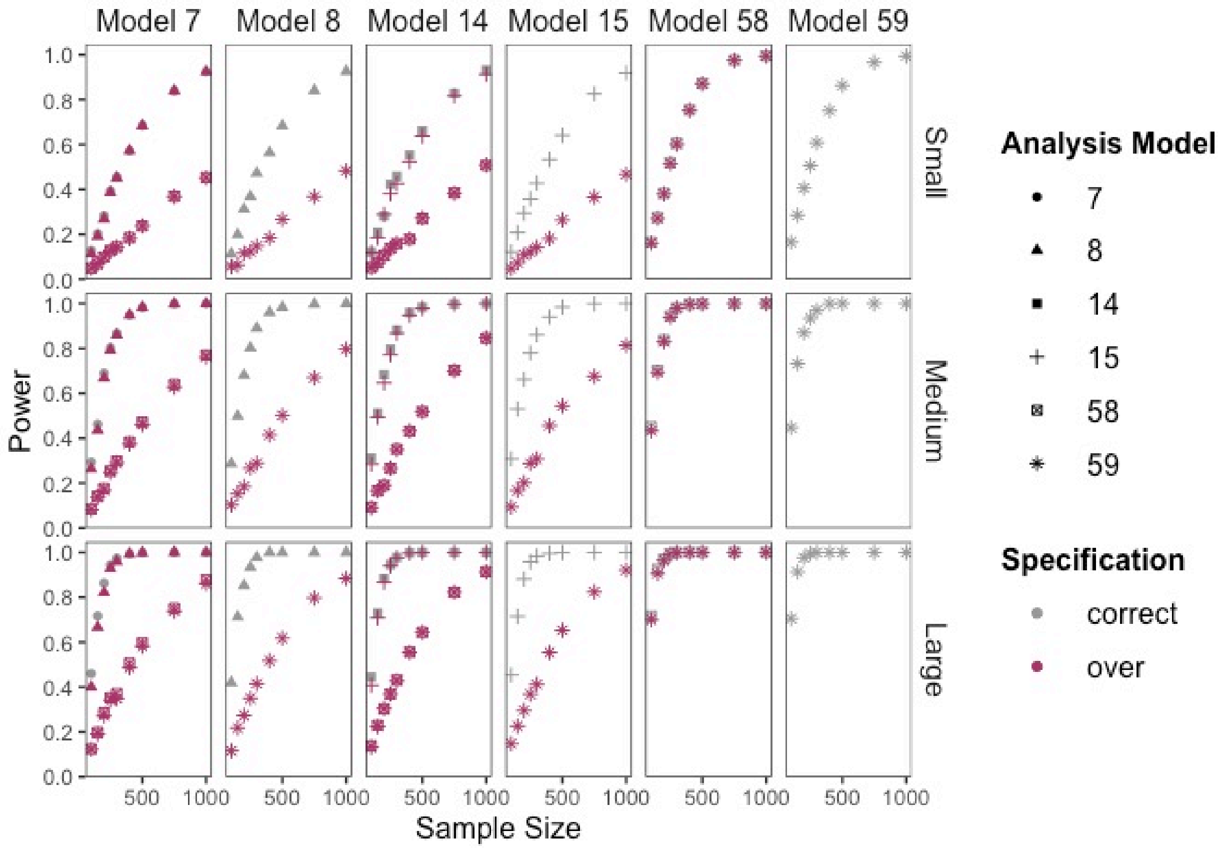

Figure 4.2: Figure displaying statistical power for moderated mediation models with a **dichotomous moderator**. The columns represent the DGP, and the rows represent effect size of the interaction(s). Sample size is displayed on the horizontal axis, and the outcome of power is displayed on the vertical axis. The results are split up by size of the model specification by color (grey for correct specification; pink for over-specification) and analysis model (shape). For clarity, the figure is shown for continuous X only. Additional figures are provided in the Appendix with dichotomous X.

H1a: Model Specification Hypothesis 1a was supported. Model specification was a significant predictor of statistical power in both the analysis using continuous moderators and dichotomous moderators, controlling for sample size, number of moderated paths in the analysis model, DGP, magnitude of the effect size, and type of X variable. For over-specified models compared to correctly specified models, we saw a significant difference in power, $b = -0.14$, $CI_{95\%}[-0.16, -0.11]$, $p < .001$, $OR = 0.87$ in the continuous moderator model, and $b = -0.69$, $CI_{95\%}[-0.72, -0.66]$, $p < .001$, $OR = 0.50$ in the dichotomous moderator model, holding all other variables constant. Power is lower for over-specified models compared to correctly specified models. With continuous moderators, the largest difference in power between a correctly specified model and an over-specified model was trivial: .896 compared to .891, respectively, occurring when the sample size was set to 300, X was continuous, the effect size was medium, the over-specified data analysis model was Model 8, and the DGP was Model 7. In contrast, with dichotomous moderators, the difference in power was much larger. The biggest difference happened between Model 14 as the DGP and Model 59 as the data analysis model, when the sample size was 300, X was dichotomous, and the effect size was medium. The correctly specified model had .980 power, while Model 58 had .320 power, and Model 59 had .308 power. Power between Model 58 and 59 when both were over-specified tended to have the smallest differences overall.

While the odds ratios seem to indicate the effect with dichotomous moderators is more salient, this appears to be confounded by the fact that the dichotomous moderators include Models 58 and 59, which seem to have the biggest decrease in power when used as the data analysis model in an over-specification. Looking at Figure 4.1, the power curves for the over-specified models look similar to the correctly specified models, though slightly lower. In Figure 4.2, the difference in power appears most prominent when the data analysis model is Models 58 or 59 and it is over-specified. Also, based on the figures, when Models 14 and 15 are used for data analysis and they are over-specified (i.e., the DGP is Model 7 or 14, respectively), power is not as affected as when Models 58 or 59 are used for data analysis. In

Figure 4.2, Model 59 does not appear to have much lower power than Model 58 when both are over-specifications (seen when the DGP is Model 7 or Model 14).

A subset of power estimates from the simulation is shown in Table 4.3, where power is shown for every combination of analysis model (column) and DGP (row). Power for both continuous moderators (top) and dichotomous moderators (bottom) are shown. For clarity, only medium interaction effect sizes are shown, with continuous X and a sample size of 300 since it was the median sample size. Consistent with the results from the analyses, over-specified models all have lower power than correctly specified models. The difference in power is greatest for Models 7, 8, 14, and 15 when compared to Models 58 and 59 when Models 58 and 59 are over-specifications. Over-specifying Model 59 when the DGP was Model 58 did not show a decrease in power.

H1b: Number of Moderated Paths Hypothesis 1b was supported. The number of moderated paths in the analysis model was a significant predictor of statistical power in both models, controlling for analysis model specification, sample size, DGP, the magnitude of the effect size, and type of X variable. Since Model 59 is the only model with three paths moderated, and it was excluded from the continuous moderator analyses, the only test for this was the comparison between two moderated paths and one moderated path in the analysis model. This test was significant, $b = -0.49$, $CI_{95\%}[-0.52, -0.46]$, $p < .001$, $OR = 0.61$, indicating that power to detect a significant index of moderated mediation is lower with more moderated paths in the data analysis model. With dichotomous moderators, Model 59 was included, so the number of paths moderated in the analysis model could be 1, 2, or 3. A Wald test for the set of fixed effects corresponding to the number of paths in the analysis model was significant in the dichotomous moderator model, $\chi^2 = 5587.00$, $p < .001$, also indicating that power is lower with more moderated paths in the analysis model.

Type of X There was no significant difference in power for continuous or dichotomous predictor X, in either the model with a continuous moderator, $b = -0.09$, $p = 0.39$,

			Analysis Model			
DGP	Model 7	Model 8	Model 14	Model 15	Model 58	Model 59
7	.846	**.839**				
	.868	**.860**			**.298**	**.291**
8	*.896*	.891				
	.896	.891			*.349*	**.289**
14			.872	**.832**		
			.881	**.868**	**.352**	**.349**
15			*.958*	.847		
			.980	.861	*.324*	**.308**
58						
	.892	*.881*	*.866*	*.842*	.980	**.980**
59						
	.881	*.879*	*.956*	*.810*	*.977*	.970

Table 4.3: Table showing statistical power (proportion of correctly rejected hypothesis tests for the index of moderated mediation) from the simulation. Power for continuous moderators (top number in each cell) and dichotomous moderators (bottom number in each cell) are shown. The columns represent the data analysis model, and the rows represent the DGP. All power is for continuous W with a medium interaction effect size at sample size 300. Completely missing cells are complete model misspecification where power was not reported, and cells missing the continuous power value were undefined (either the data analysis model or DGP was Model 58 or 59). Power for **over-specified models are bold** and power for *under-specified models are italicized*. The diagonal is correct specifications.

$CI_{95\%} = [-0.30, 0.12]$, $OR = 0.91$ or dichotomous moderator, $b = -0.02$, $p = 0.83$, $CI_{95\%} = [-0.17, 0.13]$, $OR = 0.98$. Due to there being no significant difference, only continuous X is shown in Figure 4.1 and Figure 4.2. Additional figures showing dichotomous X as well are in the Appendix.

Parameter Bias Parameter bias was only slightly worse for over-specified models compared to correctly specified models. All relative bias values were well under the threshold for acceptable bias of under .1 (Forero et al., 2009). Average relative bias across all sample sizes and effect sizes for over-specified models ranged from -0.0027 to 0.0022, compared to correctly specified models which ranged from -0.0031 to 0.0026. Relative bias was highest for over-specified models at the smallest effect size, where the highest bias was seen with Model 58 as the data analysis model and Model 7 as the DGP, at -0.011. Relative bias was also highest for over-specified models at smaller sample sizes, where there the bias with Model 58 as the data analysis model and Model 7 as the DGP was -0.014.

Exploratory: Interactions Model A model containing all possible two-way through six-way interactions was fit, but none of the interactions were significant at the $\alpha = .001$ level for either the continuous moderator model or the dichotomous moderator model.

4.5 Statistical Power in Under-specified Models

The second hypothesis examined factors affecting statistical power in under-specified moderated mediation models compared to correct specification. The factors tested in this model included model specification, DGP, sample size, magnitude of the effect size, type of X, and number of moderated paths in the analysis model. I again fit two separate multilevel logistic regression models, one for continuous W and one for dichotomous W. Table 4.3 shows that power for models with continuous moderators compared to dichotomous moderators are similar. There were 324,000 cases in the continuous model, and 1,026,000 cases in the dichotomous model. The results from both models were again similar, with a few notable exceptions described in this section. Figure 4.3 and Figure 4.4 also contain visualizations for under-specified models in the bottom row.

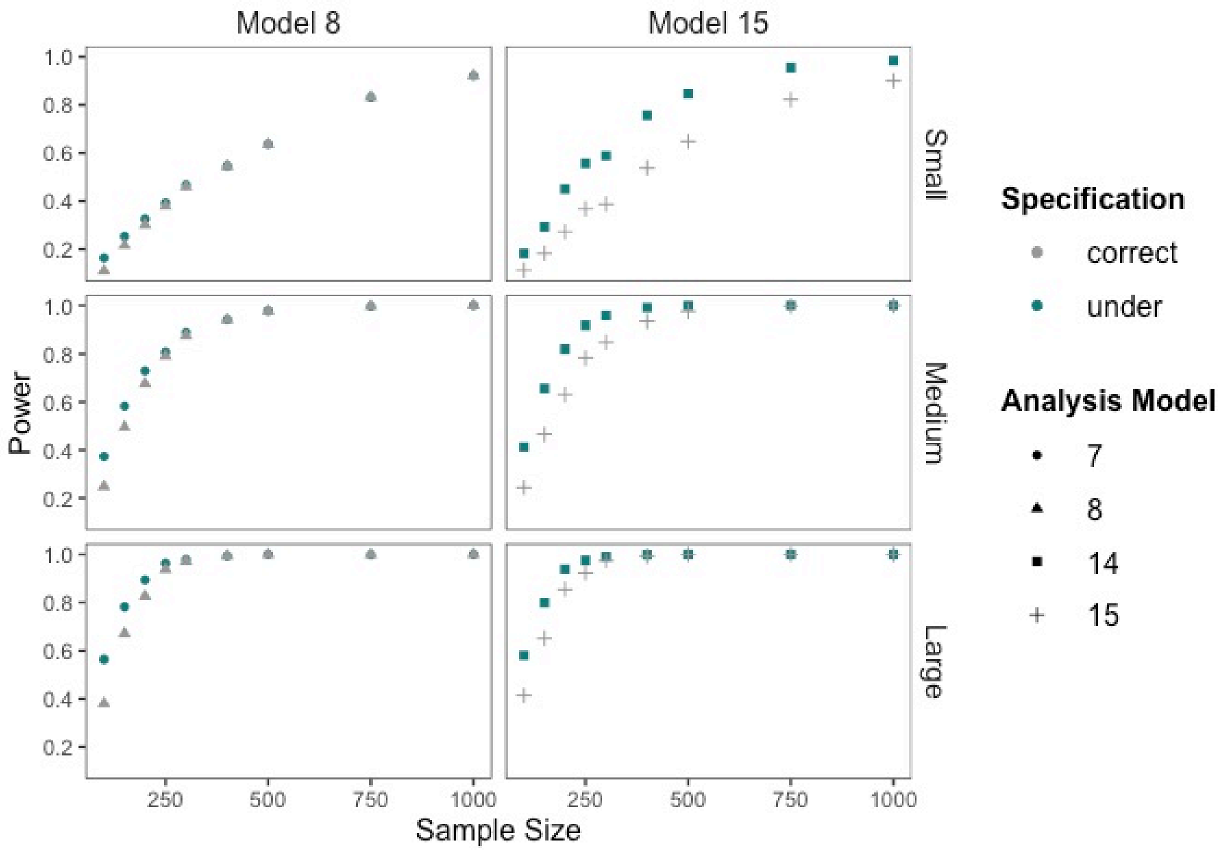

Figure 4.3: Figure displaying statistical power for moderated mediation models with a **continuous moderator**. The columns represent the DGP, and the rows represent effect size of the interaction(s). Sample size is displayed on the horizontal axis, and the outcome of power is displayed on the vertical axis. The results are split up by size of the model specification by color (grey for correct specification; teal for under-specification) and analysis model (shape). For clarity, the figure is shown for continuous X only. Additional figures are provided in the Appendix with dichotomous X.

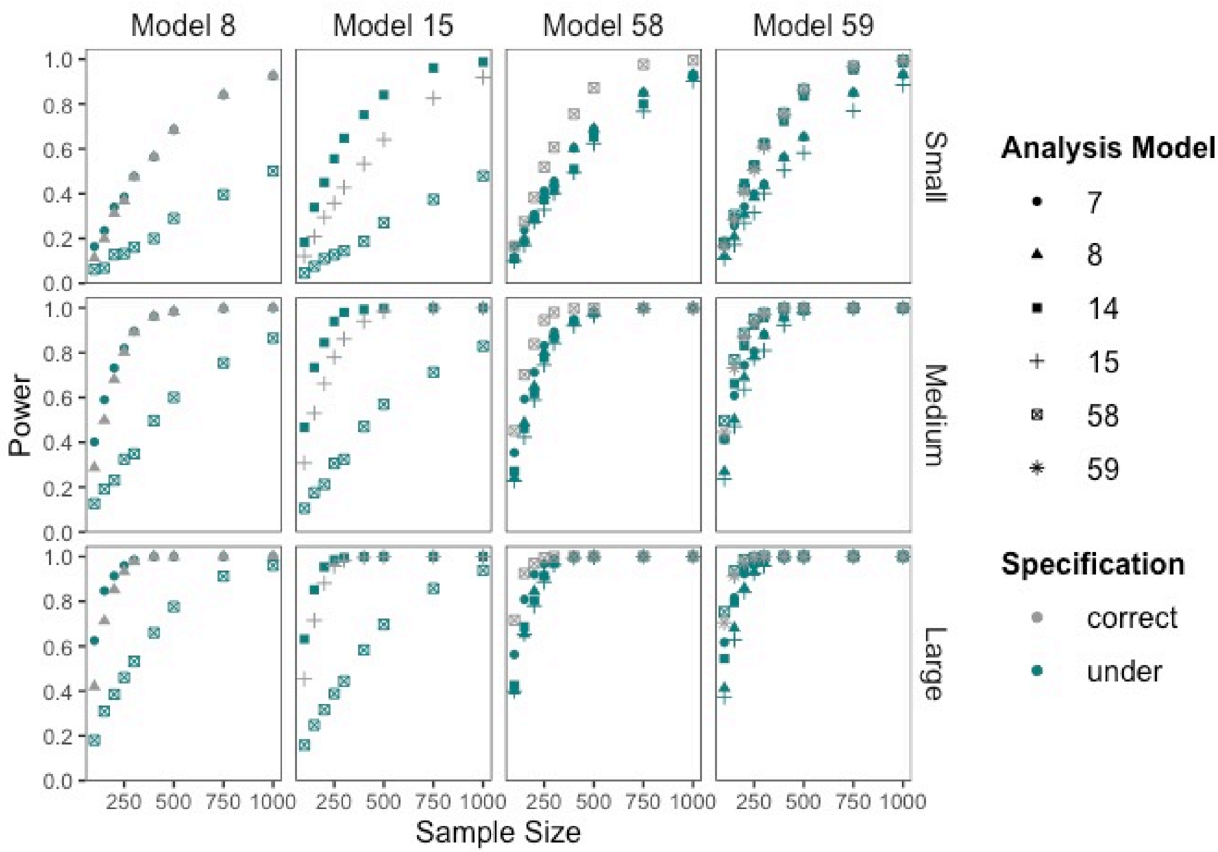

Figure 4.4: Figure displaying statistical power for moderated mediation models with a **dichotomous moderator**. The columns represent the DGP, and the rows represent effect size of the interaction(s). Sample size is displayed on the horizontal axis, and the outcome of power is displayed on the vertical axis. The results are split up by size of the model specification by color (grey for correct specification; teal for under-specification) and analysis model (shape). For clarity, the figure is shown for continuous X only. Additional figures are provided in the Appendix with dichotomous X.

H2a: Model Specification Hypothesis 2a had mixed support from the multilevel logistic regression models. Model specification was a significant predictor of statistical power in the continuous moderator model, $b = 0.55, p < .001$, $CI_{95\%} = [0.53, 0.57]$, $OR = 1.73$, controlling for sample size, number of moderated paths in the analysis model, DGP, magnitude of the effect size, and type of X variable. This suggests that the odds of detecting a significant index of moderated mediation are 1.73 times higher for under-specified models compared to correctly specified models. For under-specified models compared to correctly specified models in the dichotomous moderator model, $b = -1.41, p < .001$, $CI_{95\%} = [-1.42, -1.39]$,

$OR = 0.25$. This suggests that the odds of detecting a significant index of moderated mediation are lower for under-specified models compared to correctly specified models.

While Hypothesis 2a had mixed results based on the odds ratios, a more nuanced pattern emerged when looking at the findings by moderated mediation model. In Figure 4.3, when Model 8 was under-specified (Model 7), power tended to be the same for both specifications, though at the smallest sample sizes, the under-specified model appeared to have slightly higher power. However, when Model 15 was under-specified (Model 14), the under-specified model had higher power at most sample sizes. When the DGP has moderation in the direct effect, but the data analysis model only allows for moderation on the path involved in the indirect effect predicting Y, we see higher statistical power. Figure 4.4, the same patterns from the continuous moderator models are visible. In addition, much larger differences in power are shown with the inclusion of Models 58 and 59. Most noticeably, when Models 8 and 15 were under-specified and Model 58 was used as the analysis model, power was much lower.

This is consistent with the subset of results shown in Table 4.3. When Model 58 was an under-specification of Model 8 or Model 15, the greatest decrease in power was shown, going from .891 to .349 or from .861 to .324, respectively. When Model 58 or 59 was under-specified, power was also lower (except when Model 59 was under-specified as Model 58, where power went from .970 to .977). Under-specification showed increased power when the analysis model matched the DGP but omitted moderation on the direct effect (i.e., Model 7 as the data analysis model when Model 8 was the DGP, or Model 14 as the data analysis model when Model 15 was the DGP). However, these increases in power were very small, with gains between .005-.119. These were the only under-specifications found with a continuous moderator, so the coefficient indicating higher power for under-specified models came from this pattern. This same pattern was seen with dichotomous moderators as well, though much larger differences in power with Models 58 and 59 included and are likely responsible for the coefficient indicating decreased power with under-specified models.

H2b: Parameter Bias Hypothesis 2b was supported. The average relative bias for under-specified models ranged from -0.003 to 0.447. The only three model specifications at an acceptable level of relative bias under .1 were all under .01 as well, and these were when the data analysis model was Model 14 and the DGP was Model 58, or the data analysis model was Model 15 and the DGP was Model 58 or 59. The relative bias was negative in each case, indicating that the estimated index of moderated mediation was slightly lower than the true parameter. In these cases with low bias, the data analysis model only omits moderation of the X to M path. All other model combinations had bias over .2, which is unacceptable according to guidelines set by Forero et al. (2009). The highest bias (average relative bias = 0.447) occurred using Model 7 as the analysis model and Model 59 as the DGP. The average relative bias when Model 14 was used for data analysis and Model 59 was the DGP was only 0.289, which is similar to the cases with acceptable relative bias, except for omission of moderation of the direct effect. The next highest bias was with Model 7 as the analysis model and Model 58 as the DGP (average relative bias = 0.313). Relative bias generally worsens at larger effect sizes, with the highest average relative bias when Model 7 was used for data analysis and Model 59 was the DGP was 0.541 (and 0.354 with a small effect). Sample size appears to matter less, though relative bias is largest at smaller sample sizes, with the average relative bias when Model 7 was used for data analysis and Model 59 was the DGP being 0.475 when the sample size was 100, and 0.433 when the sample size was 1,000.

Number of Moderated Paths Number of moderated paths in the analysis model was a significant predictor of statistical power in both models, controlling for analysis model specification, sample size, DGP, magnitude of the effect size, and type of X variable. Here again, since Model 59 is the only model with three paths moderated, it was excluded from the continuous moderator analyses. The only test for this was the comparison between two moderated paths and one moderated path in the analysis model. This test was significant,

$b = -1.12$, $CI_{95\%}[-1.14, -1.11]$, $p < .001$, $OR = 0.33$, indicating that power to detect a significant index of moderated mediation is lower with more moderated paths in the data analysis model. A Wald test for the set of fixed effects corresponding to the number of paths in the analysis model was significant in the dichotomous moderator model, $\chi^2 = 6797.00$, $p < .001$. Similar to the over-specified model, both of the individual coefficients had odds ratios below 1, indicating that the power to detect a significant indirect effect is lower with more moderated paths in the data analysis model.

Type of X There was again no significant difference in power for continuous or dichotomous predictor X, in either the model with a continuous moderator, $b = -0.10$, $p = .34$, $CI_{95\%} = [-0.29, 0.10]$, $OR = 0.91$ or dichotomous moderator, $b = -0.05$, $p = .54$, $CI_{95\%} = [-0.22, 0.12]$, $OR = 0.95$.

Exploratory: Interactions Model A model containing all possible two-way through six-way interactions was fit, but none of the interactions were significant at the $\alpha = .001$ level for either the continuous moderator model or the dichotomous moderator model.

4.6 Type I Error Rate in Completely Misspecified Models

The third hypothesis examined factors affecting statistical type I error rate in completely misspecified models. The factors tested in this model included the number of moderated paths in the analysis model, DGP, sample size, magnitude of the effect size, type of X variable, and type of W variable. For the outcome of type I error rate, I only fit one multilevel logistic regression model, because the index of moderated mediation was always defined (Models 58 and 59 were never considered completely misspecified analysis models). There were 864,000 cases in this model. Figure 4.5 provides a visualization for these results. In the figure, the columns represent the data analysis model. Sample size is displayed on the horizontal axis, and type I error rate is on the vertical axis. Results are split up by the size of the interaction effect (rows) and analysis model (color).

Parameter bias is still a problem in these models because in all cases, the parameter of interest (index of moderated mediation) should be zero. Relative bias, therefore, was not calculable. The average parameters that should be zero instead ranged from less than 0.0001 to 0.0114 when the data analysis was done using Model 14 when Model 8 was the DGP. This was higher at higher effect sizes (ranging from 0.007-0.015 when separated by effect size), but was less variable by sample size (ranging from 0.011-0.013, with higher bias at lower sample sizes).

H3a: Type I Error Rates Hypothesis 3a was partially supported. Results showed very few conditions to have type I error rates that were too conservative, with only 1.27% of the 864 conditions having type I error rates below .025 (Bradley criteria) and 6.48% of conditions having type I error rates below .035 (Serlin criteria). There was no noticeable factor common across these conservative conditions. However, more conditions had type I error rates that were too liberal. Around 10.19% were too liberal (Bradley criteria), and among these, all but two are conditions where the DGP is Model 8 and the data analysis model is Model 14. As shown in Figure 4.5, the other two cases with type I error rates that exceed Bradley's criteria have Model 7 as the DGP and either Model 14 or Model 15 as the data analysis model. Both happen when the effect size is large. Using Serlin's criteria, this proportion of overly liberal type I error rates increases to 13.31%, and the only noticeable pattern is that these rates tend to occur when the sample size is higher.

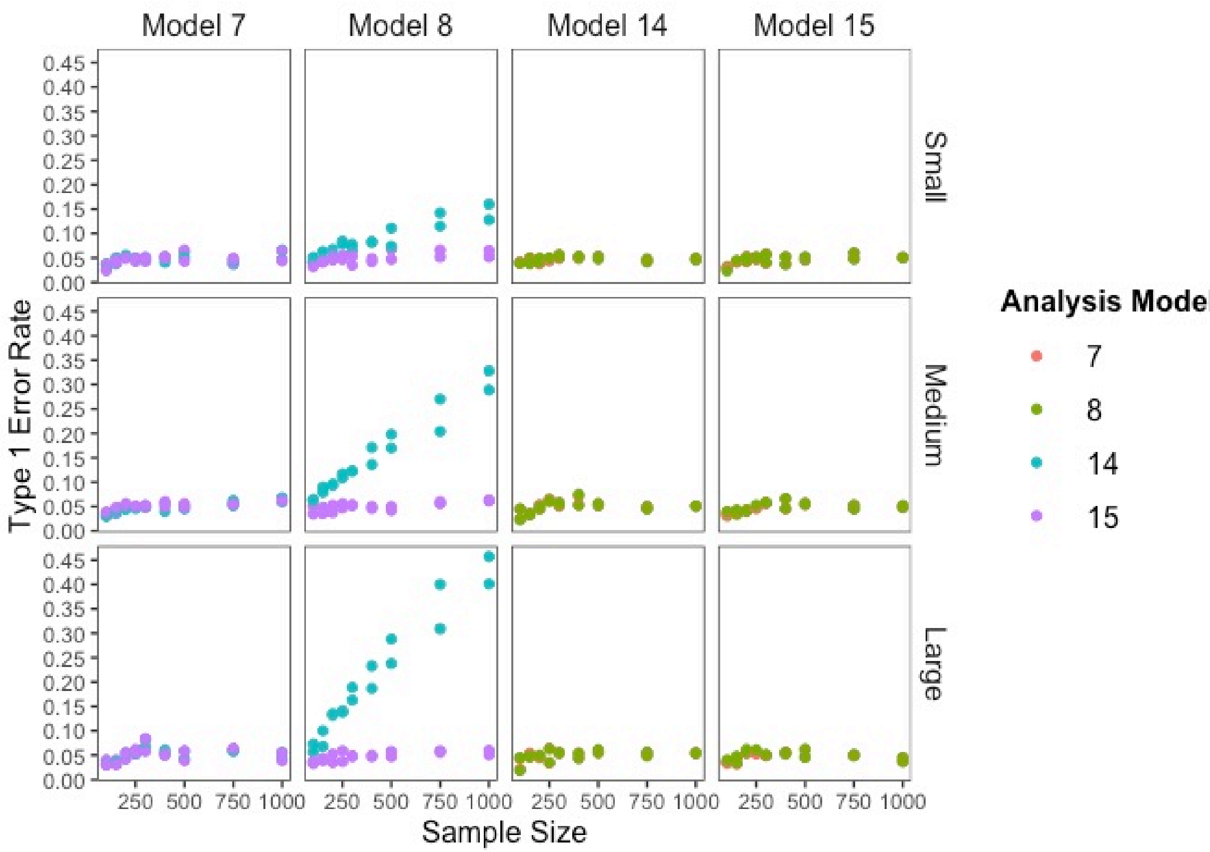

Figure 4.5: Figure displaying type I error rate for moderated mediation models. The columns represent the data analysis model. Sample size is displayed on the horizontal axis, and type I error rate on the vertical axis. Results are split up by size of the interaction effect (rows) and analysis model (color). For clarity, the figure is shown for continuous X only. Additional figures are provided in the Appendix with dichotomous X, and separated by continuous and dichotomous W.

Table 4.4 shows type I error rates by sample size. The columns are separated by the analysis model, and the rows are the DGP. Results are separated by effect size, with Table 4.4 showing results when the coefficient on the interaction term was .17 (medium). In Table 4.4, all of the type I error rates that are too liberal according to Bradley's criteria are found with Model 14 as the analysis model and Model 8 as the DGP. A few type I error rates that are either too liberal or too conservative using Serlin's criteria are also shown. Additional tables by effect size are available in the Appendix.

		Analysis Model			*Analysis Model*	
Sample Size	*DGP*	Model 7	Model 8	*DGP*	Model 14	Model 15
100	14	.045	.045	7	.030	.039
	15	*.033*	.040	8	.063	.035
150	14	*.034*	*.033*	7	.043	.048
	15	.042	.043	8	***.079***	.036
200	14	.054	.050	7	.055	.056
	15	.040	.041	8	***.093***	.053
250	14	.063	.061	7	.046	.051
	15	.049	.053	8	***.109***	.056
300	14	.058	.058	7	.052	.053
	15	.056	.058	8	***.122***	.054
400	14	*.072*	*.075*	7	.040	.050
	15	*.066*	*.066*	8	***.136***	.050
500	14	.056	.056	7	.045	.048
	15	.057	.057	8	***.170***	.042
750	14	.045	.045	7	.062	.055
	15	.045	.045	8	***.204***	.056
1,000	14	.051	.051	7	.060	.062
	15	.052	.052	8	***.289***	.063

Table 4.4: Type I error rate by sample size. The columns represent the data analysis model, and the DGP is listed in the row. Type I error rates in the table are shown only for **medium** effects with continuous X and continuous W. Type I error rates exceeding criteria set by Serlin (2000) are in *italics* and rates exceeding criteria set by Bradley (1978) are in **bold**.

H3b: Number of Moderated Paths Hypothesis 3b was supported, but in the opposite direction as hypothesized. The number of moderated paths in the analysis model was a significant predictor of statistical type I error rate, controlling for sample size, DGP, magnitude of the effect size, and type of X variable. The test for this was the comparison between two moderated paths and one moderated path in the analysis model, $b = -0.45, p < .001$, $CI_{95\%} = [-0.47, -0.43]$, $OR = 0.64$. This suggests that the odds of making a type I error are

0.64 times lower for analysis models with two moderated paths compared to one moderated path. Based on Figure 4.5, this pattern is likely attributed to the highly inflated type I error rate of using Model 14 as the analysis model when the DGP was Model 8 because Model 14 only has one moderated path.

Type of X and W There was no significant difference in type I error rate for continuous or dichotomous predictor X, $b = 0.005$, $p = .82$, $CI_{95\%} = [-0.04, 0.05]$, $OR = 1.00$. Additionally, there was no significant difference in type I error rate for continuous or dichotomous predictor W, $b = -0.02$, $p = .46$, $CI_{95\%} = [-0.06, 0.03]$, $OR = 0.98$. Plots showing these relationships for dichotomous x, and split up by type of W variable are provided in the Appendix.

Exploratory: Interactions Model A model containing all possible two-way through six-way interactions was fit, but none of the interactions were significant at the $\alpha = .001$ level.

4.7 Discussion

The simulation study in this dissertation aimed to investigate how over-specifying, under-specifying, and completely misspecifying moderated mediation models impacts statistical power and type I error rate. Using a Monte Carlo simulation with a factorial design, I systematically varied the generating models (DGPs), the magnitude of the moderated effect, sample size, type of X and moderator variable, and data analysis models to examine their effects on the outcomes. This section discusses the implications of the findings of the simulation, alignment with the hypotheses, and potential avenues for future research.

There were two outcomes of interest in this study: statistical power (for over-specified, under-specified, and correctly specified models) and type I error rate (for completely misspecified models) for the index of moderated mediation. Both were calculated as the proportion of the 1,000 generated samples that had a statistically significant indexe of moderated mediation in each condition, which was the rejection rate. The rejection rate was either statistical power or type I error rate, depending on the type of model specification. The goal was to understand which factors had the potential to increase or decrease statistical power or increase or decrease the type I error rate in cases of complete model misspecification.

Over-specified Models For Hypothesis 1, which focused on over-specified models compared to correctly specified models, the outcome was statistical power. The results showed that model specification was a significant predictor of statistical power when the moderator was continuous and when the moderator was dichotomous. Over-specified models had lower statistical power compared to correctly specified models, fitting with the literature that over-specifying regression models has the consequence of lower power to detect effects (Babyak, 2004). The number of moderated paths in the analysis model also mattered, with the power to detect a significant index of moderated mediation being lower with more moderated paths. Then, as a positive check and fitting with previous research such as Fritz and MacKinnon (2007) in the mediation literature, higher power was found in cases with higher sample sizes. As another positive check, the magnitude of the effect size was a significant predictor of statistical power, meaning that higher power can be expected for models with larger effect sizes. Finally, there was no significant difference in power for continuous or dichotomous predictor X for either the continuous or dichotomous moderator model.

Under-specified Models Model specification was again a significant predictor of statistical power, but whether under-specification had higher or lower power depended on the specific DGP and data analysis models. Under-specification predicted either higher (continuous moderator models) or lower (dichotomous moderator models) power for under-specified models compared to correctly specified models based on the multilevel logistic regression results. This complex finding was explored further through Table 4.3, Figure 4.3, and Figure 4.4. The mixed findings seem mostly due to the power of an under-specified analysis model depending on the DGP. Power was slightly higher when Model 7 was used as an under-specification of DGP Model 8, and power was even higher when Model 14 was used as an under-specification of DGP Model 15. This difference likely has to do with moderation of the direct effect, and Model 14 possibly capturing some of that effect on Y in the indirect effect. Model under-specification had lower power when Model 58 was the under-specification of DGP Model 8 or 15, similar to the lower power when Model 58 was the over-specification of DGP Model 7 or 14. In these cases, moderation of the direct effect seemed to matter less, and the moderation of both paths in the indirect effect led to decreased power. The mixed results for under-specified models could also be in part due to the large amount of parameter bias in many of these models. This parameter bias indicates that even in cases where it may seem advantageous in terms of power to under-specify a model, it will likely result in more biased estimates. Similar to previous results, number of moderated paths is a significant predictor of lower statistical power. Additionally, as positive checks, higher sample sizes and larger effect sizes were again significant predictors of higher statistical power. Finally, again, there was no significant difference in power for continuous or dichotomous predictor X for either the continuous or dichotomous moderator model.

Completely Misspecified Models Type I error rates were generally not overly conservative, though type I error rates were higher than typically seen in the mediation literature (Yzerbyt et al., 2018), and were more often overly liberal. These overly liberal conditions were mainly associated with specific combinations of the DGP and data analysis model, with the most problematic combination occurring where Model 8 was the DGP and Model 14 was the analysis model. Model 8 has moderation on both the X to M path and the M to Y path, so when these are constrained to zero in the data analysis model (Model 14; which only has moderation on the M to Y path), those effects could be showing up on the one moderated path, making the index of moderated mediation appear significant. This could help explain the severely inflated type I error rates for this combination of DGP and analysis model. One thing that makes this relationship unique is that it is the only case where there is moderation on a path predicting Y (specifically the direct effect) in the DGP, but the data analysis model has moderation on the other path predicting Y, which is part of the indirect effect. No other pairs of DGPs and analysis models have this distinction, so some of the variance in Y from the moderation could be falsely accounted for in the analysis model as part of the index of moderated mediation, pushing it to be significant. This pattern is not seen in the reverse when Model 14 is the DGP and Model 8 is the analysis model. However, this could still be explained by the moderated X to Y path taking on some of the effect from the moderated M to Y path from the DGP, but it would not matter for the index of moderated mediation so the type I error rate would not be inflated. These type I error rates between Model 8 as the DGP and Model 14 as the analysis model were far higher than other conditions, and this could be behind the relationships seen in the following analyses.

None of the other data analysis model and DGP combinations had such inflated type I error rates, even when the DGP had two moderated paths while the data analysis model only had one. This differs from the previous findings about statistical power, where the number of moderated paths did matter. For completely misspecified models, the particular paths that are moderated or omitted seemed more important than the number of paths. As positive checks, the highest type I error rates happened for larger effect sizes, and smaller sample sizes indicating that the effect size of the moderated path in the DGP and the sample size matter to an extent. Finally, there was again no significant difference in type I error rate for either continuous or dichotomous moderator W, or continuous or dichotomous predictor X.

Limitations and Future Directions

While this simulation study provides insights into the effects of model specification on statistical power and type I error rate in moderated mediation models, several limitations should be acknowledged. These limitations point to potential directions for future research to further enhance understanding of the factors affecting statistical power and type I error rates in moderated mediation models.

First, the simulation study focused on six specific moderated mediation models. In reality, there are far more moderated mediation models used in research. Based on results from the meta-analytic review, these results should cover around 85% of moderated mediation models currently being used, but that is not all models, and more complex models such as multi-level models were excluded. Future studies could explore a broader range of model specifications and DGPs, including models with two different moderators such as Model 21, to assess whether the results hold across different contexts and conditions. This could reveal additional edge cases that have drastically different results, such as the inflated type I error rate found in the simulation where the DGP was Model 8 then Model 14 was used for data analysis. These cases should be rare, but if more complex models are growing in popularity now that they are available in the most recent version of the commonly used tool PROCESS (Hayes, 2022), it is worth exploring implications for power and type I error rate for misspecifying these models.

Additionally, another limitation is the choice of equal allocation to each level when the predictor variables were dichotomous. As the difference in size between groups increases, power for detecting significant moderated effects has been shown to decrease (Aguinis & Stone-Romero, 1997). Future work could extend this to cases where allocation is unequal, which may be more representative of dichotomous variables that are not experimentally manipulated, where statistical power is expected to be lower (Aguinis & Stone-Romero, 1997). The predictor variables (X and W) were also always generated independently, so future research could examine the impacts of variable dependence.

In conclusion, while this simulation contributes insights into the effects of model specification on statistical power and type I error rate in moderated mediation models, future research can expand on these findings by including additional model specifications and exploring unequal allocation. Future work can also be done to create easy-to-use tools such as a Shiny app to simplify power analysis calculations for individual research projects. By addressing these limitations, future studies can further enhance our understanding of model specification in moderated mediation analysis and provide more comprehensive recommendations for researchers in various fields using even more complex moderated mediation models.

5 General Discussion

Overall, the aim of this dissertation was to understand how factors such as model specification affect statistical power and type I error rate in moderated mediation models. I began by introducing mediation, moderation, and moderated mediation. Next, I included a meta-analytic review of recent articles using moderated mediation analysis. This meta-analytic review had two main goals: to understand the contexts in which moderated mediation analyses are used and assist in the selection of relevant sample sizes and model specifications to explore further via a simulation study. Then, in an applied example using moral emotions data, I explored a real-world example of how power can differ based on model specification. In this example, a significant index of moderated mediation that was detectable in a model with fewer paths moderated was no longer significant when the model was over-specified with additional paths moderated. This implied a decrease in statistical power for the over-specified model. Finally, the simulation study further explored the impacts of model specification on power and type I error rate.

In the simulation study, the parameters used were chosen directly from the meta-analytic review to best represent current practices. Specifically, the nine sample sizes represented the inner deciles from the range of sample sizes found in the review, and the six moderated mediation models collectively represented over 85% of models used in the literature based on the meta-analytic review. Using these parameters, I found that over-specified models had lower power than correctly specified models, and under-specified models had mixed results for power and greater parameter bias than correctly specified models. Additionally, while the type I error rate was for the most part not overly liberal or conservative for completely misspecified models, a certain case of model misspecification where the data analysis model missed moderation on the direct effect had a highly inflated rate. In this chapter, I will discuss the results from the simulation study, informed by the meta-analytic review, and connect it to the applied example. I will also discuss some limitations and future directions for this dissertation work.

Previous research has shown several factors that affect statistical power and type I error rate in mediation and moderation analyses (e.g., Fritz & MacKinnon, 2007; McClelland & Judd, 1993; Yzerbyt et al., 2018), but research for moderated mediation analyses is limited. Compounding this problem, few tools exist to aid researchers in sample size planning for these complex designs, and the two tools that do exist are limited to providing power as the output instead of sample size (Aberson, 2019a; Zhang & Yuan, 2018). Before sample size planning through a power analysis can be done, researchers must decide on what role each variable will play in the model, specifying the moderated mediation model. Moderated mediation analysis requires careful justification for where the moderation occurs in the model as part of model specification. Model specification was hypothesized to affect statistical power and type I error rate, and this dissertation began the work of exploring how over-specifying, under-specifying, and completely misspecifying models can impact statistical power and type I error rate.

5.1 Discussion of Simulation Results

This simulation aimed to investigate the effects of model specification on statistical power and type I error rate in moderated mediation models. The results generalize to six commonly used models, accounting for 85% of published moderated mediation analyses from the meta-analytic review. Results pertaining to all three hypotheses are integrated in this section.

The results suggested that model specification matters for statistical power and type I error rate, with both continuous and dichotomous moderators. Over-specifying a model lead to decreased statistical power, indicating substantially lower power for over-specification compared to correct model specification. This is similar to findings from Babyak (2004) about over-specified regression models having decreased power to detect true effects that also led to problems with replication. Under-specifying models had mixed results, with under-specified models having higher power than correctly specified models with continuous moderators and lower power with dichotomous moderators. Exploring these results further, the higher power for under-specified models was only seen when the under-specified model omitted moderation on the direct effect, and it was only a modest difference in power. In the rest of the cases, under-specified models had lower power, with the largest differences in power seen when Model 58 was used for data analysis, moderating both paths in the indirect effect instead of just one from the DGP. Under-specification is also a challenge seen in the machine learning literature, where under-specified training models have been criticized for their instability in real-world situations when these models are deployed (D'Amour et al., 2022).

Relative bias was also high in under-specified models, though not an issue with over-specified models. For under-specified models, relative bias was highest when both paths of the indirect effect were moderated in the DGP but only Model 7 is moderated for data analysis. Relative bias was also unacceptably high where moderation was omitted on the direct effect in the data analysis model. This fits with the structural equation modeling literature, where model misspecification, in particular under-specification, was found to demonstrate greater bias in a simulation study (Yang & Green, 2010). Additionally, parameter bias varied depending on the size of the effect on the interaction term and sample size, with larger effect sizes and smaller sample sizes being associated with greater parameter bias. This is consistent with previous findings in the mediation literature, that parameter bias decreases as sample size increases (Mackinnon, Warsi, & Dwyer, 1995), and this study extends the finding to a wider range of sample sizes. Relative bias was was only at an acceptable level when moderation of the X to M path was omitted, but these cases of under-specification resulted in lower power, so there is no advantageous way to under-specify a model without it resulting in either biased estimates, lower power, or in most cases, both. Additionally, complete misspecification mattered for the type I error rate. When the direct effect was moderated in the DGP (such as Model 8), then the data analysis model only allows for moderation on the M to Y path (Model 14), type I error rate was highly inflated.

Notably, having a continuous or dichotomous X variable did not affect statistical power in any model, holding effect size (as measured by explained variance) and sample size constant. This was evidenced by the non-significant results with very small effect sizes. This was inconsistent with findings from a highly cited simulation study that suggested power was higher for interactions involving a dichotomous variable (McClelland & Judd, 1993). However, the finding from my dissertation study is consistent with Coutts (2023), where the power for moderated mediation models was not significantly affected by continuous or dichotomous predictors. This discrepancy can be attributed to the variances of the variables generated for the simulation. Dichotomous variables were coded as -1, 1, which has a variance of 1 with equal allocation. However, the variance for the continuous variable in McClelland and Judd (1993) was set to 0.5, whereas I set the variance to 1 in this dissertation study. When the variances of both variables are the same, there is no difference in power for dichotomous or continuous predictors. While there is no power benefit to using dichotomous predictors, there are still benefits to experimental manipulation which often results in a dichotomous predictor variable, such as the elimination of measurement error through the manipulation and establishing temporal precedence (Judd & Kenny, 2010). There is also still good reason to never dichotomize continuous variables (Royston, Altman, & Sauerbrei, 2006), including the reduced power to detect interactions involving a dichotomized variable (Durand, 2013).

As a positive check for the simulation results, and consistent with previous research, power was higher at higher sample sizes. This was found across all analyses of over-specified models and under-specified models. Based on how statistical power is defined, as sample size increases, statistical power should increase as well (J. Cohen, 1969). However, specific differences between individual sequentially coded sample sizes were not always significant. The power curves demonstrated diminishing returns, with the rate of increase in power becoming less pronounced at larger sample sizes, fitting with how power curves generally function with diminishing returns at higher sample sizes, to the point where an ethics paper was written arguing that often participant burden too costly to go from 80% power to 90% power (Bacchetti, Wolf, Segal, & McCulloch, 2005). Type I error rate was also higher overall for higher sample sizes from the Wald test, fitting with findings from Ma and Zeng (2014). Ma and Zeng (2014) explored power and type I error rates in multiple mediator models, and found higher type I error rates with sample sizes of 500 compared to 200, in addition to 200 compared to 100. In this dissertation simulation, at sample sizes higher than 200, the type I error rate levels out around .05. However, the result of the lowest type I error rate in the smallest sample size condition is consistent.

As another positive check, and again consistent with previous research, effect size was a consistent predictor of both power and type I error rate. Effect size goes into the calculation of statistical power (J. Cohen, 1988). From this simulation, the effect size of the interaction term was a significant predictor of statistical power, with larger effect sizes associated with higher power. Aguinis et al. (2005) found in a review of effect sizes over 30 years that effect sizes reported in the published literature with categorical moderators increased over time, which means studies should on average be trending towards higher power. The finding about type I error rate being higher at higher effect sizes is also consistent with reported results in mediation analysis, where type I error was higher when one of the paths in the indirect effect was higher and the other zero, for it to be a type I error (Yzerbyt et al., 2018). In this simulation, for example, the type I error rate was .083 for a large interaction effect and .051 for a small interaction effect. This result was specific to the DGP being Model 7 and Model 14 being used as the data analysis model, with all continuous variables, at a sample size of 300, but is an example of the general pattern of type I error rate being higher at higher amounts of variance explained by the interaction. These results as a whole confirm that researchers should still consider both sample size and effect size when designing studies, including when using moderated mediation models.

5.1.1 Support for Simulation Hypotheses

In summary, most hypotheses were supported. Hypothesis 1a was supported, with over-specified models having lower power than correctly specified models. This was seen most drastically when Models 58 and 59 were used as over-specified data analysis models. Hypothesis 1b was also supported, with additional moderated paths predicting lower power as well. Hypothesis 2a had mixed support. Model specification for under-specified models had either higher or lower power than correctly specified models, depending on the combination of DGP and which model was used for data analysis. Hypothesis 2b was supported, with parameter bias indeed being higher for under-specified models. For Hypothesis 3a, up to 13.31% of tests had inflated type I error rate (Serlin, 2000), and one case of complete misspecification reached type I error rates of nearly .50, offering partial support for this hypothesis that type I error rate is increased for completely misspecified models. Hypothesis 3b was also supported, with the number of moderated paths in the analysis model significantly predicting a higher type I error rate. Both Hypothesis 3a and 3b are largely due to misspecification of a model by including moderation on the M to Y path but omitting moderation of the direct effect found in the DGP.

5.2 Connecting the Applied Example

The goal of the simulation study was to investigate the effects of over-specifying, under-specifying, or completely misspecifying a moderated mediation model. One of the main findings was that over-specifying a moderated mediation model can lead to decreased power to detect a significant index of moderated mediation. The applied example used real data to illustrate this, showing what could potentially happen by over-specifying a moderated mediation model using moral emotions data. This was hypothesized to potentially obscure important findings in over-specified analysis models based on results from the simulation study showing that over-specified models have lower statistical power compared to correctly specified models.

The findings from the "correctly" specified model (Model 7) revealed a significant index of moderated mediation, theoretically indicating that the indirect effect of age on disclosure through repair was moderated by empathy. The results showed significant effects of age on repair, empathy on repair, and an interaction effect between age and empathy on repair. Additionally, there were significant effects of repair on disclosure and age on disclosure. The conditional indirect effect of age on disclosure through repair was significant when empathy was absent but not significant when empathy was present in the narrative. In contrast, the over-specified model (Model 59), where empathy moderated all paths, did not yield a significant index of moderated mediation. While the effects of age, repair, and the interaction between age and empathy on repair were consistent with Model 7, there was no significant effect of empathy on disclosure or significant interactions between age and empathy or repair and empathy on disclosure. Consequently, there was no evidence of moderated mediation in this over-specified model.

While this applied example highlights the importance of correctly specifying a moderated mediation model in an anecdotal form, the results from the simulation study justify that over-specification may be a major issue as hypothesized in terms of statistical power. As seen in the applied example, over-specifying the model by including unnecessary moderator variables can lead to non-significant or spurious findings, potentially obscuring the true relationships among variables. In the applied example, the over-specified model failed to capture the significant moderated mediation effect observed in the "correctly" specified model where one path was moderated. For this reason, and to a greater degree for the reason of simply having a correctly specified model for the sake of reporting correct results, researchers should use caution when deciding which paths to moderate and provide strong justifications for each moderation effect included in the model.

These findings, when held together, underscore the need for researchers to consider the theoretical rationale behind their model specifications and not over-specify a model without good reason. The applied example focused on demonstrating a possible consequence of over-specification, highlighting the need to establish clear justifications for each path in a moderated mediation model. The inclusion of moderation effects should be guided by theory and empirical evidence rather than a desire to avoid missing potentially relevant effects in any case.

In conclusion, this applied example demonstrated the potential consequences from the simulation study of over-specifying a moderated mediation model using moral emotions data. The findings emphasized the importance of correctly specifying the model and providing justifications for moderation effects, because there are risks to power and type I error rates associated with over-specifying, under-specifying, and in certain cases, completely misspecifying models. Researchers should exercise caution when deciding which paths to moderate and avoid including unnecessary moderation effects that lack strong theoretical grounding or other valid reasoning. Model specification matters for the validity and interpretability of research findings in moderated mediation analyses.

5.3 Summary

This dissertation aimed to provide insights into the effects of model specification on statistical power and type I error rate in moderated mediation models. Over-specified models had lower power compared to correctly specified models, as demonstrated in the applied example. Under-specified models had mixed results depending on the DGP, in addition to higher parameter bias than correctly specified models. The cost of over-specifying a model is the loss of power, which must be weighed relative to the cost of under-specifying a model, which is riskier in terms of getting an accurate parameter estimate and could also lead to lower power. While one possible solution is to over-specify models and recruit enough participants to still have high power, there are ethical concerns in terms of participant burden (Bacchetti et al., 2005), and there can be other resource limitations preventing larger sample sizes (Lakens, 2022). Completely misspecified models tended to have reasonable type I error rates, except where the direct effect was moderated in the DGP but not in the data analysis model, which moderated the path between the mediator and the outcome instead, possibly because the moderation of the direct effect was misattributed to the indirect effect. The riskiest form of complete misspecification for type I error rate happened when moderation of the direct effect was omitted in the data analysis, and instead moderation was incorrectly specified on the path between the mediator and the outcome. Based on the results from complete model misspecification, and since over-specified models tend to have acceptable levels of parameter bias, choosing to potentially over-specify a model to include moderation on the direct effect might be worth the risk of lower power to avoid a highly inflated risk of making a type I error.

The findings from this dissertation highlight the importance of model specification for moderated mediation analyses. The risk of over-specification is decreased power, as including unnecessary moderated paths can decrease statistical power and obscure meaningful effects. The risk of under-specification is unacceptable levels of parameter bias for most under-specifications, and often also lower power. Taken together, these findings indicate that there is no possible way increase statistical power without an unacceptable level of bias by over- or under-specifying a model. The risk of complete misspecification, aside from having no chance of finding a true effect, is a possible inflated type I error rate depending on the analysis model in comparison to the DGP. In other words, when models are completely misspecified, correct conclusions cannot be drawn, though these complete misspecifications generally do not lead to inflated type I error rates except in one case detected in the simulation study. That case, where the DGP moderated the direct effect and the analysis only moderated the M to Y path, is suggestive of issues of confounding between moderation of the direct path compared to the M to Y path as part of the indirect effect. The dissertation emphasized the role of properly specifying a model in sample size planning for adequate power in moderated mediation models.

While this dissertation contributes novel findings, some limitations should be acknowledged. The analysis focused on a specific set of model specifications, limiting how these findings can generalize to other types of moderated mediation models, and in particular more complex models including structural equation models. Future research could explore a broader range of model specifications to assess the robustness of the findings. This study also focused on observed variable systems only. Many psychological variables of interest are more complex, and cannot easily be modeled using a single score for each participant on each variable. However, while acknowledging that this is not an exact measurement of the difference, over 85% (343) of the analyses from the meta-analytic review used regression, meaning these results should apply to the majority of moderated mediation analyses. Future research could explore differences in power and type I error rate in structural equation modeling contexts, especially given the increasing popularity of SEM in positive psychological research (van Zyl & ten Klooster, 2022), and the field of psychological science as a whole.

Overall, recently published moderated mediation analyses from the literature review tended to have sufficient sample sizes to detect large moderated effects, but larger sample sizes are generally needed to detect the small (and even medium) effects typically seen with interactions (Aguinis et al., 2005). Table 5.1 gives the required sample size for 80% power (within 0.1%) for each model for the effect sizes of the interaction used in this simulation study: small (.10), medium (.17), and large (.22). All paths in the mediation model set to .26, which an effect between small and medium (Fritz & MacKinnon, 2007) according to Cohen's guidelines (J. Cohen, 1988). The meta-analytic literature review did not collect any information about effect size, but the proportion of analyses using each analysis model that would have 80% power in each case is also shown in Table 5.1.

	Model 7	Model 8	Model 14	Model 15	Model 58	Model 59
Effect Size						
Small	549	549	532	553	775	778
	.22	.25	.19	.08	.23	.25
Medium	207	206	194	190	265	268
	.64	.69	.72	.67	.54	.58
Large	152	142	138	131	164	165
	.84	.79	.84	.92	.85	.77

Table 5.1: Sample sizes required for 80% power by moderated mediation model and effect size (top number in each cell), along with proportion of analyses with that sample size or greater from the meta-analytic review by model (bottom number in each cell).

To make these results more applicable, tools should continue to be developed for power analysis for moderated mediation models. Tools are helpful for sample size planning, but the current tools for power analysis for moderated mediation are limited to Models 7, 8, 14, and 15 in the R package ‘pwr2ppl’, plus Model 58 using WebPower. While these tools cover most of the commonly used moderated mediation models found in the meta-analytic review, both tools are limited to output power given sample size, when practically it would be more useful to have a sample size as the output for a desired level of power. Future tools can be developed or enhanced to facilitate sample size planning for individual research contexts using moderated mediation models, with the added emphasis from this dissertation that assuming correct model specification is very important.

In conclusion, while this dissertation study highlights the importance of model specification in moderated mediation analysis, future research should continue to focus on providing comprehensive guidelines and tools for researchers. I hope that this dissertation formed a start to this work and that it is applicable for researchers, especially because the parameters included in this study were informed by current practices in the field. By focusing on properly specifying moderated mediation models, researchers can better navigate the complexities of these models and improve their chances of finding the effects that truly exist.

6 Appendix

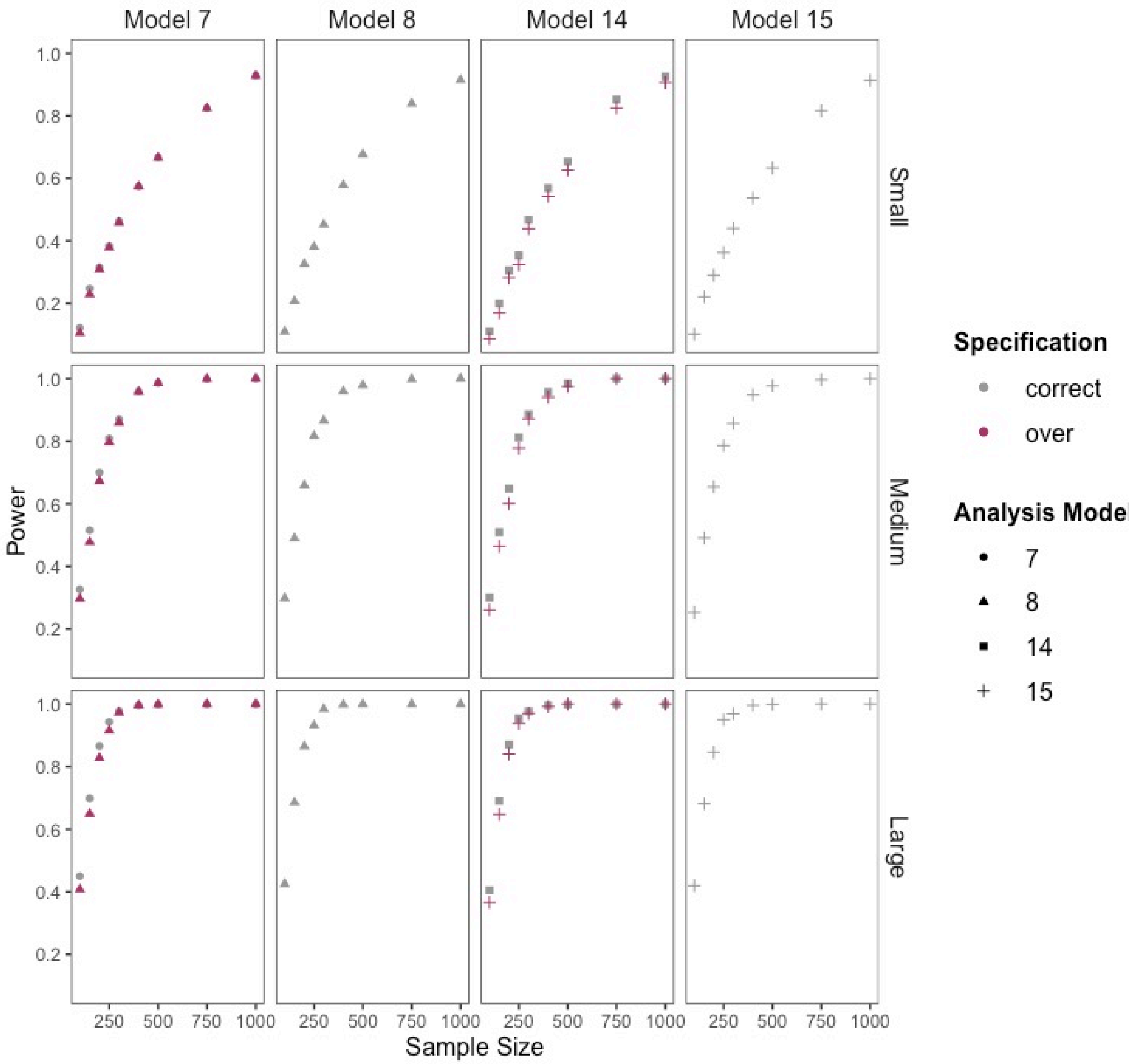

Figure 6.1: Figure displaying statistical power for moderated mediation models with a **continuous moderator**. The columns represent the DGP, and the rows represent effect size of the interaction(s). Sample size is displayed on the horizontal axis, and the outcome of power is displayed on the vertical axis. The results are split up by size of the model specification by color (grey for correct specification; pink for over-specification) and analysis model (shape). For clarity, figure is shown for dichotomous X only.

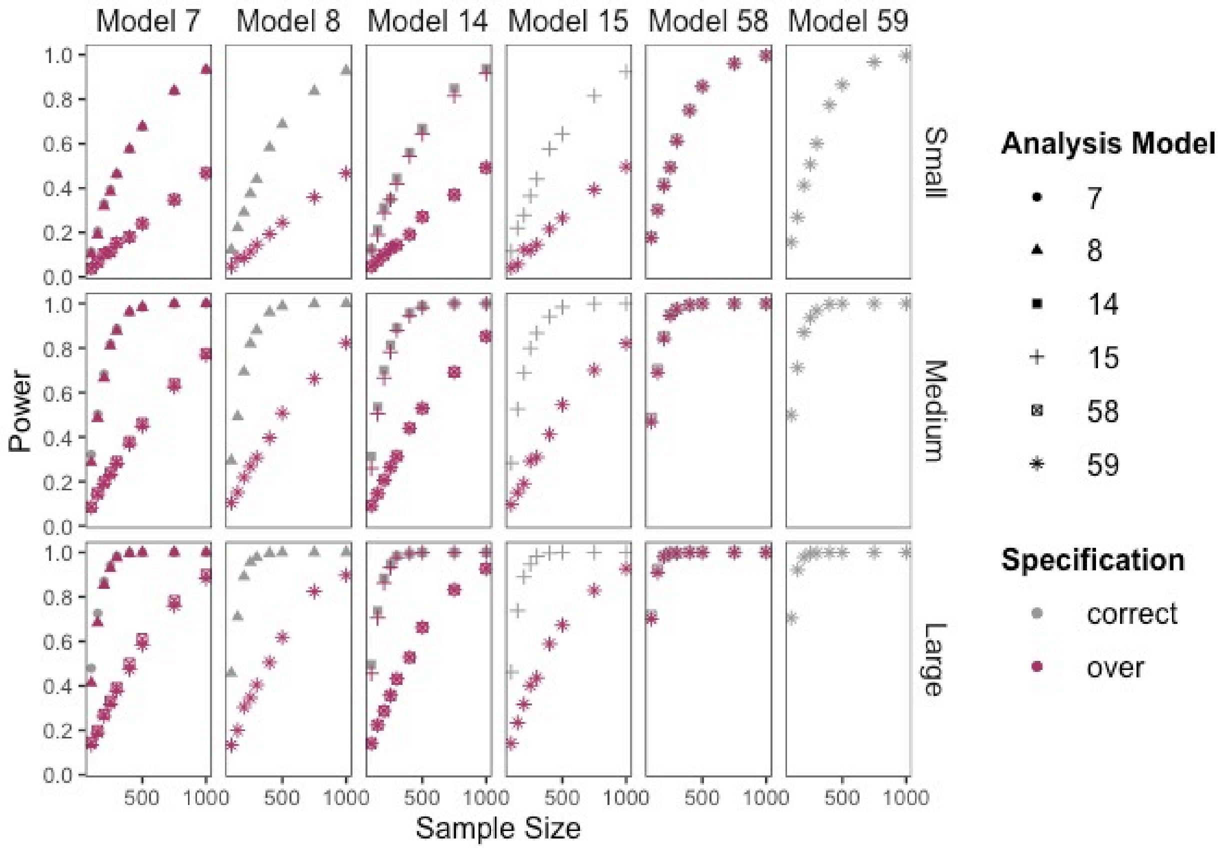

Figure 6.2: Figure displaying statistical power for moderated mediation models with a **dichotomous moderator**. The columns represent the DGP, and the rows represent effect size of the interaction(s). Sample size is displayed on the horizontal axis, and the outcome of power is displayed on the vertical axis. The results are split up by size of the model specification by color (grey for correct specification; pink for over-specification) and analysis model (shape). For clarity, figure is shown for dichotomous X only.

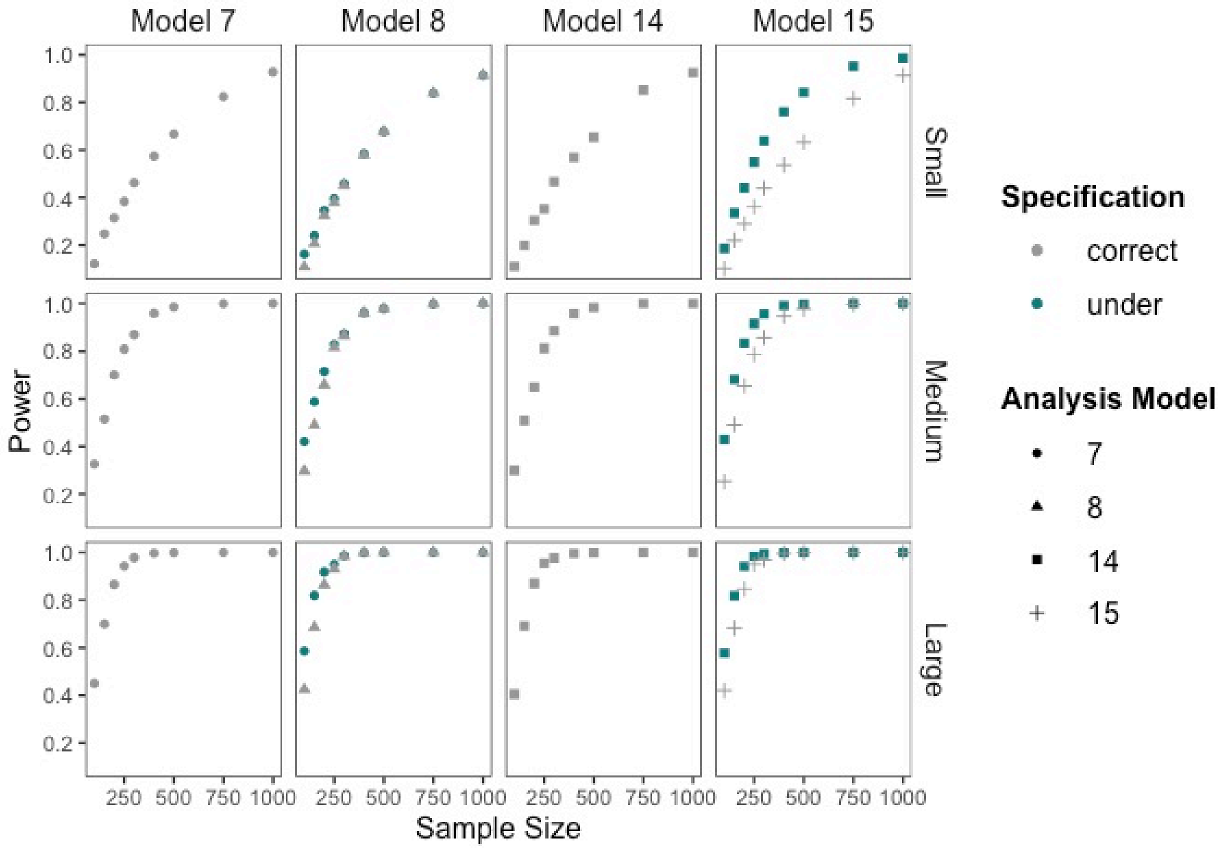

Figure 6.3: Figure displaying statistical power for moderated mediation models with a **continuous moderator**. The columns represent the DGP, and the rows represent effect size of the interaction(s). Sample size is displayed on the horizontal axis, and the outcome of power is displayed on the vertical axis. The results are split up by size of the model specification by color (grey for correct specification; teal for under-specification) and analysis model (shape). For clarity, figure is shown for dichotomous X only.

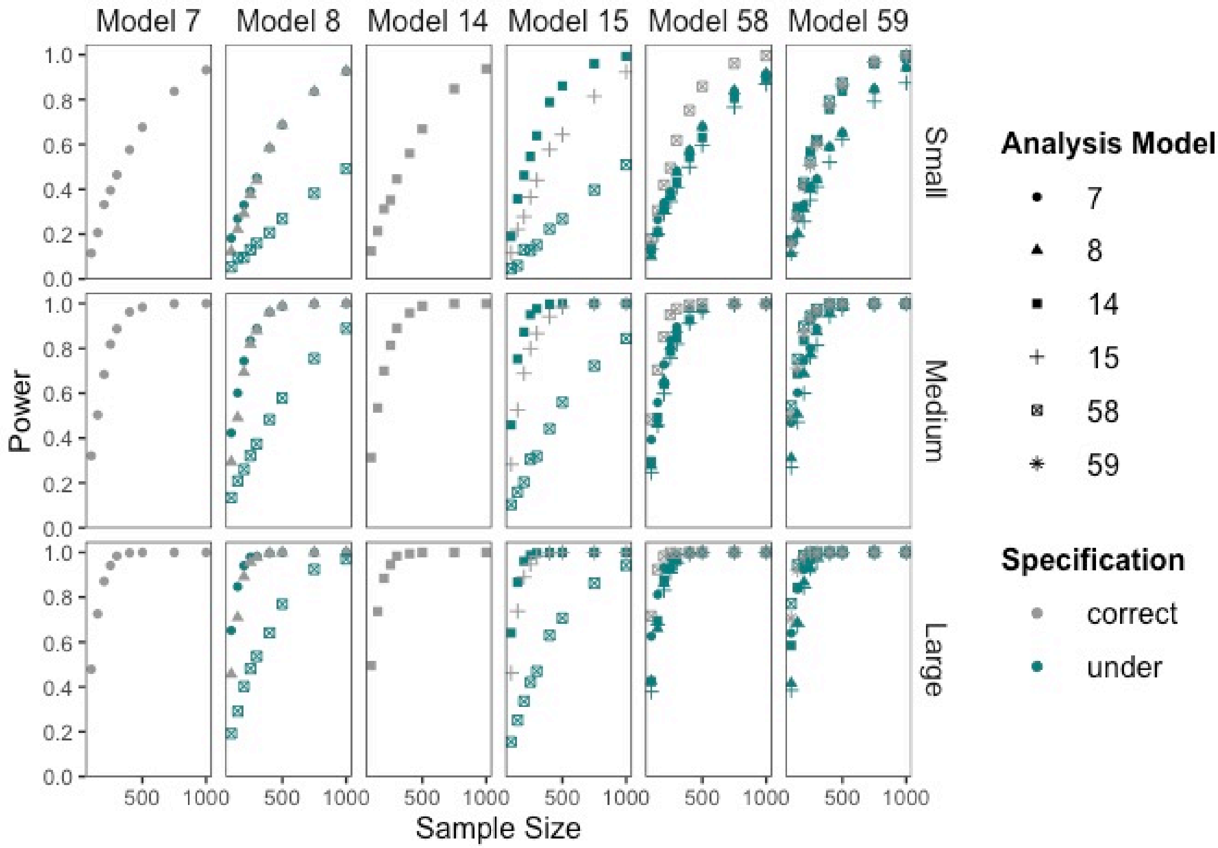

Figure 6.4: Figure displaying statistical power for moderated mediation models with a **dichotomous moderator**. The columns represent the DGP, and the rows represent effect size of the interaction(s). Sample size is displayed on the horizontal axis, and the outcome of power is displayed on the vertical axis. The results are split up by size of the model specification by color (grey for correct specification; teal for under-specification) and analysis model (shape). For clarity, figure is shown for dichotomous X only.

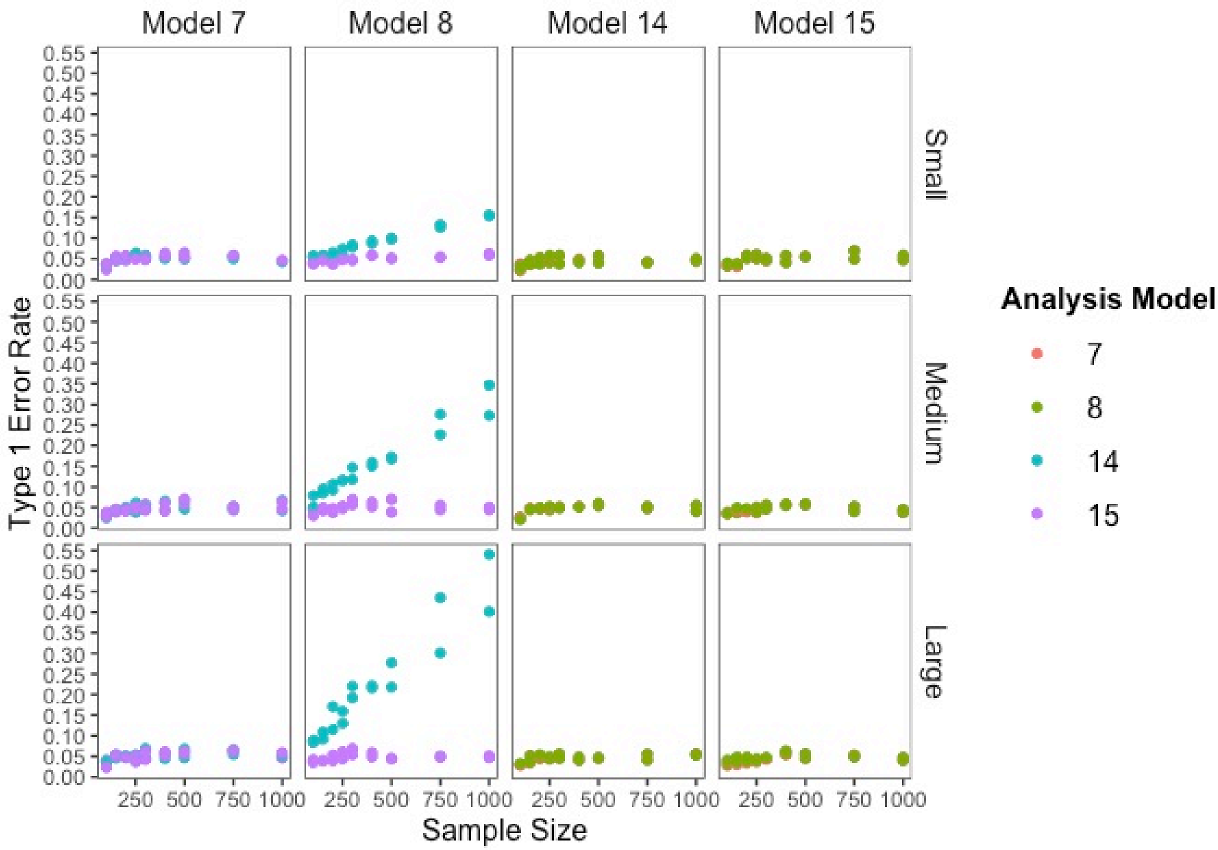

Figure 6.5: Figure displaying type I error rate for moderated mediation models. The columns represent the data analysis model. Sample size is displayed on the horizontal axis, and type I error rate on the vertical axis. Results are split up by size of the interaction effect (rows) and analysis model (color). For clarity, figure is shown for dichotomous X only.

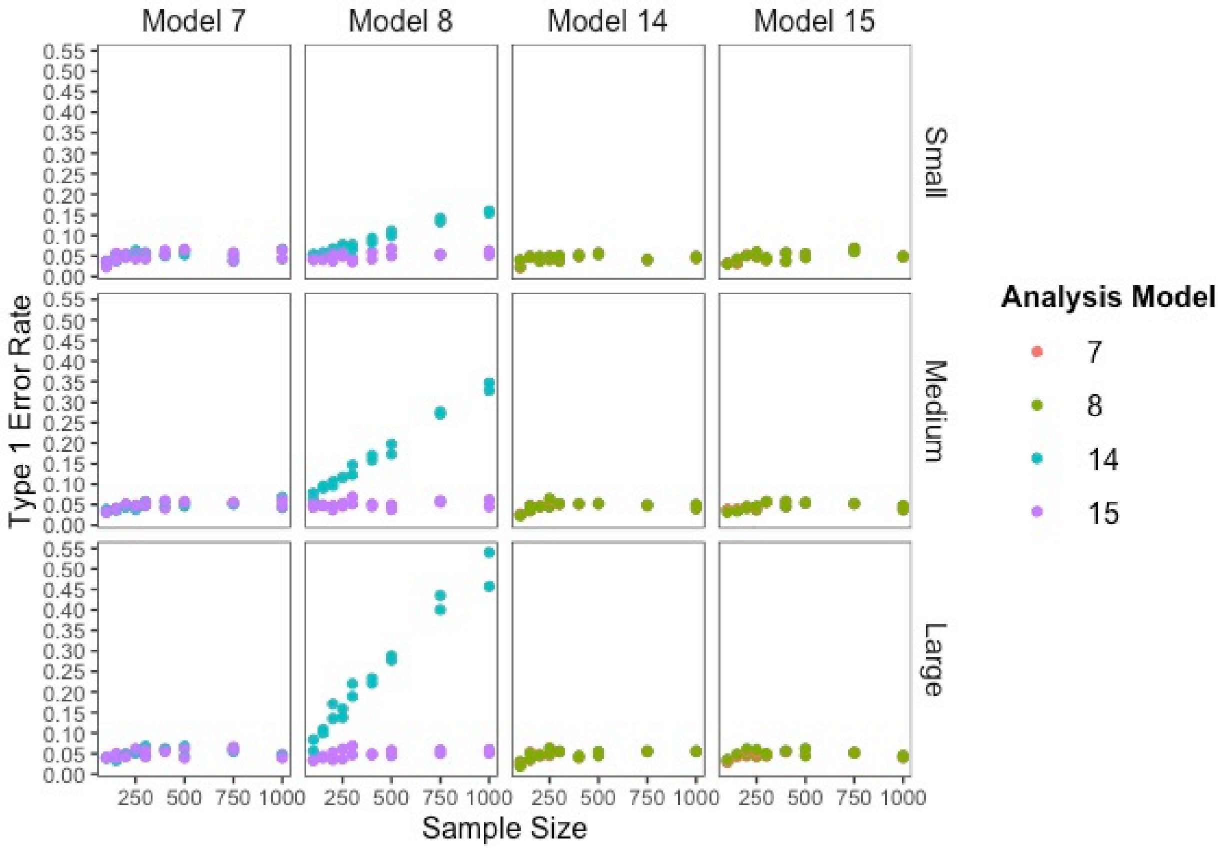

Figure 6.6: Figure displaying type I error rate for moderated mediation models. The columns represent the data analysis model. Sample size is displayed on the horizontal axis, and type I error rate on the vertical axis. Results are split up by size of the interaction effect (rows) and analysis model (color). For clarity, figure is shown for continuous W only.

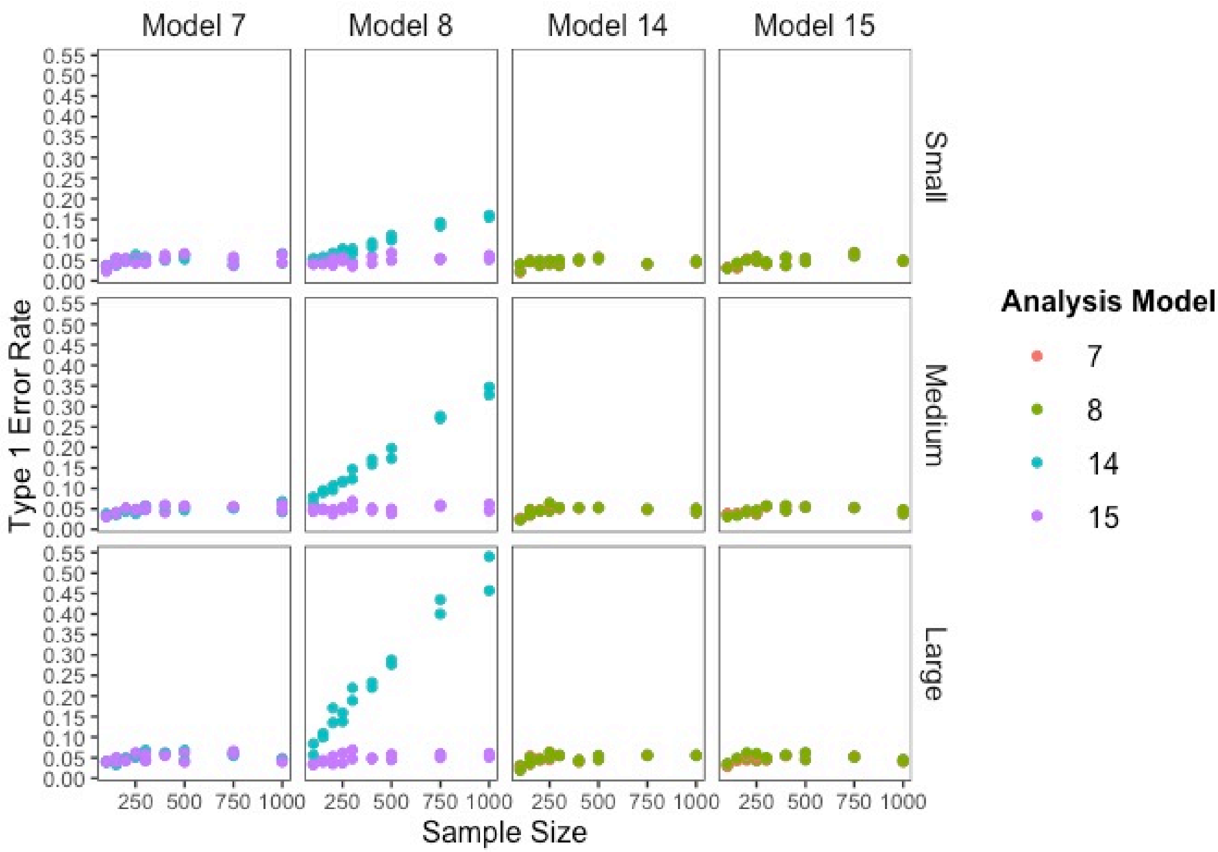

Figure 6.7: Figure displaying type I error rate for moderated mediation models. The columns represent the data analysis model. Sample size is displayed on the horizontal axis, and type I error rate on the vertical axis. Results are split up by size of the interaction effect (rows) and analysis model (color). For clarity, figure is shown for dichotomous W only.

		Analysis Model			*Analysis Model*	
Sample Size	*DGP*	Model 7	Model 8	*DGP*	Model 14	Model 15
100	14	.043	.045	7	*.031*	*.030*
	15	*.034*	.040	8	*.073*	.037
150	14	.047	.046	7	.036	*.031*
	15	*.032*	*.034*	8	*.068*	.039
200	14	.047	.050	7	.055	.055
	15	.054	.056	8	***.123***	.052
250	14	*.034*	*.035*	7	.054	.055
	15	.054	.061	8	***.141***	.059
300	14	.055	.055	7	***.083***	***.084***
	15	.050	.051	8	***.163***	.049
400	14	.054	.054	7	.052	.050
	15	.053	.053	8	***.187***	.050
500	14	.061	.061	7	.058	.059
	15	.046	.046	8	***.238***	.047
750	14	.049	.049	7	.058	.064
	15	.049	.049	8	***.309***	.056
1,000	14	.054	.054	7	.056	.055
	15	.037	.037	8	***.410***	.051

Table 6.1: Type I error rate by sample size. The columns represent the data analysis model, and the DGP is listed in the row. Only Models 14 and 15 were complete specifications for both Models 7 and 8, and only Models 7 and 8 were complete specifications for both Models 14 and 15. Type I error rates in the table are shown only for **small** effects with continuous X and continuous W. Type I error rates exceeding criteria set by Serlin (2000) are in *italics* and rates exceeding criteria set by Bradley (1978) are in **bold**.

Sample Size	*DGP*	*Analysis Model* Model 7	Model 8	*DGP*	*Analysis Model* Model 14	Model 15
100	14	.045	.045	7	.030	.039
	15	*.033*	.040	8	.063	.035
150	14	*.034*	*.033*	7	.043	.048
	15	.042	.043	8	***.079***	.036
200	14	.054	.050	7	.055	.056
	15	.040	.041	8	***.093***	.053
250	14	.063	.061	7	.046	.051
	15	.049	.053	8	***.109***	.056
300	14	.058	.058	7	.052	.053
	15	.056	.058	8	***.122***	.054
400	14	*.072*	*.075*	7	.040	.050
	15	*.066*	*.066*	8	***.136***	.050
500	14	.056	.056	7	.045	.048
	15	.057	.057	8	***.170***	.042
750	14	.045	.045	7	.062	.055
	15	.045	.045	8	***.204***	.056
1,000	14	.051	.051	7	.060	.062
	15	.052	.052	8	***.289***	.063

Table 6.2: Type I error rate by sample size. The columns represent the data analysis model, and the DGP is listed in the row. Only Models 14 and 15 were complete specifications for both Models 7 and 8, and only Models 7 and 8 were complete specifications for both Models 14 and 15. Type I error rates in the table are shown only for **large** effects with continuous X and continuous W. Type I error rates exceeding criteria set by Serlin (2000) are in *italics* and rates exceeding criteria set by Bradley (1978) are in **bold**.

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.*

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.

Aguinis, H., & Stone-Romero, E. F. (1997). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, *81*(1), 192–206.

Anand, S., Hu, J., Vidyarthi, P., & Liden, R. C. (2018). Leader-member exchange as a linking pin in the idiosyncratic deals - performance relationship in workgroups. *Leadership Quarterly*, *29*(6), 698–708. doi: 10.1016/j.leaqua.2018.07.005

Babyak, M. A. (2004, May). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*(3), 411–421. doi: 10.1097/01.psy.0000127692.23278.a9

Bacchetti, P., Wolf, L. E., Segal, M. R., & McCulloch, C. E. (2005). Ethics and sample size. *American Journal of Epidemiology*, *161*(2), 105–110. doi: 10.1093/aje/kwi014

Baranger, D. A. A. (2022). Tutorial: Power analyses for interaction effects in cross-sectional regressions. *PsyArXiv*. doi: 10.31234/osf.io/5ptd7

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

Biesanz, J. C., Falk, C. F., & Savalei, V. (2010). Assessing mediational models: Testing and interval estimation for indirect effects. *Multivariate Behavioral Research*, *45*, 661–701. doi: 10.1080/00273171.2010.498292

Bohrnstedt, G. W., & Goldberger, A. S. (1969). On the exact covariance of products of random variables. *Journal of the American Statistical Association*, *64*(328), 1439–1442.

Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, *20*, 115–140. doi: 10.2307/271084

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

Carpenter, T. P., Isenberg, N., & McDonald, J. (2019). The mediating roles of guilt- and shame-proneness in predicting self-forgiveness. *Personality and Individual Differences*, *145*, 26–31. doi: 10.1016/j.paid.2019.03.013

Chen, D., & Fritz, M. S. (2021). Comparing alternative corrections for bias in the bias-corrected bootstrap test of mediation. *Evaluation & the Health Professions*, *44*(4), 416–427. doi: 10.1177/01632787211024356

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: McGraw:Hill.

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

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

Cohen, T. R., Wolf, S. T., Panter, A. T., & Insko, C. A. (2011). Introducing the gasp scale: A new measure of guilt and shame proneness. *Journal of personality and social psychology*, *100*(5), 947–966. doi: 10.1037/a0022641

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2021). Avoid cohen's 'small', 'medium', and 'large' for power analysis. *Trends in Cognitive Sciences*, *3*, 200–207.

Coutts, J. (2023). *Contrasting contrasts: An exploration of methods for comparing indi-*

rect effects in mediation models (Unpublished doctoral dissertation). The Ohio State University.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., … Sculley, D. (2022). Underspecification presents challenges for credibility in modern machine learning. *J. Mach. Learn. Res.*, *23*(1), 61.

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.

Durand, C. P. (2013). Does raising type 1 error rate improve power to detect interactions in linear regression models? a simulation study. *PLoS ONE*, *8*(8), e71079. doi: 10.1371/journal.pone.0071079

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 and 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

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Method*, *39*(2), 175–191. doi: 10.3758/bf03193146

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.

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*. doi: 10.1177/1948550615612150

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A monte carlo study comparing dwls and uls estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 625–641. doi: 10.1080/10705510903203573

Fossum, J. L., & Carpenter, T. P. (in prep). Talk about feeling bad: How guilt- and shame-proneness predict disclosure.

Fossum, J. L., & Montoya, A. K. (2023). When to use different tests for power analysis and data analysis for mediation. *Advances in Methods and Practices in Psychological Science*.

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

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem even when there is no ”fishing expedition” or ”p-hacking” and the research hypothesis was posited ahead of time. *Unpublished*.

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. (2013). *Introduction to mediation, moderation, and conditional process analysis*. New York, NY: Guilford Press.

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. (2018a). *Introduction to mediation, moderation, and conditional process analysis* (2nd ed.). New York, NY: Guilford Press.

Hayes, A. F. (2018b). Partial, conditional, and moderated moderated mediation: Quantification, inference, and interpretation. *Communication Monographs*, *85*, 4–40. doi: 10.1080/03637751.2017.1352100

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524-532. doi: 10.1177/0956797611430953

Judd, C. M., & Kenny, D. A. (2010). *Data analysis in social psychology: Recent and recurring issues.* Hoboken, NJ: John Wiley & Sons, Inc. doi: 10.1002/9780470561119.socpsy001004

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/`

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 233–265). Boston: McGraw-Hill.

Koenker, R. (1994). Confidence intervals for regression quantiles. In P. Mandl & M. Hušková (Eds.), *Asymptotic statistics.* Heidelberg: Physica. doi: 10.1007/978-3-642-57984-4$_2$9

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PloS one*, *9*(9), e105825. doi: 10.1371/journal.pone.0105825

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267. doi: 10.1525/collabra.33267

Lavery, M. R., Acharya, P., Sivo, S. A., & Xu, L. (2019). Number of predictors and multicollinearity: What are their effects on error and bias in regression? *Communications in Statistics - Simulation and Computation*, *48*(1), 27–38. doi: 10.1080/03610918.2017.1371750

Lindsey, L. L. M., Yun, K.-A., & Hill, J. B. (2007). Anticipated guilt as motivation to help unknown others: An examination of empathy as a moderator. *Communication Research*, *34*(4), 468–480. doi: 10.1177/0093650207301319

Lomnicki, Z. A. (1967). On the distribution of products of random variables. *Journal of the Royal Statistical Society*, *29*, 513–524.

Ma, Z.-w., & Zeng, W.-n. (2014). A multiple mediator model: Power analysis based on monte carlo simulation. *American Journal of Applied Psychology*, *3*(3), 72–79. doi: 10.11648/j.ajap.20140303.15

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83-104.

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*, 99 − 128.

Mackinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*(1), 41.

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

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

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

Orth, U., Robins, R. W., & Soto, C. J. (2010). Tracking the trajectory of shame, guilt, and pride across the life span. *Journal of Personality and Social Psychology*, *99*(6), 1061. doi: 10.1037/a0019615

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–710. doi: 10.1093/biomet/82.4.669

Pearl, J. (2001). *Direct and indirect effects*. San Francisco, CA: Morgak Kaufman.

Pieters, R. (2017). Meaningful mediation analysis: Plausible causal inference and informative communication. *Journal of Consumer Research*, *44*(3), 692–716.

Preacher, K. J., & Hayes, A. F. (2004). Spss and sas procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods*, *36*, 717–731. doi: 10.3758/BF03206553

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891. doi: 10.3758/BRM.40.3.879

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

Rasoolimanesh, S. M., Wang, M., Roldán, J. L., & Kunasekaran, P. (2021). Are we in right path for mediation analysis? reviewing the literature and proposing robust guidelines. *Journal of Hospitality and Tourism Management*, *48*, 395–405. doi: 10.1016/j.jhtm.2021.07.013

Rohrer, J. M., & Aslan, R. C. (2021). Precise answers to vague questions: Issues with interactions. *Advances in Methods and Practices in Psychology*, *4*(2), 1–19. doi: 10.1177/25152459211007368

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, *25*, 127–141. doi: 10.1002/sim.2331

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of American Statistical Association*, *100*(469), 322–331. doi: 10.1198/016214504000001880

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

Selig, J. P., & Preacher, K. J. (2008). Monte carlo method for assessing mediation: An interactive tool for creating confidence intervals for indirect effects [computer software].

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

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, *7*, 422–445.

Sim, M., Kim, S. Y., & Suh, Y. (2022). Sample size requirements for simple and complex mediation models. *Educational and Psychological Measurement*, *82*(1), 76–106.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290–312. doi: 10.2307/270723

Stine, R. (1989). An introduction to bootstrap methods: Examples and ideas. *Sociological Methods and Research*, *18*(2-3), 243-291. doi: 10.1177/0049124189018002003

Sun, J., Kaufman, S. B., & Smillie, L. D. (2018). Unique associations between big five personality aspects and multiple dimensions of well-being. *Journal of Personality*, *86*(2), 158–172. doi: 10.1111/jopy.12292

Sun, S. (2011). Meta-analysis of cohen's kappa. *Health Services and Outcomes Research Methodology*, *11*(3-4), 145–163. doi: 10.1007/s10742-011-0077-3

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*, 345–372. doi: 10.1146/annurev.psych.56.091103.070145

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. doi: 10.1177/0261927X09351676

Thoemmes, F. J., & Ong, A. D. (2015). A primer on inverse probability of treatment weighting and marginal structural models. *Emerging Adulthood*, *4*(1), 40–59. doi: 10.1177/2167696815621645

Tibbe, T. D., & Montoya, A. K. (2022). Correcting the bias correction for the bootstrap confidence interval in mediation analysis. *Frontiers in Psychology*, *13*. doi: 10.3389/fpsyg.2022.810258

Tignor, S. M., & Colvin, C. R. (2019). The meaning of guilt: Reconciling the past to inform the future. *Journal of Personality and Social Psychology*, *116*(6), 989–1010. doi: 10.1037/pspp0000216

van Ggndy, S., & Van Gundy, K. (2000). The personal and social links between age and self-reported empathy. *Social Psychology Quarterly*, *63*(2), 152–174. doi: 10.2307/2695889

van Zwet, E. W., & Goodman, S. N. (2022). How large should the next study be? predictive power and sample size requirements for replication studies. *Statistics in Medicine*, *41*(16), 3090–3101. doi: 10.1002/sim.9406

van Zyl, L. E., & ten Klooster, P. M. (2022). Exploratory structural equation modeling: Practical guidelines and tutorial with a convenient online tool for mplus. *Frontiers in Psychiatry*, *12*, 795672. doi: 10.3389/fpsyt.2021.795672

Vize, C., Baranger, D. A. A., Finsaas, M. C., Goldstein, B. L., Olino, T. M., & Lynam, D. R. (2023). Moderation effects in personality disorder research. *Personality Disorders: Theory, Research, and Treatment*, *14*(1), 118–126. doi: 10.1037/per0000582

Vize, C., Sharpe, B. M., Miller, J., Lynam, D., & Soto, C. J. (2022, June 28). Do the big five personality traits interact to predict life outcomes? systematically testing the prevalence, nature, and effect size of trait by trait moderation. (Preprint) doi: 10.31234/osf.io/b5ph8

Williams, J., & MacKinnon, D. P. (2008). Resampling and distribution of the product methods for testing indirect effects in complex models. *Structural Equation Modeling*, *15*, 23 – 51.

Yang, Y., & Green, S. B. (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(1), 66–81. doi: 10.1080/10705510903438963

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.

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