The Power of Design: Impact of Experimental Design Outweighs Impact of Inferential Methods on Statistical Power to Detect Indirect Effects

Amanda K. Montoya

University of California, Los Angeles

Author Note

Amanda Kay Montoya, Department of Psychology, UCLA.

Correspondence concerning this article should be addressed to Amanda Montoya, Department of Psychology, University of California, Los Angeles, 502 Portola Plaza, Los Angeles, CA 90095. Email: akmontoya@ucla.edu.

The Power of Design: Impact of Experimental Design Outweighs Impact of Inferential Methods on Statistical Power to Detect Indirect Effects

Abstract

Mediation analysis is a popular statistical method throughout psychology, communication, business and other behavioral sciences. However, previous research has demonstrated that that typical sample sizes are too small to have high power to detect indirect effects in between-subject designs; however, some inferential methods have higher power than others. Montoya and Hayes (2017) developed new methods for estimating and conducting inference on indirect effects in within-subject designs. For tests of means, within-subject designs boast power advantages over between-subject designs; however, this advantage has not been demonstrated for mediation analysis. Monte Carlo simulation is used to compare six inferential methods, two designs, and a broad range of sample sizes, effect sizes, and correlations among repeated measurements. The results suggest within-subject designs require about half the sample size of between-subject designs to detect indirect effects of the same size, and the effect of design is much greater than the effect of inferential method. Factors which can impact power (e.g. highly correlated measurements) are discussed, and how to determine if a within-subject design is appropriate for a given research question. MEMORE is an easy to use tool for estimating and conducting inference on indirect effects in two-condition within-subject designs.

Keywords: Mediation analysis, indirect effect, within-subject design, power, type I error, Monte Carlo simulation.

Word Count: 10,392

Much of the debate in mediation analysis regarding statistical power is focused on using different inferential methods (e.g., bootstrapping, product of the coefficients method, delta method; Biesanz, Falk, & Savalei, 2010; Fritz & MacKinnon, 2007; Hayes & Scharkow, 2013; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). However, there has been very limited discussion about how experimental design can impact power for mediation analysis. This may be largely due to a lack of comparable approaches to mediation analysis across different designs. A two condition within-subject design serves as the within-subject counterpart to a two-condition between-subject design. These two designs are frequently compared with respect to the effect of condition on an outcome (paired vs. independent samples t-test). Using a within-subject design can improve statistical power for tests comparing two conditions by collecting observations on each participant in each experimental condition. This allows the participant to act as their own control, eliminating between participant variability from the standard error of the estimate of difference between conditions. The ability for within-subject designs to detect effects with greater precision than between-subject designs is well-documented for tests comparing condition means (Maxwell & Delaney, 2004; Senn, 1993; Venter & Maxwell, 1999). However, the advantage of within-subject designs is less clear for tests of mediation. The recent development of a path analytic approach to conducting mediation analysis for two condition within-subject designs (Montoya & Hayes, 2017), as well as the release of a free macro for SPSS and SAS to conduct these designs, MEMORE (available at akmontoya.com) has resulted in a rapid rise in the use of these models in psychology and other behavioral sciences (e.g., Damen, 2019; Gunia & Levine, 2019; Nguyen, Carnevale, Scholer, Miele, & Fujita, 2019; Pazda & Thostenson, 2019; Rousselet, Brial, Cadario, Béji-Bécheur, 2018; Thorstenson, Pazda, & Lichtenfeld, 2019). With this sudden increase in adoption, it is important to understand the implications of design in

quality of statistical decision making, as well as what factors may impact the quality of inference in mediation analysis.

Mediation analysis examines the indirect effect of some causal antecedent variable (*X*) on an outcome (*Y*) through a proposed intermediary variable (*M*, a mediator). Mediation analysis is exceedingly common in psychology research and other behavior sciences (Hayes & Scharkow, 2013; Yzerbyt et al., 2018). This analysis is particularly useful for understanding the mechanisms by which a manipulation may influence an outcome. For example, clinical psychology is very focused on understanding the psychological mechanisms by which behavioral therapies impact psychological disorders. Education researchers may want to understand the mechanisms by which certain educational practices influence student learning. Understanding mechanisms is central to many of the research questions throughout psychological and behavioral research, and mediation analysis is often useful for providing insight into these mechanisms.

This article focuses on two specific experimental designs called the *two condition between-subject design* and the *two condition within-subject design*. In the two condition between-subject design, *X* is randomly assigned to participants (i.e., 50% of participants are assigned to Condition 1 and 50% are assigned to Condition 2), and for the purposes of mediation analysis each participant is observed once on the proposed mediator *M* and the proposed outcome *Y*. This design is very common in psychology and other behavioral sciences, and it is one of the most common designs for the use of mediation analysis. For comparison, consider the two condition within-subject design, which is the repeated-measures equivalent of the two condition between-subjects design. In the two condition within-subject design all participants experience both of two experimental conditions and are observed in both of these two conditions on the mediator (*M*) and the outcome (*Y*). The aim of this paper is to determine the impact of experimental design

(between vs. within-subject) on the statistical power to detect an indirect effect. The impact of experimental design is compared to the impact of inferential method, as much previous research has suggested that different inferential methods for detecting indirect effects can have different statistical power (Biesanz, Falk, & Savalei, 2010; Fritz & MacKinnon, 2007; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002, Yzerbyt, et al., 2018). Additionally, little is known about what characteristics in a within-subject design might impact power of mediation analysis, as the performance of methods for statistical inference on indirect effects has largely gone unstudied for this design (but see Yzerbyt et al., 2018). Monte Carlo simulation is used to investigate these factors, assisting researchers in identifying when power would be maximized in within-subject designs.

This paper begins with a brief overview of how to estimate indirect effects for the two condition between- and within-subject designs. I next discuss a variety of inferential methods for indirect effects which are well studied for between-subject designs. Monte Carlo simulations are used to compare the performance of the different inferential tests in both within- and between-subject designs under a variety of realistic conditions. The results of these simulations indicate that the inferential methods for between-subject designs which balance type I error and power well, also perform well for within-subject designs. Additionally, it is clear that the impact of design on power is very large: within-subject designs have much higher power to detect indirect effects than between-subject designs, and this effect is much larger than the differences between inferential methods. However, unique cases when power is greater in between-subject designs are also discussed. Additionally, the correlation among mediators can have a non-monotonic effect on power to detect the indirect effect (i.e., power increases as correlation increases to a point, but after that point power decreases as correlation increases); whereas, correlation among

outcomes also impacts power but in a monotonic manner. This paper concludes with recommendations for assessing mediation in within-subjects design, cautions about using within-subject designs, and suggestions for future directions.

Estimating Indirect Effects

Between Subject Designs

In a two condition between-subject design, individuals are randomly assigned to an experimental condition (*X*), and after they've experienced the manipulation they are measured on the mediator (*M*) and the outcome (*Y*). In this article, the focus is on studies with only two possible conditions, therefore *X* is a dichotomous variable. *M* and *Y* are continuous variables with normally distributed error terms. It is possible for *M* and *Y* to be categorical or non-normal variables; however, alternative estimation procedures would be needed. For more information about when *M* and *Y* are categorical or non-normal see Preacher (2015) for a review.

Indirect effects in two condition between-subject designs are typically estimated using either a linear regression (Ordinary Least Squares) or structural equation modeling (Maximum Likelihood) framework. The descriptions in this paper will rely on the OLS approach; however, the extension to structural equation models is straight forward, and for observed variable models the parameter estimates are exactly the same for OLS and ML approaches (Ledgerwood & Shrout, 2011; Hayes, Montoya, & Rockwood, 2017). A mediation model for two condition between-subject designs can be estimated using two equations:

$$M_i = a_0 + aX_i + e_{M_i} \quad (1)$$

$$Y_i = c_0 + c'X_i + bM_i + e_{Y_i} \quad (2)$$

Where X_i, M_i, and Y_i are the observations for the experimental manipulation, the mediator, and the outcome (respectively) for individual i. The intercepts are represented using a_0 and c_0. The errors are represented using e_{M_i} and e_{Y_i} and each assumed to be normally distributed with a mean of zero and a unique variance. The effect of the experimental manipulation on the mediator is represented by the coefficient a, the effect of the experimental manipulation on the outcome controlling for the mediator is represented by the coefficient c' (the direct effect), and the effect of the mediator on the outcome controlling for the experimental manipulation is represented by the coefficient b. These equations can be represented as a single path diagram (See Figure 1). Based on the diagram there are two pathways by which *X* can influence *Y*: through the direct effect, *c'*, and through the indirect effect where *X* influences *M* which then influences *Y*. The indirect effect is quantified as the product of the *a* and *b* paths: *ab*. The focus of mediation analysis is typically on the presence or absence of the indirect effect as well as its magnitude. An important property of mediation analysis with continuous mediator and outcome is that the two pathways by which *X* affects *Y* are a partition of the total effect, *c*, which is a quantification of the overall effect of *X* on *Y* (without controlling for any other variables). The total effect, *c*, can be estimated in its own regression equation:

$$Y_i = c_0 + cX_i + e_{Y_i^*}$$

The sum of the indirect and direct effects equals the total effect (at both the sample and population level):

$$c = c' + ab \quad (3)$$

Mediation analysis can be conducted using experimental or correlational data. However, mediation analysis requires an assumption of causal order such that the independent variable (*X*)

causes the mediator (*M*) which then causes the outcome (*Y*). Only if this specific ordering of variables is true is mediation analysis informative. Therefore, mediation analysis is most appropriate when *X* is an experimentally manipulated variable and the mediator and outcome (*M* and *Y*) are measured downstream outcomes. However, even when *X* is experimentally manipulated, there is no guarantee that the mediator occurs before the outcome, and so the order assumptions of mediation analysis are not necessarily met through experimental manipulation of *X*. Pirlott and MacKinnon (2015) provide additional ways to improve the validity of causal claims through mediation analysis.

Within-Subject Designs

An equivalent experiment can be run where instead of randomly assigning individuals to a condition, each individual experiences both conditions and is measured on the mediator and the outcome in both conditions. Judd, Kenny, and McClelland (2001) were the first to outline a procedure for evaluating mediation in such models, and Montoya and Hayes (2017) showed how to represent these models as path models, estimate the indirect effect, and conduct inference using this estimate.

For these models, each individual has an observation on the mediator in both of the two experimental conditions (M_{1i} and M_{2i}). Similarly, each individual has an observation on the outcome in both of the two experimental conditions (Y_{1i} and Y_{2i}). To estimate the influence of the experimental manipulation on the mediators, an intercept-only model of the difference between the mediators can be estimated where the intercept, *a*, is the estimate of the average difference between M_{1i} and M_{2i}.

$$M_{2i} - M_{1i} = a + e_{M_i} \quad (4)$$

Similarly, a model for the difference in the outcomes provides an estimate of the influence of the mediator on the outcome controlling for experimental condition (*b*), and the intercept (*c'*) estimates the influence of the experimental manipulation on the outcome controlling for the mediators.

$$Y_{2i} - Y_{1i} = c' + b(M_{2i} - M_{1i}) + d(M_{2i} + M_{1i})^* + e_{Y_i} \quad (5)$$

In the above equation the * indicates that the sum of the mediators has been grand mean centered (i.e., $(M_{2i} + M_{1i})^* = (M_{2i} + M_{1i}) - \overline{(M_{2\iota} + M_{1\iota})}$, where $\overline{(M_{2\iota} + M_{1\iota})} = \sum_{i=1}^{N}(M_{2i} + M_{1i})/N$ is the sample mean of the sum of the mediators). This grand mean centering is what allows the intercept to be interpreted as the direct effect (Judd, Kenny, and McClelland, 2001; Montoya & Hayes, 2017). The coefficient *d* represents the degree to which the relationship between M_1 and Y_1 differs from the relationship between M_2 and Y_2. This coefficient is described as a moderation parameter because it reflects the degree to which the *M*-*Y* relationship differs across conditions. Similar to the between-subjects case, these two equations can be represented as a path model (See Figure 2). Just as in the between-subjects case, the focus of inference for mediation is on the indirect effect, *ab*. In the within-subjects case, the total effect can be estimated by using an intercept only model to predict the difference between the outcomes:

$$Y_{2i} - Y_{1i} = c + e_{Y_i^*}$$

The total effect as estimated above perfectly partitions into the direct effect and indirect effect such that Equation 3 also holds for the within-subject case.

Methods of Inference

There are a number of popular methods of inference for between-subject mediation and a growing body of research examining the performance of these different methods of inference for

mediation in between-subject designs. This section describes the most common methods of inference for between-subject mediation, outlines how they are conducted in between-subject designs, and how they may be generalized to within-subject designs. All of the described methods for testing indirect effects in within-subjects designs can be implemented in MEMORE a freely-available macro for SPSS and SAS (Montoya & Hayes, 2017). This section concludes with a short review of the simulation research examining the performance of these inferential methods, which helps guide expectations for how these methods might perform in within-subject designs.

Causal steps method. The causal steps method was introduced by Baron & Kenny (1986). According to this method, if the *c*-path, *a*-path, and *b*-path are all significantly different than zero then the researcher has evidence for mediation. This method has been criticized for a variety of reasons, one of which is the requirement that the total effect be significant to make claims of mediation (Collins et al., 1998; Kenny, Kashy, Bolger, 1998; MacKinnon, 2008; MacKinnon et al., 2000; Shrout & Bolger, 2002). It is possible for the signs of the direct and indirect effects to be opposite each other, producing a non-significant total effect, which does not indicate a lack of mediation, but rather that opposing forces are operating through *X* making the total effect small or nonexistent (MacKinnon, 2008; O'Rourke & MacKinnon, 2018; Shrout & Bolger, 2002). Additionally, in some circumstances there is more power to detect an indirect effect than a total effect even when the magnitude of the two effects is the same (i.e., $ab = c$, Shrout & Bolger, 2002; Judd & Kenny, 2014). Another criticism of this approach is that it is not based on an estimate of the indirect effect, rather the inferential test is a logical progression of tests on single paths rather than a single test on an estimate of the indirect effect. This method does not provide estimates of the standard error of the indirect effect, nor can a researcher create a confidence

interval for the indirect effect using this method (Hayes, 2018; MacKinnon, 2008; MacKinnon, et. al., 2002).

Judd, Kenny, and McClelland (2001) provided instructions on how to conduct these steps in two condition within-subject designs. Similar to Baron and Kenny (1986), the *c*-, *a*-, and *b*-paths must all be statistically significant to support a claim of mediation. All of the criticisms of the causal steps method as it is used in between-subject designs also apply to its use in within-subject designs.

Test of joint significance. The test of joint significance is very closely related to the causal steps method (Kenny et al., 1998; MacKinnon, 2002). This test does not impose the requirement that the total effect be significantly different from zero, rather both the *a* and *b* paths are required to be significantly different from zero. The criticisms of the causal steps method related to the restrictive requirement that *c* be significantly different from zero have been alleviated through the use of this test; however, others criticisms still remain: in particular, the test relies on a series of logical steps (i.e., if *a* is not zero and *b* is not zero then the *ab* is not zero) rather than conducting the test on the effect of interest, the indirect effect (but see Yzerbyt, et al., 2018).

To conduct this test in a within-subjects design, significance tests are used for *a* and *b*. If *a* and *b* are both significantly different from zero, a claim of mediation can be made. There is no requirement that there be a significant effect of condition on the outcome variable (i.e., *c* need not be significant). This would eliminate some of the criticisms of the Judd, Kenny, and McClelland (2001) method, but does not reduce concerns about conducting a test on a direct quantification of the indirect effect, or the lack of standard errors and confidence intervals for the indirect effect.

Normal theory tests. The normal theory tests (commonly known as Sobel tests or Delta method) are tests of significance on the indirect effect (*ab*) which assume that the sampling distribution of *ab* is normal. Normal theory tests use an asymptotic standard error derived from the multivariate delta method to compute a Z-score (ab/se_{ab}) which can be used to compute a *p*-value which can be compared to a previously determined significance level (e.g. $\alpha = .05$). There are multiple derived standard errors for this test, most popularly the Sobel (Sobel, 1982, 1986) and Aroian standard errors (Aroian, 1947; based on the first and second degree Taylor expansions, respectively). One of the main criticisms for the normal theory tests is the assumption of normality, which has been shown to be violated in finite samples (Stone & Sobel, 1990); however, MacKinnon and Dwyer (1993) showed that both the Sobel (Equation 6) and Aroian (Equation 7) estimates of the standard error were very close to simulated population standard errors in samples of size 50 for one set of population parameters *a*, *b*, and *c'*.

$$se_{ab1} = \sqrt{a^2 se_b^2 + b^2 se_a^2} \tag{6}$$

$$se_{ab2} = \sqrt{a^2 se_b^2 + b^2 se_a^2 + se_b^2 se_a^2} \tag{7}$$

Though the standard errors may be accurate, the shape of the distribution is still non-normal; the distribution of the indirect effect has skew and kurtosis which can be calculated as a function of population parameters for the *a* and *b* paths. The distribution of the product of two regression coefficients is not estimated accurately using a normal distribution with a limited sample size, and thus this method may not perform as well as other methods with different or fewer assumptions about the sampling distribution of the indirect effect.

The *a* and *b* path estimates along with their standard errors from the path analytic model for within-subjects designs can be used in either Equation 6 or 7 to compute the estimated standard error of the indirect effect. The standard error for *a* and *b* can be found in the output of any statistical package used to estimate the models. Overall, the extension of the normal theory tests to within-subject designs is easily implemented by using the path estimates and standard errors from the path analytic model proposed for a single mediator within-subject mediation model. MEMORE can be used to conduct this test by calculating the standard error of the indirect effect (using Equation 7), the Z-value, and the *p*-value.

Bootstrap confidence intervals. Another inferential test for the indirect effect is a bootstrap confidence interval (Preacher & Hayes, 2004; Shrout & Bolger, 2002). This method generates *K* bootstrap estimates of the indirect effect by sampling from the observed data with replacement, generating *K* bootstrap samples of size *N*, where *N* is the number of cases in the original data and *K* is recommended to be a large number (e.g. 1,000). An estimate of the indirect effect is calculated from each bootstrap sample using the regression methods in Equations 1 and 2, generating *K* estimates of the indirect effect. The distribution of the bootstrapped estimates provides an approximate sampling distribution for the indirect effect (*ab*). There are a few different methods for generating confidence intervals from this estimated sampling distribution, including percentile bootstrap and bias-corrected bootstrap. The percentile bootstrap method sorts the estimates from smallest to largest and selects the upper and lower bounds for the confidence interval such that $\frac{\alpha}{2}$% of the bootstrapped estimates lie below the lower bound and $(100 - \frac{\alpha}{2})$% of the bootstrapped estimates lie above the upper bound, where α is the predetermined level of the inferential test (Efron & Tibshirani, 1993). The percentile bootstrap assumes that the estimate of indirect effect is unbiased (Mooney & Duval, 1993), so the bias-

corrected bootstrap can be used to create a confidence interval without relying on this assumption. The bias-corrected bootstrap takes into account the percentage of the bootstrap estimates below the estimate of the indirect effect from the observed sample and adjusts the confidence interval to correct for bias of the estimated sampling distribution (Mooney & Duval, 1993).

Applying bootstrap confidence intervals to a within-subjects design uses the same procedure for generating *K* bootstrap samples, where the observed data are the difference in outcomes, difference in mediators, and average of the mediators. The average of the mediators should be mean-centered in each bootstrap sample if the researcher would like bootstrap estimates of any function of the direct effect *c'*. The methods outlined in Equations 4 and 5 can be used to estimate the *a* and *b* paths, then the product of these paths can be calculated to estimate the indirect effect. This procedure will generate *K* estimates of the indirect effect, and the percentile or bias-corrected methods for selecting the upper and lower bounds for the confidence interval can be used to calculate a confidence interval for the indirect effect in a within-subjects design.

Monte Carlo confidence intervals. The final method of inference discussed in this paper is the Monte Carlo confidence interval (Preacher & Selig, 2012). This method uses the estimates of the *a* and *b* paths and their standard errors to simulate sampling distributions for both *a* and *b*. These simulated estimates are then used to create an estimated sampling distribution for the indirect effect. To create a Monte Carlo confidence interval, a sample of size *K*, where *K* is recommended to be large (e.g. 1,000), is generated from a random normal distribution with mean *a*, where *a* is the point estimate of the *a*-path from the observed sample, and standard deviation se_a, where se_a is the estimated standard error of the *a*-path calculated from observed sample. These estimates approximate the sampling distribution of the *a*-path under the assumptions of

linear regression. Next, a sample of size *K* is generated from a random normal distribution with mean *b*, where *b* is the point estimate of the *b*-path from the observed sample, and standard deviation se_b, where se_b is the estimated standard error of the *b* path calculated from the original sample. These estimates approximate the sampling distribution for the *b* path under the assumptions of linear regression. The estimates of the *a* path and *b* path are then multiplied together listwise, such that the first estimate from the *a* distribution is multiplied by the first estimate from the *b* distribution, to generate a sample of *K* estimates of the indirect effect. This distribution can be used as an approximation of the sampling distribution of the indirect effect. The estimates of the indirect effect are sorted from smallest to largest, selecting the upper and lower bounds for the confidence interval such that $\frac{\alpha}{2}\%$ of the estimates lie below the lower bound and $(100 - \frac{\alpha}{2})\%$ of the estimates lie above the upper bound of the confidence interval.

Monte Carlo confidence intervals have a few unique advantages and disadvantages. First, they do not require the original data to generate confidence intervals for the indirect effect. Only the path estimates and their standard errors are required to create the confidence interval. However, Monte Carlo confidence intervals rely on the assumption that the distributions of the *a*-path and *b*-path are normally distributed, which is true in large enough samples without excessive violations of the assumptions of linear regression. Any violations of regression assumptions that would bias the estimates of the standard errors (e.g., outliers, heterskedasticity) would influence the Monte Carlo confidence interval method. As such, methods that do not rely on the normality assumption of the individual path estimates and their standard errors may perform better than the Monte Carlo confidence interval, particularly when there are assumption violations.

Similar to the extension of the normal theory tests in within-subject designs, extending the Monte Carlo confidence interval to within-subjects designs only requires the researcher use

estimates of the paths and the standard errors of the paths which are appropriate to the design. As such, *a* would be the mean difference between the two mediator variables, and *b* would be the partial regression coefficient *b* from Equation 5. The standard errors to be used are the same as those to be used in calculating the asymptotic variance of the indirect effect for the normal theory tests (Equations 6 and 7). The rest of the Monte Carlo procedure can be carried out as it was in the between-subjects case.

Previous simulation results for methods of inference. Because of the large number of methods of inference, it is important for researchers to understand which methods of inference perform well in different situations and in what situations methods fail. A variety of simulation studies have compared the performance of different tests of inference for the indirect effect in between-subjects designs (Biesanz Falk, Savalei, 2010; Fritz & MacKinnon, 2007; Hayes & Scharkow, 2013; MacKinnon et al., 2002; Stone & Sobel, 1990; Yzerbyt, Muller, Batailler, & Judd, 2018). The following results only explore between-subject studies, and the purpose of this study is to examine if these same differences among inferential methods hold in within-subject designs.

Some studies exclude the causal steps method because of its limitations: no estimate of the indirect effect or its standard error as well as a general agreement among mediation researchers that there should be no requirement that the total effect be significant in order to support a claim of mediation (Collins et al., 1998; Hayes, 2013; MacKinnon, 2008; MacKinnon et al., 2000; Hayes & Rockwood, 2017; Shrout & Bolger, 2002). One study which did examine the causal steps method found it had conservative type I error and was underpowered (MacKinnon et al., 2002). The joint significance test has been shown to have conservative type I error but good power in a variety of conditions (Biesanz, Falk, & Savalei, 2010; MacKinnon et al., 2002; Yzerbyt, et al., 2018).

Normal theory tests have lower power than bootstrapped confidence intervals and Monte Carlo confidence intervals (Biesanz, Falk, Savalei, 2010; Fritz & MacKinnon, 2007; MacKinnon, Lockwood, and Williams, 2004). Additionally, normal theory tests are slightly biased, particularly in small samples and with small non-zero indirect effects (MacKinnon, Lockwood, Williams, 2004; Stone & Sobel, 1990). The different standard errors used for the normal theory tests do not differ greatly in performance and are nearly equivalent (MacKinnon & Dwyer, 1993; Stone & Sobel, 1990).

The Monte Carlo confidence interval works well as an inferential test for the indirect effect in the between-subject mediation literature (Hayes & Scharkow, 2014; MacKinnon, Lockwood, & Williams, 2004; Preacher & Selig, 2012). The Monte Carlo confidence interval and percentile bootstrap confidence interval have been shown to have lower than expected (α) type I error and lower power than some other methods like the bias-corrected bootstrap confidence interval, suggesting the confidence intervals are too wide resulting in slightly conservative tests and reduced power (Hayes & Scharkow, 2013; MacKinnon, Lockwood, Williams, 2004).

Different bootstrapping methods have been compared by not only type I error and power, but also bias in coverage, which is when the confidence interval misses the true value more often on one side than the other. Though the bias-corrected bootstrapping method has been shown to have higher power than the percentile bootstrap and lower coverage bias, this comes at the cost of an increase in type I error (Biesanz, Falk & Savalei, 2010; Hayes & Scharkow, 2013; MacKinnon, Lockwood, & Williams, 2004). The percentile bootstrap performs well under non-normality and with incomplete data (Biesanz, Falk, and Savalei, 2010). Biesanz and colleagues (2010) recommend the percentile bootstrap or hierarchical Bayes (a method not discussed in this paper) due to their stability in type I error, coverage, and power, under conditions of incomplete data

and non-normality. Similarly, in their review of a variety of methods Hayes and Scharkow (2013) concluded that the percentile bootstrap and the Monte Carlo confidence interval succeed in providing the best balance of type I error, power, and coverage bias. Yzerbyt and colleagues (2018) add to this list the joint significance test which performs similarly but still has certain weaknesses with regard to completeness of information.

Because tests of the indirect effect have a complex null hypothesis, either *a* or *b* is zero, researchers have focused on how different combinations of the *a* and *b* paths influence type I error, specifically comparing when either *a* or *b* is zero (but not both) to when both *a* and *b* are zero. Researchers have found that most methods are very conservative when both the *a* and *b* paths are zero (Biesanz, Falk, & Savalei, 2010; MacKinnon et al., 2002); however, performance differences between tests are more noticeable when either *a* or *b* is zero but not both. For example, Biesanz, Falk, and Savalei (2010) found that percentile bootstrapping was one of the only methods to have an acceptable type I error under a variety of conditions where either the *a* or the *b* path was zero but not both. Researchers have found that methods like the joint-significance test and the causal steps method, which rely on multiple testing, tend to have conservative type I error rates when both *a* and *b* are zero, but when *a* or *b* is non-zero the type I error rate increases as the nonzero path increases, even though the size of the indirect effect is not actually increasing but rather remains zero (Biesanz, Falk, & Savalei, 2010; MacKinnon et al., 2002). Because of this important role of the two types of null hypotheses, both situations are investigated using Monte Carlo simulations.

There are many approaches to inference for indirect effects which have been heavily studied for between-subject designs. Each of these popular approaches to inference in between-subject mediation can be generalized to within-subject designs. Though the logic behind these methods

is fairly clear, often the computations can be arduous if they are not already implemented as commands in common statistical packages. Tools for between-subject mediation, such as PROCESS for SPSS and SAS and RMediate and Psych packages for R, have made calculation of complicated tests of the indirect effect quick and easy for researchers of all programming levels. MEMORE for SPSS and SAS fills this role for the two condition within-subject case (Montoya & Hayes, 2017).

Monte Carlo Simulation: Inferential Methods and Design

Many previous studies have compared the relative performance of methods of inference for mediation in between subjects designs, but only one previous study has done so for within-subjects designs (Yzerbyt, et al., 2018). Additionally, no previous studies have compared the relative performance of between and within-subject designs for mediation analysis. Yzerbyt, et al. (2018) explored the performance of Monte Carlo confidence intervals, bootstrapping methods (percentile, bias-corrected, and bias-corrected and accelerated), and the joint significance test for two condition within-subject designs. The present study expands on the methods and conditions explored by Yzerbyt, et al. (2018) in a variety of ways: additional inferential methods, broader range of sample size, variation of correlation among repeated-measurements, and variation of the moderation parameter, and comparing within- vs. between-subject designs. The present study adds the causal steps approach as proposed by Judd, Kenny, and McClelland (2001) and normal theory tests using both first and second-order Taylor expansions. This study does not examine bias-corrected or bias-corrected and accelerated methods as they have been shown to perform poorly in the within-subject case (Yzerbyt et al., 2018) and between-subjects case (Biesanz, Falk & Savalei, 2010; Hayes & Scharkow, 2013; MacKinnon, Lockwood, & Williams, 2004). This study expands the range of investigated sample sizes by Yzerbyt et al. (2018) by looking at both

smaller and larger samples. Additionally, a factor which was ignored in Yzerbyt et al. (2018) is the degree to which repeated measurements are correlated with each other. In this simulation, the correlation between the two measurements of the mediators and the correlation between the two measurements of the outcomes is systematically varied. Additionally, the moderation parameter, *d*, was always set to 0 in Yzerbyt et al. (2018); in this simulation, *d* is systematically varied and its impact on type I error and power is investigated. This study investigates how different methods of inference performed under a variety of reasonable circumstances regarding these unique characteristics of within-subject designs, as these have never been investigated before. Additionally, the simulation provides the opportunity to compare the effect of design to the effect of inferential method with regard to type I error and power.

For statistical analyses comparing group means, within-subject designs have distinct power advantages over between-subject designs for the same sample size. However, within-subject mediation analysis presents a unique case: the *a*-path is a within participant comparison, but the *b*-path is a between participant comparison. Thus, it is not clear if there will be power advantages of within-subject designs, and if there are power advantages, how large they would be. This study aims to determine if there are power advantages for within-subject designs, quantify how large they may be, and identify any cases where there are not power advantages.

Simulation Methods

A Monte Carlo simulation was used to assess the performance of the tests of inference on the indirect effect for both within- and between-subject designs. The simulation generates a large number of samples from conditions where the researcher assigns the population values of a variety of parameters. Each sample is analyzed using the statistical tests of interest. By

examining how these tests perform under repeated use in a known population, the researcher can make assessments about test characteristics like type I error and power in a variety of situations. For this simulation study seven parameters were manipulated, each of which might influence the rejection rates of the tests of inference on the indirect effect. Two thousand samples were generated in each condition (1,000 for within-subject design and 1,000 for between subject designs). There were 16,384 conditions leading to 32,768,000 samples total (See Footnote 1). There were seven factors: sample size, *a*-path size, *b*-path size, *c'*-path size, *d*-path size, correlation among mediators (ρ_M), correlation among outcome variables (ρ_Y) (See Table 1 for generating values). Values for the *a*-path, *b*-path, *c'*-path, and *d*-path were selected based on previous simulations for between-subject mediation (Biesanz, Faulk, Savalei, 2010; Hayes & Scharkow, 2013). The values for these paths correspond to small (2%), medium (13%), and large (26%) effects with respect to explained variance as defined by Cohen (1988). Values for ρ_M and ρ_Y were selected to span the range from uncorrelated repeated measures to very highly correlated repeated measures.

For each combination of possible conditions, an implied population correlation matrix and pattern of means were calculated. Some correlation matrices were not positive definite, so these conditions were dropped from the simulation[1]. From each population correlation matrix and pattern of means, samples were generated using a Cholesky decomposition and standard normal deviates. For each condition 2,000 samples were generated: 1,000 samples were used as within-subject designs where each individual had a measure for both mediators and both outcomes, and 1,000 samples were used as between-subject designs where each individual had a measure of one

[1] 9536 conditions had non-positive definite matrices. This resulted in 6848 valid conditions and 13,696,000 samples. It is important to note that these matrices were not equally distributed among the conditions, but rather were overrepresented when certain parameters were large (e.g. $\rho_M = .9$).

mediator and one outcome. For each sample the six tests of inference for the indirect effect were applied. For each test, whether or not the test rejected the null hypothesis that the indirect effect is zero was recorded, providing an overall rejection rate for each test in each condition.

There was a subset of conditions where the indirect effect in the population is zero. These include all conditions where the *a*-path or the *b*-path was set to zero at the population level. By examining the results for this subset, type I error rates of each test can be estimated. Type I error is calculated by computing the proportion of samples which were deemed significant by a specific method in each condition where the population indirect effect equals zero. Each method is specified at an α level of 0.05, so each method is expected to reject 5% of samples under conditions where there is no indirect effect. Two common criteria are used to evaluate empirical type I error rates when $\alpha = 0.05$: Bradley's (1987) liberal criteria (0.025 to 0.075) and Serlin's (2000) robustness criteria (0.035 to 0.065).

Additionally, by examining all the conditions where neither *a* nor *b* is zero, the relative power of each test can be estimated. Power is calculated by computing the proportion of samples which were deemed significant by a specific test in each condition where the population indirect effect does not equal zero. There is no specific power required for a test, though it is generally accepted that more power is better. The results are aggregated across manipulated conditions, but also useful and informative subsets of the data are broken apart to investigate cases manipulated factors may interact to influence type I error or power.

Inferential Tests

Causal steps method. The causal steps method endorsed by Judd, et al. (2001) or Baron & Kenny (1986) was used due to its popularity in both within- and between-subject designs. For this method, a hypothesis test was used to test if the total effect, *c*, was significantly different

from zero, if the *a*-path was significantly different from zero, and if the *b*-path was significantly different from zero. The null hypothesis that the indirect effect is zero was rejected if all three paths were statistically different from zero.

Test of joint significance. Both the *a* and *b* paths were tested using the regression methods described earlier (Equations 1 & 2 for between-subject and Equations 4 & 5 for within-subject). The null hypothesis that the indirect effect was zero is rejected if both paths were significantly different from zero.

Normal theory tests. Normal theory tests are tests of significance on the indirect effect (*ab*) which assume that the sampling distribution of *ab* is normal. There are multiple standard errors that can be used for normal theory tests depending on the order of the Taylor series expansion. Previous research suggests that choice of standard error does not have a large effect on decisions made in between-subject designs (MacKinnon & Dwyer, 1993). However, to explore this issue in within-subject designs as well, two different standard errors were included: the delta method first-order Taylor series approximation (Equation 25, Sobel, 1982) and the delta method second-order Taylor series exact solution (Equation 26, Aroian, 1947).

Bootstrap confidence intervals. Another method for testing the significance of the indirect effect is using a bootstrap confidence interval (Preacher & Hayes, 2004; Shrout & Bolger, 2002). For each sample, 1,000 bootstrap estimates of the indirect effect were generated. For this study, 95% confidence intervals using the percentile bootstrap method were calculated (Efron & Tibshirani, 1993). The bootstrap estimates are sorted from smallest to largest and used the 25th and 976th samples to generate a 95% confidence interval for the indirect effect. The indirect effect was deemed statistically different from zero if the bootstrap confidence interval did not include zero.

Monte Carlo confidence intervals. The final method was the Monte Carlo confidence interval, a method that has been shown to work well for tests of indirect effects in the between-subject mediation literature (Hayes & Scharkow, 2014; Preacher & Selig, 2012). In this simulation, 1,000 estimates of the indirect effect were generated for each sample, then estimates were sorted from smallest to largest, and the 25th and 976th element in the list were selected to generate a 95% confidence interval for the indirect effect. The indirect effect was deemed significantly different from zero if the Monte Carlo confidence interval did not include zero.

Simulation Results

The rejection rate was calculated for each method in each condition in order to examine type I error and power. Finally, the rejection rate for each individual path was calculated for each condition. Complete simulation data and code is available at osf.io/3njw6/. Overall, many of the simulation findings in the within-subject design replicate findings from simulations for between-subject mediation. Additionally, this simulation provides an opportunity to examine trends specific to within-subjects, for example, how correlation among repeated measures may effect power of different tests.

Within-Subject Test Performance

Type I error. All samples generated from populations where the indirect effect is zero (e.g., a = 0 and b = .39) were used to calculate type I error. Correlation among the mediators and among outcomes had no notable influence on Type I errors and so the results are not described in this article and results are aggregated across these factors. Complete simulation results are available online (osf.io/3njw6/). Type I error was calculated as the proportion of the generated samples where tests of the null hypothesis were rejected. For each sample size there were 812 conditions total with 1,000 samples per condition. The conditions were grouped by the type of true null

hypotheses: either *a* and *b* were both zero, or *a* or *b* (but not both) was zero and the other path was non-zero (e.g., *a* = .59, *b* = 0). Type I error across these two null hypotheses are reported in Tables 2 and 3. Type I errors outside of Bradley's criteria (worst performers) are marked in italics, and those inside Serlin's (2000) criteria are marked in bold (best performers). In Tables 2 and 3 the inferential test with type I error rate closest to .05 is underlined on each row, to examine trends of "best" performance. In general, type I error increased as sample size increased. No methods exceeded Serlin's robustness criteria, indicating that no methods had overly high Type I Error, with the highest type I error rate as 0.0592 (Percentile bootstrap, *N* = 100, *b* = 0, *a* = .59). Based on these results, it seems that all of the tests are either accurate or conservative. Additionally, percentile bootstrap, Monte Carlo confidence intervals, and the joint significance test clearly perform best with respect to closest desired level of the test. Percentile bootstrap was closest to the desired level in 61% of cases, Joint Significance in 28% of cases, and Monte Carlo confidence interval in 11% of cases.

Type I error rates can be examined among two meaningful groups: conditions in which both *a* and *b* are zero and cases in which *a* or *b* but not both are zero. In the simulation, when *a* and *b* are both zero, all tests are very conservative, with all Type I Errors at or below .0038 (See Tables 2 and 3). Type I error increases slightly as sample size increases for the causal steps method and the test of joint significance (.0007 to .002 and .0024 to .0028 respectively). Alternatively, the Type I error rate decreases for the normal theory tests (.0004 to .0001 [Sobel] and .0003 to .0001 [Aroian]), bootstrap confidence intervals (.0038 to .0019), and Monte Carlo confidence interval (.0029 to .0017) with increasing sample size (20 to 200). In comparison, the performance across tests varies more when *a* or *b* but not both are zero. Comparing Type I error when both *a* and *b* are zero to when one path is zero and the other is large (e.g., *a* = 0, *b* = .59), for all tests this

differences is very large, often multiple orders of magnitude, and increases with increasing sample size. For example, the Monte Carlo confidence interval with sample size of 20 has a type I error of .0029 when both *a* and *b* are zero, but a type I error of .0412 when *a* = 0 and *b* = .59 and .0374 when *a* = .59 and *b* = 0, a difference of .0383 (*a* = 0) and .0345 (*b* = 0). This difference is .0493 (*a* = 0) and .0526 (*b* = 0) when sample size is increased to 200. When either *a* or *b* but not both are zero, the causal steps method and the two normal theory tests are more conservative than the test of joint significance, the bootstrap confidence interval, and the Monte Carlo confidence interval, which all seem to approach the desired rejection rate of 0.05 as sample size increases.

Power: Aggregate power rate was calculated for each inferential test by averaging the rejection rates across all conditions where the population indirect effect did not equal zero, grouped by sample size (592 conditions per cell). Unsurprisingly, power increases as sample size increases (See Figure 7) and as the indirect effect increases for all tests (See Table 6). Overall, the test of joint significance, percentile bootstrap confidence interval, and Monte Carlo confidence intervals have similarly high power, and the causal steps method and normal theory tests are under-powered (typically 4 – 7% lower).

Factors which Impact Power. Based on the simulation results, a number of factors impact the power of each test to detect the presence of the indirect effect. Similar to previous findings from between-subject mediation simulations, as sample size increases, power increases. Additionally, as the indirect effect increases, power increases for all methods of analysis. The *c'* path only affected power for the causal steps method, because the causal steps method requires a significant test of the *c* path (where $c = c' + ab$) to conclude there is an indirect effect. As *c'* increases the power of the causal steps method increases.

Correlation among Mediators. One factor in this simulation that is unique to within-subject designs is the correlation among the repeated measurements. Because the between-subject designs only have one measurement for each participant, all measurements are independent, but in within-subject designs this is not the case. In this study, power is investigated by varying the correlation among the mediator variables in 0.3 increments (0.0 - 0.9).

Power was evaluated by selecting all conditions with a non-zero indirect effect. The rates of rejection were averaged over all conditions with the same level of correlation among mediators and sample size (See Table 4). Figure 3 provides a visualization of the pattern of power across correlations among the mediators for $N = 100$. For all tests, power increases up to $\rho_M = .6$ then the recorded power is lower for $\rho_M = .9$. An exception to this pattern are the two normal theory tests at N = 20 and N = 50 which show monotonic increases in power. However, these methods exhibit the same non-monotonic pattern as the other methods at higher sample sizes. Generally, though, across methods, there seems to be a point at which increased correlation among the mediators is detrimental to power for all tests. Supplemental simulations suggest the peak of power is around $\rho_M = .75$.

Correlation among Outcome Variables. Similar to the correlation among the mediators, there can also be correlation among the outcome variables. In this study, power is investigated by varying the correlation among the outcome variables in 0.3 increments (0.0 - 0.9).

Power was evaluated by selecting all conditions with a non-zero indirect effect. The rates of rejection were averaged over all conditions with the same level of correlation among outcome variables and sample size (See Table 5). Overall as correlation among outcome variables increases so too does power. This effect seems fairly uniform among all methods of analysis, and

the increase is reasonably linear. Though it would be reasonable to expect this trend to asymptote at 1 as the correlation approaches 1.

Moderation Effect. The moderation parameter did not affect type I error or power to detect the indirect effect (See Table 8).

Size of Indirect Effect. Though the power rates in Tables 4 and 5 may seem quite low, it is important to note that these power rates are averaged across the size of the indirect effect (ranging from 0.0196 to .3481). Table 6 shows that across indirect effects and sample sizes the power of most tests approaches 1.0 at large samples and/or large indirect effects. Because the data generated for the outcome variable were standard normal, the indirect effects generated can be interpreted as standardized mean differences between the outcome variables. These power estimates are very idealistic considering the data is perfectly normal and generated with equal variances for the outcome variables and mediators. With these precautions in mind, it seems that 50 participants would be sufficient to achieve the traditionally recommended 0.8 power for an effect of .35 using a bootstrap or Monte Carlo confidence interval. As can be seen in Table 6, much larger sample sizes are needed to detect very small effects.

From these simulation results it is clear that inferential methods for indirect effects adapted from the between-subject literature can be applied to within-subject designs. All tests performed in a relatively similar manner to previous simulation research with between-subject designs (For reference see Hayes & Scharkow, 2013 and MacKinnon et al., 2002). Additionally, based on these simulation results, I recommend researchers use a percentile bootstrap confidence interval, Monte Carlo confidence interval, or joint significance test for inferential decisions in within-subject mediation. Notably, the joint significance test does not provide a single *p*-value or a confidence interval, and so for that reason either of the interval estimates may be preferred. For

reasons previously explained, the Monte Carlo confidence intervals and the joint significance test may be sensitive to common violation of multiple regression, such as heteroskedasticity.

Between vs. Within-Subject Designs

Comparing across all conditions in the simulation, within-subject inferential methods have clear dominance over the between-subject inferential methods with respect to power. Table 7 shows that in more than 90% of all conditions, the within-subject method has higher power than the between-subject counterpart. Two questions emerge from this result: (1) Which cases have higher power for the between-subject design, and (2) How much more power is gained by using a within-subject design across different sample sizes and effect sizes?

In order to understand which cases have higher power for between-subject designs, these unique cases were examined. There were 86 conditions for which all within-subject methods had lower power than their between-subject counterparts. These cases came from all examined sample sizes and all indirect effects as well as direct effects; however, there were distinct patterns with the correlation parameters and moderation. These cases came primarily from conditions with high correlations among the mediators, low correlations among the outcomes, and 0 or 0.14 moderation parameters. These conditions align with the findings above regarding unique aspects of within-subject parameters, suggesting that power is lowest when these 3 conditions align, so much so that the benefits of a within-subject design are not present. However, it is unclear whether this combination of situations would regularly occur in substantive applications of this type of analysis.

Figure 8 shows power for each of the six inferential method in both within- and between-subject designs across sample size and indirect effects. It is clear that across sample sizes, inferential methods, and indirect effects, within-subject designs have higher power on average than between

subject designs for mediation. In general, there seems to be a larger benefit of within-subject designs for larger indirect effects and larger sample sizes, unless power gets very close to 1 where it asymptotes, and the methods get closer in power. The benefit for within-subject designs can be quite large, up to about 0.3 increase. Notably, in Figure 8 it is fairly common for the between subject line at a specific *N* to run parallel or near parallel to the within-subject line at *N/2*. For example, tests seem to have similar power with a between subject design at N = 200 and within-subject design at N = 100, or between subject design at N = 100 and within-subject design at N = 50. In general, there is a clear advantage of the within-subject design in power and this benefit seems to be approximately equal to a doubling of sample size within the range of sample sizes and effect sizes explored. However, in specific cases this benefit can be quite small (e.g., small effect size and small N, or large effect size and large N), so researchers should evaluate the appropriateness of within-subject versus between-subject designs for their specific case.

Discussion

The simulation study provides results on power and type I error for a variety of inferential tests on the indirect effect for both within- and between-subject designs. Many findings from previous simulation research examining mediation analysis in between-subject designs was replicated in this within-subject context. For example, the causal steps method and the normal theory tests were very conservative and underpowered. Additionally, the two standard error estimates for the normal theory tests did not differ greatly in their performance. The percentile bootstrap confidence interval, Monte Carlo confidence interval, and joint significance tests were also slightly conservative, but had good power. Based on the simulation results, the recommended inferential approach would a percentile bootstrap confidence interval or, if the original data is not available, the Monte Carlo confidence interval. The computational tool MEMORE (MEdiation

and MOderation for REpeated measures designs) for SPSS and SAS can be used to estimate these within-subject mediation models and provide inferential tests including percentile bootstraps and Monte Carlo confidence intervals (Montoya & Hayes, 2017). In most cases, a within-subject design is beneficial with regard to power; however, in cases when correlation among repeated-measurements (both the mediator and the outcome) is low, this benefit may be minimal.

Though the methods of inference were generally conservative in the simulation results, this is not a reason to ignore concerns about type I errors in the mediation literature. The conservative nature of the tests is less prominent in cases when one of the paths involved in the indirect effect is non-zero and the other is zero. This seems like a much more likely case for when mediation analysis might be implemented. In practice, mediators are often introduced into models when there is strong previous evidence of either the association between the manipulation and the mediator or the mediator and the outcome. Additionally, the data generated in this simulation come from perfect normal distributions, do not contain notable outliers, and were not subject to any researcher degrees of freedom, all factors which can increase type I error rates in real data application settings (Finch, West, & MacKinnon, 1997; Zu & Yuan, 2010; Simmons, Nelson, & Simonsohn, 2011).

Why does power drop at high correlation among mediators?

A particularly interesting finding in the simulation was that power drops when the mediators have a very high correlation with each other (See Figure 3). This is a characteristic unique to mediation analysis, as in other contexts higher correlation among outcomes often results in a monotonic increase in power. However, because mediators serve both as outcomes and as predictors, there is a trade-off which occurs with the increase in correlation. In particular, as the

correlation among repeated-measures increases, the variance in the error term in Equation 4 will decrease. In general, this improves the precision of the estimation of the *a*-path, as there is a reduction in the error in estimation. This reduction in error in estimation means greater power to detect the *a*-path. However, the variance of the mediator difference score will also decrease as the correlation among the mediators increases. The differences between the mediators are used as a predictor of differences between the outcomes. The standard error for a regression coefficient in a multiple regression model depends on many factors, one of which is the variance of the predictor. All else being equal, as the variance of the predictor variable decreases, the standard error of the regression coefficient will increase. In this particular case, increasing the correlation among the mediators decreases the variance of the difference between the mediators, and the standard error for the *b*-path increases. So, as the correlation among the mediators increases, the precision of the estimate of the *b*-path decreases, and thus the power to detect the *b*-path decreases. Altogether, increasing the correlation between the mediators increases the power to detect the *a*-path but decreases power to detect the *b*-path. Ultimately this results in a non-monotonic function with respect to power of the indirect effect, *ab*. In general, higher correlation among the mediators is beneficial to a point; however, researchers should be aware that if the correlation among the repeated measurements of the mediators is too high, statistical power may be limited.

Evaluating Appropriateness of Within-Subject Designs

The findings of this simulation research suggest that in many cases within-subject designs will provide higher statistical power than between-subject designs. However, it is important to acknowledge that statistical power is not the only factor that should weigh in on the decision of selecting an experimental design. Though there are well documented statistical advantages of

using within-subject designs, they may not be appropriate for all types of research questions. Greenwald (1976) describes specific considerations for deciding between a within- or between-subjects design, including context effects and external validity. In particular, carry over effects may impact the validity of within-subject designs. Though counterbalancing the order of stimulus presentation can eliminate any bias in the parameter estimates, carry over effects can still add to the variability in the data, causing lower statistical precision. Additionally, if carry over effects are present this may threaten the external validity of the findings. For example, if an effect is only present when participants are exposed to both stimuli, but in a real world setting it would be unrealistic for individuals to encounter both stimuli, then the findings of the study have little implication for real-world settings. Within-subject designs are only recommended if they are equally or more ecologically valid than a between-subject design. Within the context of mediation analysis, it is important to consider each of the effects involved in an indirect effect, and whether the effect estimated using the current design reflects the effect of interest by the researcher.

Another factor that should weigh in on a researcher's decision regarding design is cost: money, time, and other resources. There are cases when within-subject designs may realistically cost more than between-subject designs. In particular, if bringing the same participants back to the lab costs more than bringing many new participants into the lab, the optimal strategy may be to use a between-subject design with many more participants than a within-subject design. This may be particularly relevant if drop out is high, such that initial responses from individuals who do not return to the lab cannot be used.

Limitations & Future Directions

Though the current research provides new and useful information to researchers interested in using mediation analysis, there are a variety of limitations. Many of these limitations suggest avenues for future research which were outside of the scope of the current research.

One particularly notable limitation of this work is that the simulations assume that between- and within-subject effects are equal. In multilevel modeling when within and between effects differ these effects are called contextual effects (Enders & Tofighi, 2007; Raudenbush & Bryk, 2002; Kreft et al., 1995). In order to reasonably compare within and between-subject designs, the assumption that these two effect are equal is required; however, this does not necessarily mean that these effects will be equal in a given research context. The issue of within and between effects is also reflected in the discussion of ecological validity above, and researchers should reflect on whether they expect these effects to differ across levels (within and between) and which of these effects is of primary interest. Recent developments in multilevel mediation analysis allow for the estimation of both types of effects, with appropriate designs (Rockwood, 2018; Preacher; Zyphur, Zhang, 2010)

Throughout this paper, only experimental designs have been discussed, where individuals are randomly assigned to conditions or all individuals experience both conditions. Mediation analysis is used for other designs; however, as previously noted, the assumptions needed for causal inference are reduced by using an experimental design with random assignment. All of the statistical methods discussed in this paper do not rely on experimental design, but the validity of the causal claims is reduced when random assignment is not present. With respect to the claims about power, balanced conditions were always assumed, allowing there to be alignment across between- and within-subject designs. Situations in which the causal antecedent, X, is an observed variable are very common, and mediation methods are commonly used to analyze data from

studies like these. In within-subject studies, pre-post designs are frequently analyzed using the described mediation models. In the case of pre-post designs where all participants go through the same intervention but are measured before and after the intervention, the researchers must assume that all change observed is due to the intervention and there is no natural change over time. For cases where individuals are randomly assigned to intervention or control and observed pre-post, a moderated mediation model is most appropriate (See Hayes & Montoya, *under review*).

There are a variety of conditions which went uninvestigated in the current simulation studies. Each of these provides a potential avenue for expansion of understanding the performance of the proposed methods under different circumstances. Previous simulation research in between-subject design mediation analysis has examined the effects of non-normality (Biesanz, Falk, & Savalei, 2010; Finch, West, MacKinnon, 1997), missing data (Biesanz, Falk, & Savalei, 2010), and heteroskedasticity (Yuan & MacKinnon, 2014). Each of these issues may present particularly interesting quandaries in a within-subject designs as well. For example, non-normality can manifest in different ways, and the power and performance of different statistical methods may depend on the degree to which error distributions are the same across conditions. These conditions should also be examined for within-subject designs. Similarly, power in within-subject designs may be impacted by whether variances are equal across conditions and whether repeated-measures are positively or negatively correlated. All of these conditions could be investigated in future simulations. In particular, because the data generated in this simulation are largely coming from very ideal situations, future research should examine required sample sizes for adequate power under less than ideal circumstances, such as non-normality, heteroscedasticity, missing data, and unequal variances for the outcomes and mediators.

Summary

This study investigates whether there are power advantages of using a within-subject versus between-subject design when conducting mediation analysis. In most cases, within-subject designs have higher power than between-subject designs. Generally, for between-subject designs with a sample size of *N*, within-subject designs with a sample size of *N*/2 have comparable power. However, within-subject designs have lower power when correlation among outcomes is low and correlation among mediators is high. This study replicated many of the findings from between-subject mediation analysis, suggesting that percentile confidence intervals, Monte Carlo confidence intervals, and the joint significant test all perform similarly with reasonable type I error and power. Based on the results of this study, researchers should consider whether a within-subject design would be appropriate for their study investigating mechanisms, and if they can use a within-subject design to achieve greater statistical power while using the methods developed by Montoya & Hayes (2017) as well as the macro MEMORE for SPSS and SAS to conduct their analyses.

Acknowledgements: Special thanks to Dr. Andrew Hayes, Dr. Craig Enders, Jessica Fossum, and Tristan Tibbe for providing feedback on iterations of this manuscript. Parts of this manuscript were presented at the annual meetings for the Southeastern Psychological Association (2016) and the Society for Personality and Social Psychology (2019). This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant DGE-1343012.

Declaration of Interest Statement: Dr. Montoya reports grants from National Science Foundation, during the conduct of the study.

References

Aroian, L. A. (1947). The probability function of the product of two normally distributed variables. *Annals of Mathematical Statistics*, *18*, 265–271.

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.

Biesanz, J. C., Falk, C. F., & Savalei, V. (2010). Assessing mediational models: Testing and interval estimation for indirect effects, *Multivariate Behavioral Research*, *45*(5), 661-701.

Bradley, J. V. (1987). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144 – 152.

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

Collins, L. M., Graham, J. J. & Flaherty, B. P. (1998) An alternative framework for defining mediation, *Multivariate Behavioral Research*, *33*(2), 295 - 312, DOI: 10.1207/s15327906mbr3302_5

Damen, T. G. E. (2019). Sense of agency as a predictor of risk-taking. *Acta Psychologica*, *197*, 10 – 15, DOI: 10.1016/j.actpsy.2019.04.015

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*, Chapman & Hall/CRC.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121 – 138, DOI: 10.1037/1082-989X.12.2.121

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling*, *4*, 87-107.

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18(3), 233-239.

Fritz, M. S., Cox, M. G., & Mackinnon, D. P. (2015). Increasing statistical power in mediation models without increasing sample size. *Evaluation & Health Professions*, *38*(3), 343 – 366. DOI:10.1177/0163278713514250

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, *83*(2), 314 – 320.

Gunia, B. C., & Levine, E. E. (2019). Deception as competence: The effect of occupational stereotypes on the perception and proliferation of deception. *Organizational Behavior and Human Decision Processes*, *152*, 122 – 137. DOI: 10.1016/j.obhdp.2019.02.003

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis*. 2nd ed. Guilford Press.

Hayes, A. F., & Montoya, A. K. (*Under Review*). Mediation analysis in the two condition pretest-posttest design: A treatment-as-moderator conditional process approach.

Hayes, A. F., & Rockwood, N. J. (2017). Regression based statistical mediation and moderation analysis in clinical research: Observations, recommendations, and implementation. *Behaviour Research and Therapy*, *98*, 39–57. doi: 10.1016/j.brat.2016.11.001

Hayes, A. F., & Scharkow, M. (2013) The Relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological Science*, *24*(10), 1918-1927.

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.

Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001) Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, *6*(2), 115-134.

Kenny, D. A. & Judd, C. M. (2014). Power anomalies in testing mediation. *Psychological Science*, *25*(2), 334 – 339, DOI: 10.1177/0956797613502676.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In *Handbook of social psychology* (4 ed., Vol. 1).

Kreft, I. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30* (1), 1- 21.

Lawrence Erlbaum Associates.

Ledgerwood, A., & Shrout, P.E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology*. *101*,1174–1188.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Routledge.

MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, *17*(2), 144-158.

MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Personality and Social Psychology Review*, *19*(1), 30-43. DOI: 10.1177/1088868314542878

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, *1*, 173-181.

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*(1), 99-128.

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*, 83 – 104.

Maxwell, S. E. & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2 ed.), Mahwah, NJ: Lawrence Erlbaum Associates.

Montoya, A. K. & Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, *22*(1), 6 – 27. DOI: 1082-989X/17

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Newbury Park, CA: Sage.

Nguyen, T., Carnevale, J. J., Scholer, A. A., Miele, D. B., & Fujita, K. (2019, May 23). Metamotivational Knowledge of the Role of High-Level and Low-Level Construal in Goal-Relevant Task Performance. *Journal of Personality and Social Psychology*. Advance online publication. http://dx.doi.org/10.1037/pspa0000166

O'Rourke, H. P. & MacKinnon, D. P. (2018). Reasons for testing mediation in the absence of an intervention effect: A research imperative in prevention and intervention research. *Journal of Studies on Alcohol and Drugs*, *79*(2), 171–181, DOI:10.15288/jsad.2018.79.171

Pazda, A. D., & Thorstenson, C. A. (2019). Color intensity increases perceived extraversion and openness for zero-acquaintance judgements. *Personality and Individual Differences*, *147*, 118 – 127. DOI: 10.1016/j.paid.2019.04.022

Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology, 66*, 825 – 852. DOI: 10.1146/annurev-psych-010814-015258

Preacher, K. J., & Hayes, A. F. (2004) SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 717-731.

Preacher, K. J., & Selig, J. P. (2012) Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*(2), 77-98.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*(3), 209 – 233. DOI: 10.1037/a0020141

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Rockwood, N. J. (2017). Advancing the formulation and testing of multilevel mediation and moderated mediation models. (Unpublished master's thesis). The Ohio State University, Columbus, OH.

Rousselet, E., Brial, B., Cadario, R., Béji-Bécheur, A. (2018). Moral intensity, issue characteristics, and ethical issue recognition in sales situations. *Journal of Business Ethics*. DOI: 10.1007/s10551-018-4020-1.

Senn, S. (1993). *Cross-over trials in clinical research*. Ed 2. John Wiley & Sons, LTD.

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.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359 – 1366. DOI: 10.1177/0956797611417632

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological Methodology*, (pp. 290 – 293). Washington, DC: American Sociological Association.

Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance structure models. In N. Tuma (Ed.), *Sociological Methodology*, (pp. 159 – 86). Washington, DC: American Sociological Association.

Stone, C. A., & Sobel, M. E. (1990). The robustness of estimates of total indirect effects in covariance structure models estimated by maximum likelihood. *Psychometrika*, *55*(2), 337-352.

Thorstenson, C. A., Pazda, A. D., & Lichtenfeld, S. (2019). Facial blushing influences perceived embarrassment and related social functional evaluations. *Cognition and Emotion*.

Venter, A. & Maxwell, S. E. (1999). Maximizing power in randomized designs in N is small. In R. H. Hoyle (Ed.), *Statistical Strategies for Small Sample Research*. Thousand Oaks: Sage.

Yuan, Y., & MacKinnon, D. P. (2014). Robust mediation analysis based on median regression. *Psychological Methods*, *19*, 1-20.

Yzerbyt, V., 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

Zu, J., & Yuan, K.-H. (2010). Local influence and robust procedures for mediation analysis. *Multivariate Behavioral Research*, *45*(1), 1 – 44. DOI: 10.1080/00273170903504695

Table 1

Simulation Population Condition Values

Variable	Population values			
Sample Size	20	50	100	200
a path	0	.14	.39	.59
b path	0	.14	.39	.59
c' path	0	.14	.39	.59
d path	0	.14	.39	.59
ρ_m	0	.3	.6	.9
ρ_y	0	.3	.6	.9

Table 2
Type I Error When a = 0, across Sample Size and Values of b

	Method					
b	CS	JS	Sobel	Aroian	Bootstrap	MC
			N = 20			
0	*0.0007*	*0.0024*	*0.0004*	*0.0003*	*0.0038*	*0.0029*
0.14	*0.0035*	*0.0080*	*0.0022*	*0.0017*	*0.0102*	*0.0098*
0.39	*0.0077*	*0.0205*	*0.0087*	*0.0074*	0.0273	0.0258
0.59	*0.0131*	0.0320	*0.0208*	*0.0188*	**0.0440**	**0.0412**
			N = 50			
0	*0.0012*	*0.0027*	*0.0002*	*0.0002*	*0.0024*	*0.0020*
0.14	*0.0086*	*0.0157*	*0.0043*	*0.0035*	*0.0163*	*0.0150*
0.39	*0.0186*	**0.0369**	*0.0180*	*0.0156*	**0.0406**	**0.0388**
0.59	*0.0213*	**0.0427**	0.0300	0.0279	**0.0504**	**0.0478**
			N = 100			
0	*0.0015*	*0.0025*	*0.0001*	*0.0001*	*0.0018*	*0.0015*
0.14	*0.0140*	*0.0228*	*0.0080*	*0.0068*	*0.0223*	*0.0215*
0.39	0.0255	**0.0439**	0.0269	0.0246	**0.0475**	**0.0464**
0.59	0.0272	**0.0449**	**0.0356**	0.0340	**0.0493**	**0.0490**
			N = 200			
0	*0.0020*	*0.0028*	*0.0001*	*0.0001*	*0.0019*	*0.0017*
0.14	*0.0215*	0.0325	*0.0134*	*0.0117*	0.0309	0.0307
0.39	0.0306	**0.0477**	**0.0360**	0.0338	**0.0502**	**0.0497**
0.59	0.0329	**0.0487**	**0.0417**	**0.0404**	**0.0514**	**0.0510**

Note. Type I errors in *italics* are outside Bradley's liberal criteria (0.025 – 0.075), and type I errors in **bold** are within Serlin's robustness criteria (0.035 – 0.065). In each row the test with the type I error rate closest to .05 is underlined.

Table 3
Type I Error When b = 0, across Sample Size and Values of a

	Method					
a	CS	JS	Sobel	Aroian	Bootstrap	MC
			N = 20			
0	*0.0007*	*0.0024*	*0.0004*	*0.0003*	*0.0038*	*0.0029*
0.14	*0.0016*	*0.0048*	*0.0011*	*0.0008*	*0.0073*	*0.0057*
0.39	*0.0060*	*0.0178*	*0.0075*	*0.0063*	0.0252	*0.0226*
0.59	*0.0091*	0.0295	*0.0152*	*0.0132*	**0.0418**	**0.0374**
			N = 50			
0	*0.0012*	*0.0027*	*0.0002*	*0.0002*	*0.0024*	*0.0020*
0.14	*0.0046*	*0.0094*	*0.0012*	*0.0009*	*0.0095*	*0.0082*
0.39	*0.0156*	0.0314	*0.0116*	*0.0098*	**0.0359**	0.0313
0.59	*0.0220*	**0.0458**	*0.0238*	*0.0210*	**0.0558**	***0.0497***
			N = 100			
0	*0.0015*	*0.0025*	*0.0001*	*0.0001*	*0.0018*	*0.0015*
0.14	*0.0087*	*0.0146*	*0.0028*	*0.0021*	*0.0138*	*0.0123*
0.39	0.0252	**0.0435**	*0.0189*	*0.0166*	**0.0479**	**0.0436**
0.59	0.0291	**0.0501**	0.0341	0.0311	**0.0592**	**0.0545**
			N = 200			
0	*0.0020*	*0.0028*	*0.0001*	*0.0001*	*0.0019*	*0.0017*
0.14	*0.0146*	*0.0225*	*0.0056*	*0.0045*	*0.0204*	*0.0190*
0.39	0.0309	**0.0483**	0.0291	0.0263	**0.0533**	**0.0503**
0.59	0.0331	**0.0508**	**0.0422**	**0.0399**	**0.0573**	**0.0543**

Note. Type I errors in *italics* are outside Bradley's liberal criteria (0.025 – 0.075), and type I errors in **bold** are within Serlin's robustness criteria (0.035 – 0.065). In each row the test with the type I error rate closest to .05 is underlined.

Table 4
Power Across Sample Size and Correlation among Mediators

	Method					
ρ_M	CS	JS	Sobel	Aroian	Bootstrap	MC
			$N = 20$			
0.0	0.0754	0.1175	0.0792	0.0722	**0.1375**	*0.1358*
0.3	0.0807	0.1271	0.0808	0.0727	**0.1447**	*0.1445*
0.6	0.1000	0.1519	0.1040	0.0953	*0.1701*	**0.1710**
0.9	0.0946	0.1370	0.1178	0.1116	*0.1567*	**0.1592**
			$N = 50$			
0.0	0.0754	0.1175	0.0792	0.0722	**0.1375**	*0.1358*
0.3	0.0807	0.1271	0.0808	0.0727	**0.1447**	*0.1445*
0.6	0.1000	0.1519	0.1040	0.0953	*0.1701*	**0.1710**
0.9	0.0946	0.1370	0.1178	0.1116	*0.1567*	**0.1592**
			$N = 100$			
0.0	0.4148	0.5213	0.4677	0.4581	**0.5263**	*0.5239*
0.3	0.4490	0.5525	0.5014	0.4918	**0.5570**	*0.5548*
0.6	0.4655	0.5584	0.5125	0.5035	**0.5638**	*0.5621*
0.9	0.4148	0.5213	0.4677	0.4581	**0.5263**	*0.5239*
			$N = 200$			
0.0	0.5696	0.6748	0.6386	0.6316	**0.6764**	*0.6757*
0.3	0.5954	0.6902	0.6568	0.6506	**0.6925**	*0.6913*
0.6	0.6192	0.7069	0.6713	0.6652	**0.7087**	*0.7076*
0.9	0.5696	0.6748	0.6386	0.6316	**0.6764**	*0.6757*

Note. In each row the test with highest power is highlighted in bold. On each line the test with the second highest power is highlighted in italics.

Table 5
Power Across Sample Size and Correlation among Outcome Variables

	Method					
ρ_Y	CS	JS	Sobel	Aroian	Bootstrap	MC
			$N = 20$			
0.0	0.0423	0.0888	0.0490	0.0426	**0.1037**	*0.1032*
0.3	0.0618	0.1079	0.0672	0.0599	**0.1254**	*0.1250*
0.6	0.1100	0.1542	0.1203	0.1126	**0.1788**	*0.1770*
0.9	0.1772	0.2170	0.1711	0.1615	*0.2364*	**0.2411**
			$N = 50$			
0.0	0.1852	0.2830	0.2113	0.1990	**0.2928**	*0.2890*
0.3	0.2243	0.3111	0.2466	0.2349	**0.3216**	*0.3179*
0.6	0.2999	0.3738	0.3241	0.3151	**0.3870**	*0.3825*
0.9	0.4324	0.5007	0.4232	0.4086	**0.5104**	*0.5087*
			$N = 100$			
0.0	0.2948	0.3943	0.3465	0.3379	**0.3998**	*0.3970*
0.3	0.3852	0.4846	0.4318	0.4222	**0.4902**	*0.4872*
0.6	0.4736	0.5593	0.5115	0.5032	**0.5642**	*0.5623*
0.9	0.6146	0.6937	0.6565	0.6479	**0.6993**	*0.6978*
			$N = 200$			
0.0	0.4933	0.6013	0.5621	0.5551	**0.6040**	*0.6025*
0.3	0.5465	0.6452	0.6068	0.5993	**0.6471**	*0.6466*
0.6	0.6372	0.7214	0.6920	0.6864	**0.7235**	*0.7226*
0.9	0.7405	0.8127	0.7977	0.7941	**0.8157**	*0.8145*

Note. In each row the test with highest power is highlighted in bold. On each line the rest with the second highest power is highlighted in italics.

Table 6
Power Across Sample Size and Size of Indirect Effect

	Method					
ab	CS	JS	Sobel	Aroian	Bootstrap	MC
			N = 20			
0.0196	0.0062	0.0139	0.0043	0.0033	**0.0179**	*0.0168*
0.0546	0.0225	0.0441	0.0203	0.0173	**0.0547**	*0.0526*
0.0826	0.0432	0.0793	0.0457	0.0404	**0.0965**	*0.0942*
0.1521	0.0875	0.1438	0.0883	0.0790	*0.1662*	**0.1666**
0.2301	0.1631	0.2428	0.1773	0.1635	*0.2739*	**0.2760**
0.3481	0.2973	0.3963	0.3337	0.3145	*0.4359*	**0.4413**
			N = 50			
0.0196	0.0270	*0.0434*	0.0143	0.0121	**0.0443**	0.0413
0.0546	0.1026	0.1583	0.0865	0.0776	**0.1650**	*0.1603*
0.0826	0.1621	0.2395	0.1705	0.1592	**0.2521**	*0.2474*
0.1521	0.3319	0.4665	0.3564	0.3353	**0.4825**	*0.4782*
0.2301	0.4927	0.6310	0.5641	0.5467	**0.6482**	*0.6463*
0.3481	0.6994	0.8044	0.7833	0.7751	**0.8184**	*0.8161*
			N = 100			
0.0196	0.0701	**0.1047**	0.0463	0.0408	*0.1023*	0.0998
0.0546	0.2338	0.3332	0.2397	0.2249	**0.3381**	*0.3342*
0.0826	0.3019	0.4073	0.3563	0.3447	**0.4181**	*0.4145*
0.1521	0.6015	0.7684	0.7186	0.7057	**0.7764**	*0.7748*
0.2301	0.7468	0.8617	0.8522	0.8485	**0.8671**	*0.8665*
0.3481	0.8938	0.9383	0.9379	0.9374	*0.9407*	**0.9407**
			N = 200			
0.0196	0.1746	**0.2467**	0.1434	0.1308	*0.2401*	0.2377
0.0546	0.4116	0.5504	0.4919	0.4790	**0.5555**	*0.5543*
0.0826	0.4725	0.5881	0.5692	0.5636	**0.5948**	*0.5936*
0.1521	0.8045	0.9327	0.9289	0.9277	**0.9343**	*0.9342*
0.2301	0.9051	0.9564	0.9560	0.9558	**0.9579**	*0.9573*
0.3481	0.9742	0.9805	0.9806	0.9805	*0.9809*	**0.9810**

Note. In each row the test with highest power is highlighted in bold. On each line the test with the second highest power is highlighted in italics.

Table 7.
Proportion of Cases for which Power is Greater in Each Design.

	Method					
Design	CS	JS	Sobel1	Sobel2	Boot	MC
Within	0.94	0.91	0.92	0.92	0.91	0.92
Equal	0.02	0.00	0.01	0.02	0.00	0.00
Between	0.04	0.08	0.07	0.07	0.09	0.08

Table 8.
Rejection Rate (Power and Type I Error) across Moderation parameter, N = 100

	Method					
d	CS	JS	Sobel	Aroian	Bootstrap	MC
			ab = .1521			
0	0.5966	0.7454	0.6985	0.6861	0.7536	0.7516
0.14	0.6275	0.8204	0.7553	0.7399	0.8263	0.8261
0.39	0.5943	0.7707	0.7353	0.7261	0.7803	0.7783
0.59	0.5370	0.7715	0.7290	0.7145	0.7853	0.7813
			ab = .00			
0	0.0185	0.0308	0.0176	0.0161	0.0338	0.0318
0.14	0.0185	0.0309	0.0172	0.0156	0.0339	0.0319
0.39	0.0165	0.0299	0.0144	0.0128	0.0320	0.0300
0.59	0.0155	0.0284	0.0132	0.0116	0.0296	0.0278

Figure 1.

Path diagram of simple mediation model two-condition between-subjects design.

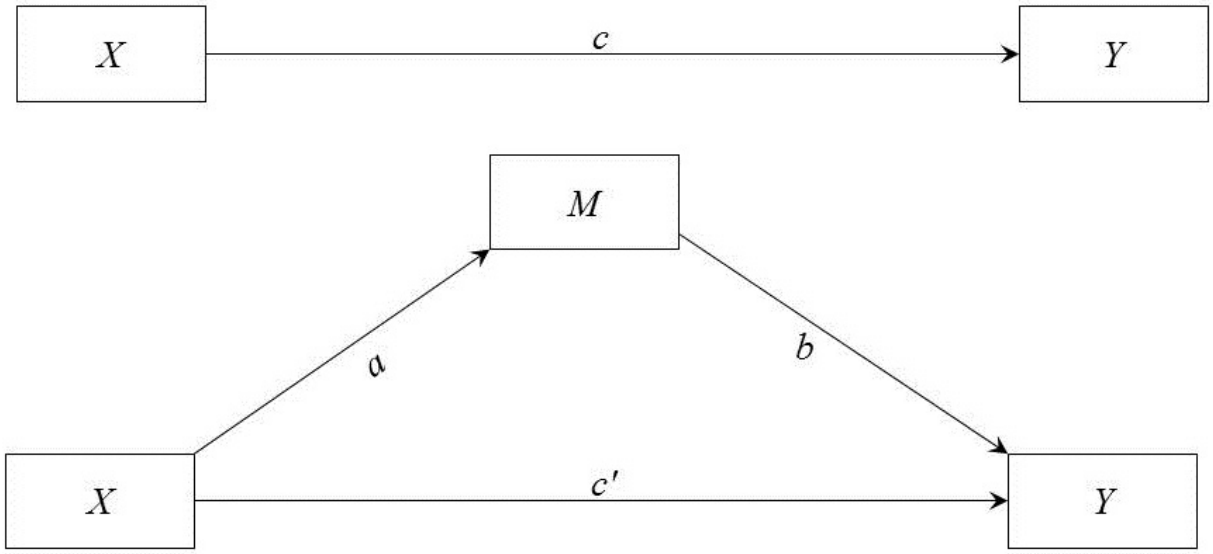
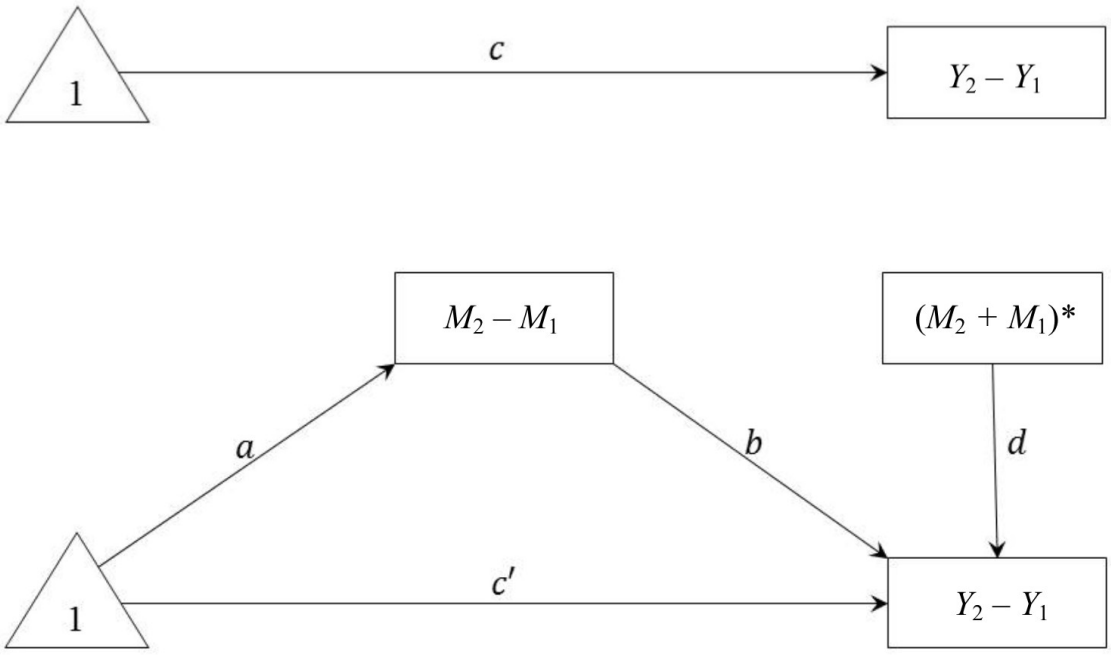


Figure 2. *Path diagram of simple mediation model for two-condition within-subject design.*



Note: A 1 contained in a triangle indicates an intercept. The * indicates grand-mean centered.

Figure 3. Power of inferential methods across levels of correlation among mediators, $N = 100$.

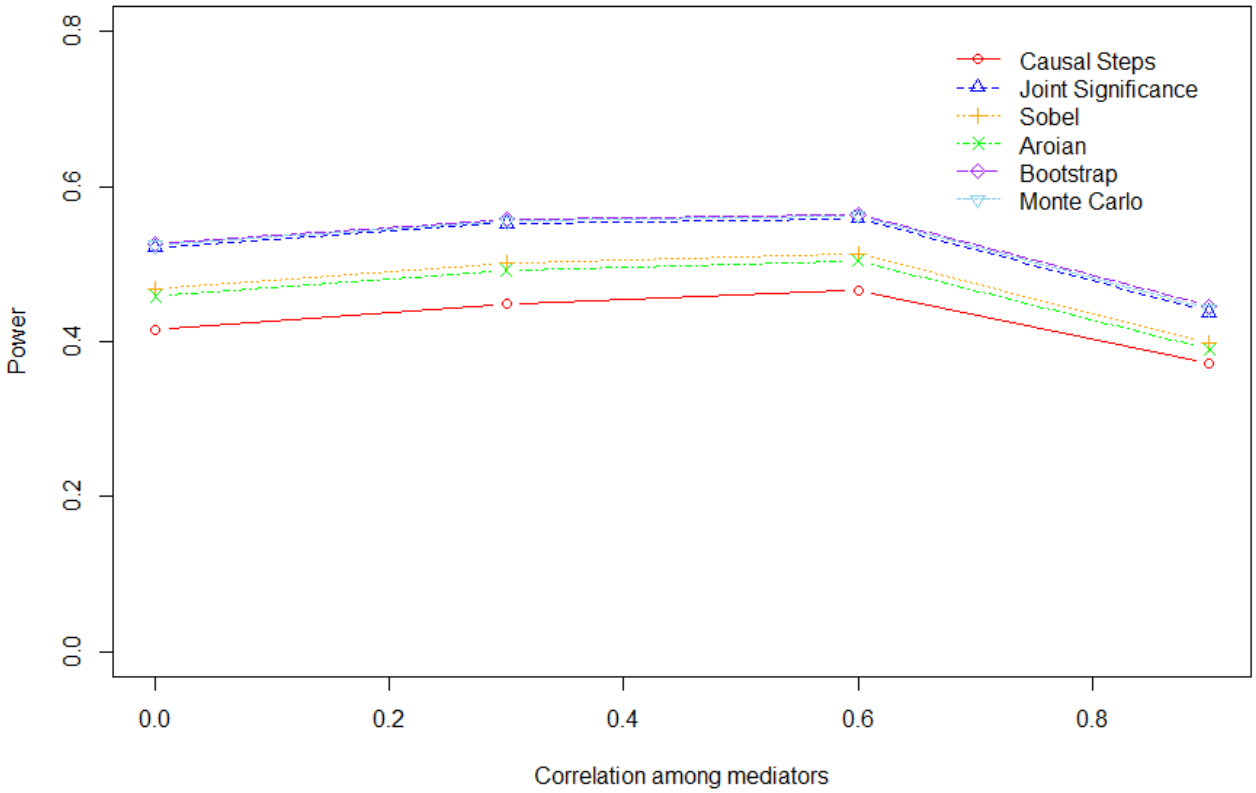


Figure 4. Power of inferential methods across correlation of the outcomes, $N = 100$.

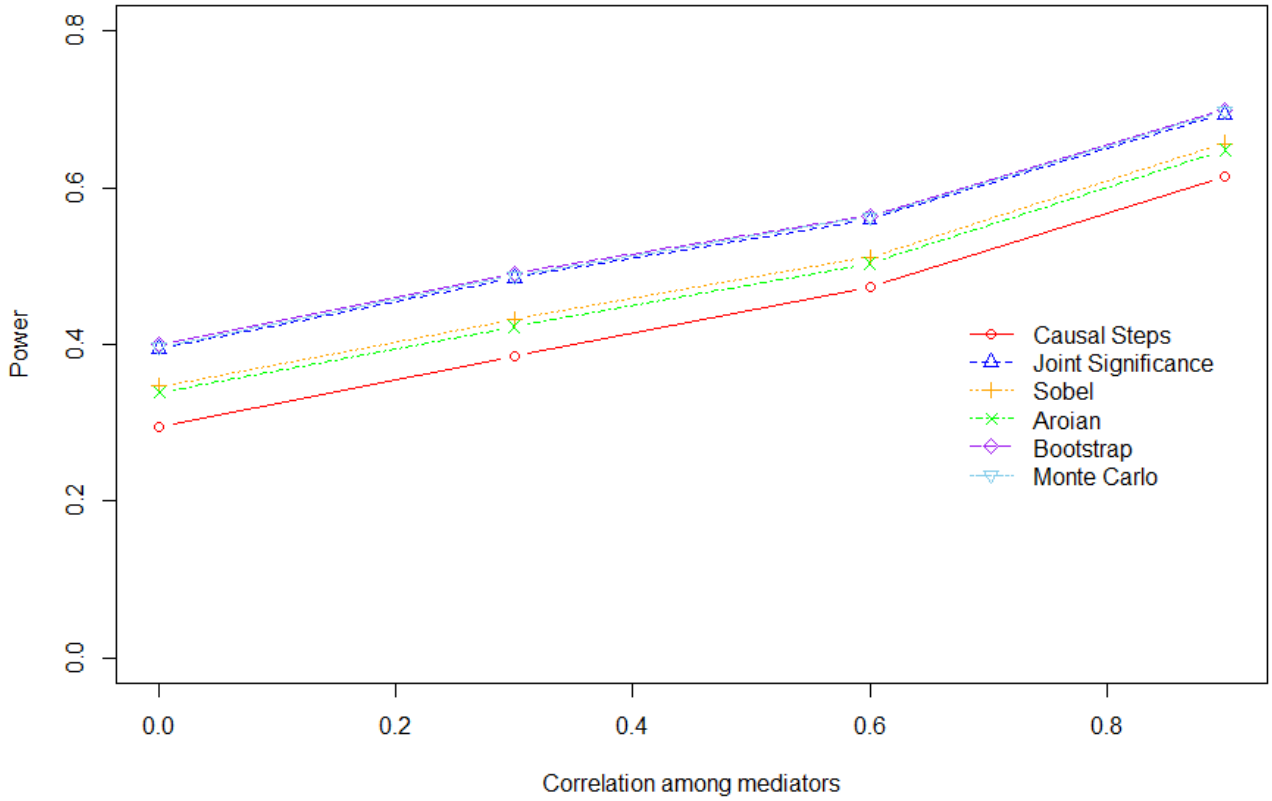


Figure 7. Power of within-subject design across sample size.

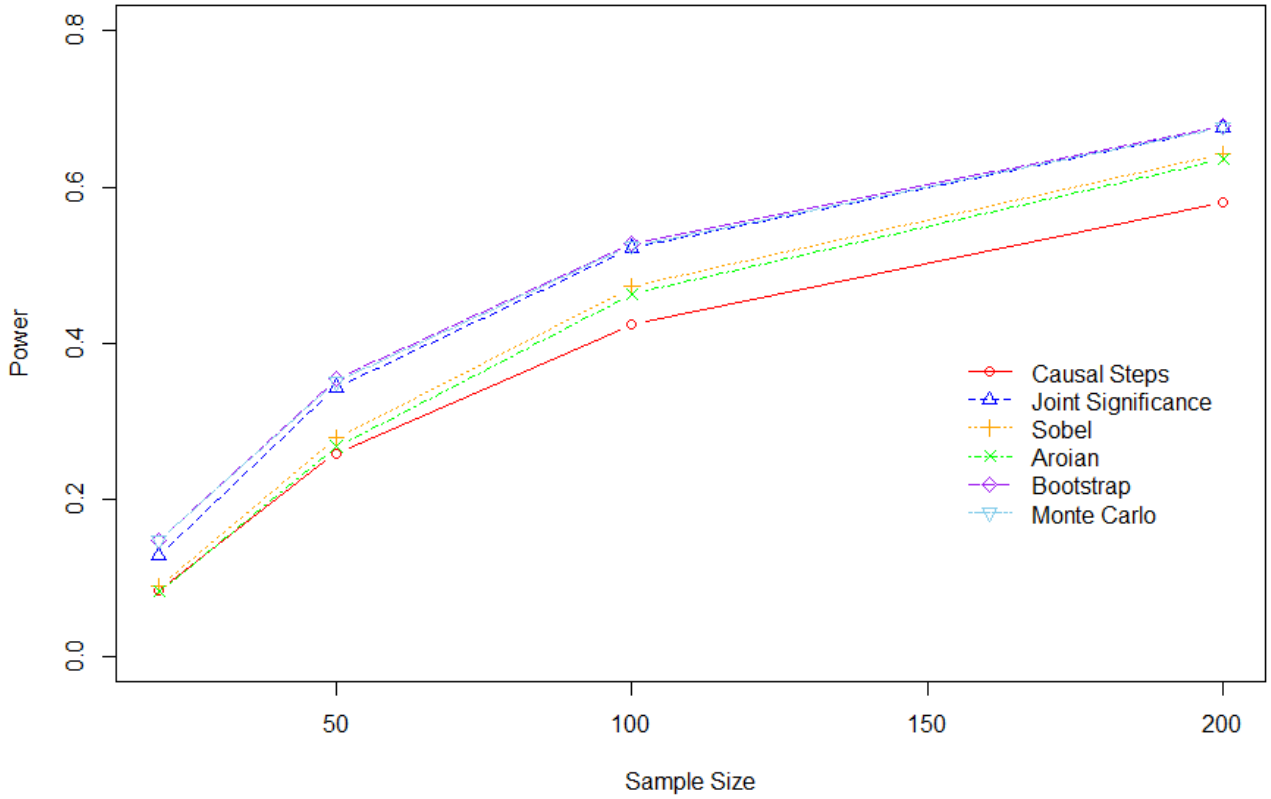


Figure 8. Power across Design, Method, Sample Size, and Indirect Effect