

# Practical Problems Estimating and Reporting Power When Hypotheses Are Embedded in Complex Statistical Models

David A. Cole<sup>1</sup>, George Abitante<sup>1</sup>, Hoi Kan<sup>1</sup>, Qimin Liu<sup>1</sup>,  
 Kristopher J. Preacher<sup>1</sup>, and Scott E. Maxwell<sup>2</sup>

<sup>1</sup>Department of Psychology and Human Development, Vanderbilt University, Nashville, Tennessee, and

<sup>2</sup>Department of Psychology, University of Notre Dame, South Bend, Indiana

Advances in Methods and  
 Practices in Psychological Science  
 January-March 2025, Vol. 8, No. 1,  
 pp. 1–17  
 © The Author(s) 2025  
 Article reuse guidelines:  
 sagepub.com/journals-permissions  
 DOI: 10.1177/25152459241302300  
 www.psychologicalscience.org/AMPPS



## Abstract

In both grant proposals and published studies, many research hypotheses (represented by primary parameters) are tested in the context of complex statistical models. The power to detect these primary parameters depends on the values of multiple secondary model parameters that often go unexamined, unreported, and unjustified. The result is that many a priori power analyses are incomplete, ambiguous, potentially subjective, and nearly impossible for others to evaluate. In the current article, we encourage researchers to use plausible values for secondary parameters (PVSPs), in addition to the hoped-for effect sizes for their primary parameters, in their power calculations. More specifically, upper and lower bounds for power are generated based on the highest and lowest plausible values for secondary model parameters. In the article, we demonstrate how to conduct and describe such power estimates for four increasingly complex statistical methods. We further demonstrate how such analyses can inform decisions about resource allocation in ways that improve power in the context of complex statistical models, often revealing that power can be enhanced by methods other than increasing sample size. The PVSP approach can enhance power, improve the transparency of power analyses in grant proposals, reduce the likelihood of funding underpowered research, and alleviate at least one problem underlying the replication crisis.

## Keywords

power estimation, complex models, Monte Carlo simulations, grant proposals, replication crisis

Received 4/11/23; Revision accepted 5/7/24

Since Cohen's (1977) classic book, power analyses have become a common step in most research designs and a standard part of most grant proposals in psychology. However, since 1977, the complexity of statistical methods used in social-science research has grown dramatically. Cohen's methods (and many recently developed online apps<sup>1</sup>) are often insufficient to estimate power when complex analyses are involved. Instead, Monte Carlo approaches to power estimation have become increasingly popular (e.g., Muthén & Muthén, 2002; Wang & Rhemtulla, 2021). A PsycInfo search revealed a 50-fold increase in nonmethodological journal articles that contain the combination of "Monte Carlo" and "power" over the past 4 decades. Typically, researchers assign plausible values to all the parameters that

comprise their model. These include the primary model parameter(s), representing the hypothesized effect(s) of interest, and the secondary parameters, representing all the rest of the model. (For heuristic purposes, in the current article, we assume that the research hypothesis is represented by a single model parameter; however, the proposed methods can be generalized to multiparameter hypotheses.) Methodological changes that affect the secondary parameters can dramatically change estimates of one's power to detect the primary parameter

## Corresponding Author:

David A. Cole, Department of Psychology and Human Development,  
 Vanderbilt University, Nashville, Tennessee

Email: david.cole@vanderbilt.edu



of interest—often without increasing sample size. At present, researchers lack a systematic, explicit, and efficient Monte Carlo method that takes these highly influential secondary parameters into account when estimating power to detect the primary parameter.

Building on the noteworthy recommendations of Muthén and Muthén (2002) and Wang and Rhemtulla (2021), our goals were to describe and demonstrate what we call a “plausible values for secondary parameters” (PVSP) approach to power estimation for testing hypotheses that are embedded in complex statistical models. This approach involves the identification of a range of plausible values for secondary model parameters and the empirical power estimation for different combinations of these values to obtain a range of plausible power estimates (not just a single estimate). This approach can be applied to any statistical model.

With the emergence of increasingly sophisticated statistical methods, key hypotheses are often represented by a relatively few primary parameters in models that contain a relatively large number of secondary parameters. In grant proposals, the power-analysis section typically contains substantial justification for the expected size of the primary parameter(s). Prior research is often invoked to support the researcher’s expectation of a small, medium, or large primary effect. Although the magnitude of the primary effect will always affect power, it may be one of the more difficult effect sizes to justify with prior research. After all, the purpose of the proposed research is probably to test something novel—something not already supported by prior research. One could argue that it is better to justify primary parameters on the basis of what is interesting, important, and worth detecting (e.g., Lakens, 2014; Lakens et al., 2018).

What can often be justified with prior research, however, are the magnitudes of the secondary parameters in the model. The power to detect the primary effect can vary enormously as a function of other model parameters. The key, then, is to identify what range of values is plausible for the secondary model parameters. Frequently, these ranges can be justified by prior research. One example might be a longitudinal study testing the effect of  $X$  at Time 1 on  $Y$  at Time 2, where important secondary parameters are the correlations between  $X$  and other covariates. Plausible high and low values for such correlations could be based on prior cross-sectional research. A second example might be a latent-variable mediation model, in which important secondary model parameters include the manifest variable factor loadings onto their respective factors. Plausible high and low values for such loadings could be based on prior psychometric research. In these examples, the identification of justified or at least plausible secondary characteristics is a critical first step. In basing high and low PVSPs on

prior research, however, one runs some risk that those values may be biased because of either upward or downward publication pressures (Anderson & Maxwell, 2017; Gelman & Carlin, 2014). To safeguard against such possibilities, one can use PVSP brackets that go somewhat beyond previously published estimates.<sup>2</sup>

In the current article, we propose a PVSP method for conducting and efficiently reporting power analyses that pertain to relatively complex models. This approach can be implemented whenever one estimates a priori power for reasonably sophisticated data-analytic models. For complex statistical models (and simpler ones as well), we propose that researchers should examine power to detect their primary model parameters under a variety of conditions—conditions represented by high and low PVSPs. Such a practice would generate an upper limit for power that is based on the optimal pattern of PVSP estimates and a lower limit based on the least optimal pattern. The term “optimal PVSPs” refers to the pattern of secondary parameters that imparts the greatest power to the primary test while still being realistic (e.g., supported by previous research). Conversely, the idea of “least optimal PVSPs” refers to the pattern of (still plausible) secondary parameters that imparts the lowest power to test the primary hypothesis while still being justified by prior studies. In some models, the magnitudes of secondary model parameters may not be monotonically related to the power to test the primary model parameter. In such cases, researchers may want to estimate power at more than just high and low values of the secondary parameters.

The PVSP approach has three advantages. First, it makes explicit the context for the power analysis. This is especially important in publications or grant proposals, in which reporting the conditions under which power is calculated will better enable reviewers to evaluate the assumptions on which power, sample size, funding decisions, and publication were made. (These conditions consist of both the primary parameter values and the PVSPs.) Second, reporting a range of power estimates (based on the optimal and suboptimal patterns of PVSPs) is more honest because researchers rarely know the exact values of their secondary parameters in advance. And third, systematic examination of PVSPs can inform the strategic allocation of resources to improve power (e.g., if factor loadings have a particularly strong effect on power, the researcher may want to consider methods for improving measurement).

The PVSP approach differs from Cohen’s (1988) classic approach in at least two ways. First, Cohen’s is an analytic approach to power (e.g., Cohen, 1977, 1992; Murphy et al., 2014; O’Brien & Muller, 1993), which becomes impractical when models are complex. Ours is an empirical approach (although PVSPs can be

embedded in an analytic approach as well). Second, Cohen's approach regards best-guess estimates of effect sizes as though they were the population values, whereas the PVSP method takes parameter uncertainty into account by considering a range of plausible values. Third, Cohen's approach focuses on the magnitude of the primary effect, whereas the PVSP approach considers the myriad ways that secondary parameters affect power. Other approaches also incorporate the observed uncertainty around effect sizes into account (e.g., McShane & Böckenholt, 2015; Perugini et al., 2014; Sakaluk, 2016). Most similar to our approach are the Bayesian-hybrid approaches that consider a set of effect-size estimates (e.g., from meta-analytic studies). Such approaches (e.g., Du et al., 2017; Du & Wang, 2016; Liu & Wang, 2019; Pek & Park, 2019) formally consider both assurance level (i.e., the likelihood for a researcher to obtain power equal to or greater than desired power under planned sample size) and power variability (i.e., how power estimates vary given research design). However, these approaches focus on how uncertainties in research design relate to a complete distribution of power values. Our approach emphasizes the careful consideration of a few high and low yet plausible secondary model parameters. Our approach encourages researchers to consider how anticipated ranges for various PVSPs can influence power. In practice, some researchers will be so familiar with certain PVSPs that the range of their plausible values may be extremely narrow, almost constant. In other cases, the range of PVSPs could be so large as to range from positive to negative. In the current article, we assume that high and low values for each parameter have moderate range and retain the same sign.

Our approach can be summarized in a kind of to-do list. The following items should be considered in the application of this method to any model; however, simpler models may end up requiring fewer steps:

1. Model: Construct the model you will eventually use when analyzing your data to test the hypotheses of interest.
2. Primary paths: Identify all primary paths (i.e., those reflecting your hypotheses) and assign them "smallest effect size of interest" (i.e., values that if significant, would be just large enough to be meaningful; Anderson, 2019; Lakens, 2014).
3. Measurement: Identify features that represent the reliability or validity of your measures and select their lowest and highest plausible values, ideally based on prior research with similar samples.
4. Collinearity: Identify paths that represent or contribute to correlations among substantive exogenous variables (reflecting collinearity among your predictors) and select their lowest and highest plausible values, ideally based on prior research.
5. Nuisance covariance: Identify correlations among residuals (e.g., those reflecting shared method variance) and select their lowest and highest plausible values.
6. Other paths: Identify other structural paths (if any) that do not represent key hypotheses and identify their lowest and highest plausible values.
7. Develop factorial design with high and low sets of values for each of these clusters.
8. Run Monte Carlo simulations to estimate power under each combination of the features represented in your design. If the model were to contain all of these features, this could result in 16 simulations: 2 (Measurement)  $\times$  2 (Collinearity)  $\times$  2 (Nuisance Covariance)  $\times$  2 (Other Paths). Note that the number of analyses is often fewer, as when all features are not contained in your design; however, the number could also be larger if you elect to subdivide these features or examine more than two conditions per feature.

These steps can be implemented with any of several recently developed tools that are relatively easy to use. One tool is embedded in Mplus. Muthén and Muthén (2002) described how Mplus can be used to conduct Monte Carlo studies to estimate power to test specific structural-equation-modeling (SEM) parameters based on one's model, plausible values for model parameters, and sample size (and certain variable and sample characteristics). The procedure merely requires researchers to add a few commands to the same Mplus program they use to analyze their data. A second tool is pwrSEM, a Shiny app created and described by Wang and Rhemtulla (2021). This app generates R code for Monte Carlo power estimates based on model syntax and plausible parameter values. The app is freely accessible at <https://yilinandre.wang.shinyapps.io/pwrSEM/> and supported by an online manual. These approaches assume that researchers can generate plausible values for all parameters in their models. We have found that most applied researchers can easily speculate about such model parameters. In some cases, however, researchers may prefer to base their power analyses on the underlying correlations among their variables. Indeed, obtaining estimates of such correlations from previous studies may be easier than finding estimates of specific model parameters. In such cases, apps such as WebPower may be especially useful (Zhang et al., 2018; Zhang & Yuan, 2018). In still other cases, especially those involving simpler models, one may be able to calculate power mathematically; however, such analytic approaches often become intractable when

models become more complex, providing opportunities for applications of Monte Carlo methods (Muthén & Muthén, 2002).

In the current article, we apply our PVSP approach to four statistical methods, demonstrating the use of Mplus, pwrSEM, or R for heuristic purposes (for examples of each available, see the supplemental materials at <https://osf.io/vymsx/>). The first example involves multiple regression. We start with this model because it enables us to demonstrate the procedure in a relatively simple case. (In addition, it demonstrates that power depends on secondary parameters in simple and complex models.) The second is a SEM approach to mediation. Because mediation processes and hypotheses are embedded in many SEM studies, the points made in this example will pertain to more complex models. Third, we consider a random intercept cross-lagged panel analytic model (RI-CLPM). Conventional cross-lagged panel analytic models (CLPMs) have been a methodological mainstay of longitudinal research for decades (Rogosa, 1980); however, Hamaker et al.'s (2015) relatively new RI-CLPM promises to replace them in many applications. Fourth, we apply our method to an especially complex model with a large number of secondary parameters to demonstrate how such parameters can be clustered or combined to make the power analyses more feasible. We use these four statistical models merely as examples. In no way do we mean to imply that they exhaust the domain of modern complex statistical analyses. Our intent is that readers can use these examples to inform their application of the PVSP method to other statistical models.

For each of these models, we demonstrate how to generate high and low PVSPs, ascertain their effects on power, and then report the upper and lower plausible bounds for power given a particular sample size, nominal Type I error rate ( $\alpha$ ), and primary effect size. We are mindful that space is precious in journal articles and grant proposals, likely leaving insufficient room for justifying multiple PVSPs and reporting multiple power estimates. Consequently, for each of our examples, we demonstrate how PVSP-based power estimates can be efficiently reported while retaining sufficient information to allow for critical review.

## Multiple Regression

We begin with a fundamental linear multiple regression model, where  $X$  at Time 1 predicts outcome  $Y$ , after controlling for prior levels of the outcome measured at Time 1: That is,  $Y_2 = \beta_1 X_1 + \beta_2 Y_1 + e$ , where  $X_1$ ,  $Y_1$ , and  $Y_2$  are standardized (see Fig. 1). For example, one might be interested in the effect of self-esteem at Time 1 ( $X_1$ ) on depression at Time 2 ( $Y_2$ ) while controlling for depression at Time 1 ( $Y_1$ ). In this example, the parameter

of primary interest would be  $\beta_1$ . More specifically, the researchers may want sufficient power to reject  $H_0: \beta_1 = 0$  when the population value of  $\beta_1$  is .30 (i.e., an a priori value large enough to be of theoretical or practical importance; see Lakens et al., 2018). In this model, the “secondary” parameters might be the correlation ( $\rho$ ) between the two predictors and the auto-regression of depression over time ( $\beta_2$ ). Even though these values do not reflect the primary question of interest, they are nevertheless critical to power estimation. As collinearity among the predictors increases, so do the standard errors around the regression weight estimates (e.g.,  $\beta_1$ ), potentially reducing power to detect these effects (Gujarati & Porter, 2009).<sup>3</sup> Conversely, a high autoregressive coefficient for the covariate (i.e.,  $\beta_2$ ) can increase the power to detect  $\beta_1$  (all other things being equal) because the covariate explains more variance in the dependent variable and reduces the denominator of the statistical test (Maxwell et al., 2017). Of course, all other things are not always equal given that higher autoregression for the criterion variable is often associated with smaller effects for the predictor of interest.

For a Monte Carlo approach to power to be ecologically valid, these secondary parameters must be plausible; that is, they should be justified by prior research. The correlation between predictors is one such secondary parameter. Even though the primary analyses are longitudinal, previously conducted cross-sectional studies could document the range of correlations that have been found between self-esteem and depression range (e.g., Nguyen et al., 2019; Sick et al., 2020). The autoregression of the depression variable is another secondary parameter. Previous longitudinal studies of depression (especially those with similar lag times) could be used to document the range of autoregression estimates that have been reported (e.g., Ju & Lee, 2018; Reed-Fitzke et al., 2021; Takagishi et al., 2011).

## Simulations

To demonstrate the effect of these secondary characteristics on the power to test the primary parameter, we began by conducting a Monte Carlo simulation with the model parameters set to what prior methodological research (Gujarati & Porter, 2009; Maxwell et al., 2017) would suggest would be the most optimal of the potential PVSP values (i.e., those plausible secondary parameter values that maximize the power to detect  $\beta_1$ )—in this case, low collinearity ( $\rho = .10$ ) and high autoregression ( $\beta_2 = 0.50$ ). In addition, we set  $\beta_1 = 0.30$ ,  $\sigma_{X_1}^2 = \sigma_{Y_1}^2 = \sigma_{Y_2}^2 = 1.0$ ,  $\alpha = .05$ , and  $N = 200$ . Using maximum-likelihood (ML) estimation in R, we ran 10,000 replications (similar results emerge for least-squares estimation). For each replication, we drew 200 random deviates from a multivariate normal population with the above

**Table 1.** Monte Carlo Power Estimates for Detecting  $\beta_1$  in Multiple-Regression Model for Low and High PVSPs, Where  $\beta_1 = 0.30, \alpha = .05, N = 200$

PVSPs		
Stability $\beta_2$	Collinearity $\rho$	Power to detect $\beta_1$
High: 0.50	Low: .10	.999
Low: 0.10	Low: .10	.991
High: 0.50	High: .80	.973
Low: 0.10	High: .80	.784

Note: Power estimates will vary if the seed or the number of replications is changed. PVSPs = plausible values for secondary parameters.

characteristics. The result of primary interest is the power to reject  $H_0: \beta_1 = 0 | \beta_1 = 0.30$ . Results appear in the first row of Table 1, showing excellent power at .999. (R code for this analysis is contained in supplemental materials. Similar results emerge from using either Mplus or pwrSEM, which are also included in supplementary materials.)

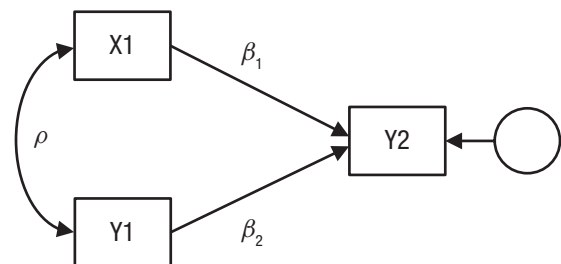
In a second Monte Carlo simulation, we set all PVSPs to values we anticipated would generate the worst power (i.e., their suboptimal values)—in this case, high collinearity ( $\rho = .80$ ) and low autoregression ( $\beta_2 = 0.10$ ). After changing these values, we also adjusted the error variance so that the total model-implied variance of  $Y$  remained 1.0 and  $\beta_1$  remained the same in the population. Results from this analysis appear in the last row of Table 1. These PVSP values yielded a much lower power estimate of .784 even though the population value for  $\beta_1$  (the parameter of interest) remained exactly the same. In other words, power calculations might reveal exceptionally good power (.999) or barely adequate power (.784) depending on one’s selection of values for model parameters that are not of primary interest, are rarely justified by prior research, and are often not even reported in most power analyses. Underreporting secondary model parameters in power analyses opens the door for the “strategic” selection of secondary model parameter values that could paint an overly optimistic picture about power.

In more complex models, anticipating whether changing a particular secondary parameter will increase or decrease power is not always straightforward. Therefore, we recommend estimating differences in power that result from changing several secondary characteristic values at a time, ideally in a full factorial design. (When the number of PVSPs is large, a full factorial design may be impractical. Reduced factorial designs [Collins et al., 2009] may provide a viable option in these situations.)

We conducted these analyses examining different combinations of optimal and suboptimal PVSPs. In Table 1, Rows 2 and 3 contain power estimates that emerged when only one secondary characteristic (either autoregression or collinearity) was set to its worst plausible value. These changes had a relatively small impact on power. Setting both characteristics to their suboptimal values, however, revealed that the combination of low autoregression and high collinearity (Table 1, Row 4) had a more serious effect on power. Unfortunately, changing the research design in ways that reduce collinearity may be impossible. Fortunately, these results also suggest solutions. For example, one might change the research design to increase the autoregression coefficient (perhaps by improving the reliability of the criterion variable). More specifically, comparing Rows 3 and 4 in Table 1 reveals that increasing autoregression from 0.1 to 0.5 increases power from .784 to .973. In our experience, researchers often behave as though increasing their sample size is the only way to improve power. However, the PVSP approach reveals that power can be enhanced by changing other aspects of the research design as well. For example, Fritz et al. (2015) demonstrated how improved measurement strategies in one part of a structural equation model can enhance power in other parts.

If power analyses are being conducted for inclusion in a grant proposal or journal article, not all of the above information needs to be reported. For the current case, an efficient and sufficient power paragraph might simply be

To estimate power for the test of our primary hypothesis (that  $\beta_1 \neq 0$ ), we used a PVSP Monte Carlo approach. For  $\beta_2$ , the lowest and highest plausible values were .10 and .50; for  $\rho$ , the lowest and highest plausible values were .10 and .80 (e.g., Ju & Lee, 2018; Saint-Georges & Vaillancourt, 2020). Assuming a true effect of  $\beta_1 = 0.30, \alpha = .05$ , and  $N = 200$ , power ranges from .784 (for suboptimal yet plausible secondary characteristics) to .999 (for optimal plausible secondary characteristics).



**Fig. 1.** Multiple-regression model.

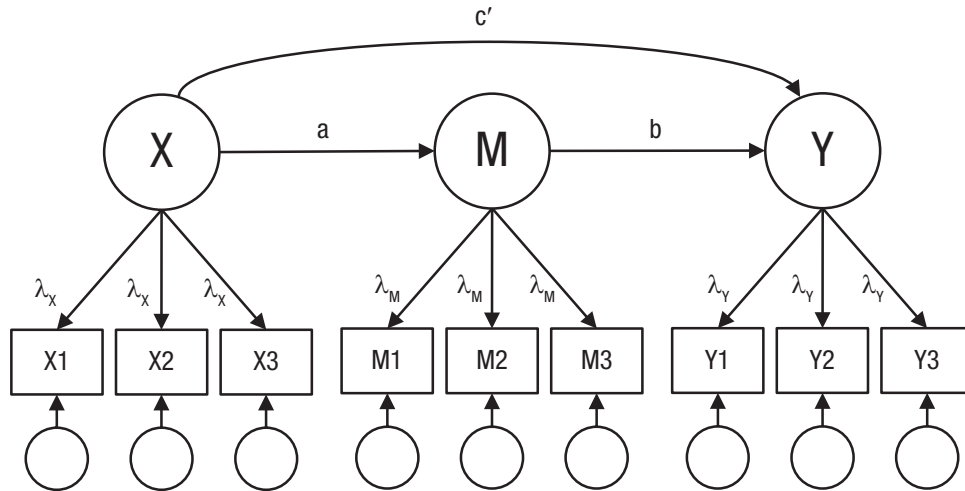


Fig. 2. Latent-variable mediation model.

### An Aside

In a multiple-regression analysis such as this, one might question the need for a Monte Carlo approach to power estimation when analytic approaches have long been available (Cohen, 1977). We submit at least two good reasons. First, Monte Carlo simulations allow researchers to take sample characteristics into account, such as attrition, violation of distributional assumptions, differences in measurement reliability, and so on. These features can affect not only power but also Type I error rates, the accuracy of parameter estimates, and the integrity of standard errors, all of which can be examined via Monte Carlo methods.

Second, in many applications of multiple regression, the researcher may be testing a variety of hypotheses or effects (e.g.,  $\beta_1, \beta_2, R^2, \Delta R^2$ ). Power computational formulas for these effects may exist but become cumbersome as the number of predictors increases (Maxwell, 2000). Alternatively, Monte Carlo methods can be easily expanded to estimate power to detect multiple effects either individually or collectively.

### Latent-Variable Mediation Model

Our second example is a latent-variable mediation model using SEM. Not only are explicit tests of mediation extremely common, but mediation processes are also implicit in many structural equation models. Here, we focus on the relatively simple mediation SEM depicted in Figure 2, where latent  $X$  (represented by indicators  $X_1$ ,  $X_2$ , and  $X_3$ ) predicts latent  $M$  (represented by  $M_1$ ,  $M_2$ , and  $M_3$ ), which then predicts latent  $Y$  (represented by  $Y_1$ ,  $Y_2$ , and  $Y_3$ ). For example, the first factor might

be Classroom Milieu, and its measures are three classroom-observation sessions. The second factor might be Self-Worth, and its measures are three separate questionnaires. And the third factor could be Performance, measured via three tasks. The product of paths  $a$  and  $b$  represents the indirect effect of  $X$  on  $Y$  through  $M$ . Path  $c'$  represents the direct effect of  $X$  on  $Y$  (i.e., that part of the  $X \rightarrow Y$  relation that is not mediated by  $M$ ).

Again, we consider two types of study characteristics. The first type consists of the primary parameters of interest: paths  $a$  and  $b$  and the  $ab$  product. All represent key hypotheses. Next are the secondary parameters. In this example, we regarded the factor loadings onto  $X$ ,  $M$ , and  $Y$  as secondary parameters. Previous research has shown that (all other things being equal) smaller factor loadings are associated with larger standard errors around the structural paths of primary interest and thus lower power (Ledgerwood & Shrout, 2011). For the purposes of the current article, we also regarded path  $c'$  as a secondary parameter. The existence of path  $c'$  makes  $X$  a covariate in the regression of  $Y$  onto  $M$ . If all other path coefficients are held constant, increasing  $c'$  means that  $M$  will explain a larger proportion of the unexplained variance in  $Y$ , thus increasing the power to detect path  $b$  (Maxwell et al., 2017). Other characteristics could also be examined, including various amounts of missing data, deviations from multivariate normality, variances and covariances of the residuals, and so on.

To ensure that our secondary characteristic values were plausible, we conducted a small literature review of mediation studies in which parent-adolescent relation predicted self-esteem, which predicted depression (Hu & Ai, 2016). Of particular interest were the factor loadings ( $\lambda$ ). Based on this review, we set the lowest and highest PVSPs

for  $\lambda$ s to be .40 and .70. In general, lower factor loadings occurred when the researchers used qualitatively different methods to represent the latent variable (e.g., parent, teacher, and child reports of child problem internalization; Muhtadie et al., 2013), and higher loadings emerged when researchers used similar methods (e.g., multiple self-report questionnaires; see Orcutt, 2006). For the direct effect, we elected to set the low value of path  $c'$  to 0 (sometimes conceptualized as the ideal outcome of a mediation study) and the high value to be .39, a medium-sized effect according to MacKinnon et al. (2002). In all simulations, we set  $\alpha = .05$  and  $N = 350$ .

**Simulations**

We conducted Monte Carlo simulations to demonstrate the effect of our PVSPs ( $\lambda$ s and path  $c'$ ) on the power to test the primary parameters (path  $a$ , path  $b$ , and the  $ab$  indirect effect). In our first analysis, we set population parameters equal to the optimal PVSP values:  $\lambda_X = \lambda_M = \lambda_Y = .7$  and path  $c' = .39$  (additionally, path  $a = \text{path } b = .39, ab = .15, \sigma_X^2 = \sigma_M^2 = \sigma_Y^2 = 1.0$ ). Direct-effect values are based on values commonly used to represent small, medium, and large effects ( $a = b = .14, .39, .59$ ) in Monte Carlo research on mediation (MacKinnon et al., 2002). Best and worst plausible values for factor loadings were set to .30 and .90, based on values commonly reported for latent-variable mediation models (e.g., Clarkson-Smith & Hartley, 1990; Gustafsson et al., 2012; Lucas-Thompson & Hostinar, 2013; Mega et al., 2014; Von Korff & Grotevant, 2011; Wei et al., 2005). When loadings changed, we adjusted the measurement residual variances to maintain variances of the manifest variables at 1.0. When the direct effect changed, we adjusted the structural residual variances to maintain the latent-variable variances at 1.0 as well.

We ran 10,000 replications, drawing multivariate normal random deviates from a population with the above characteristics. We used ML estimation in the *lavaan* package of R (Version 4.1.0; R Core Team, 2021; Rosseel, 2012) to implement the Monte Carlo confidence method to test mediation (MacKinnon et al., 2004; Preacher & Selig, 2012). R code for this analysis appears in the Supplemental Material available online.<sup>4</sup> Results appear in the first row of Table 2. Power estimates were excellent, with  $>.99$  power to detect path  $a$ , path  $b$ , and the  $ab$  indirect effect.

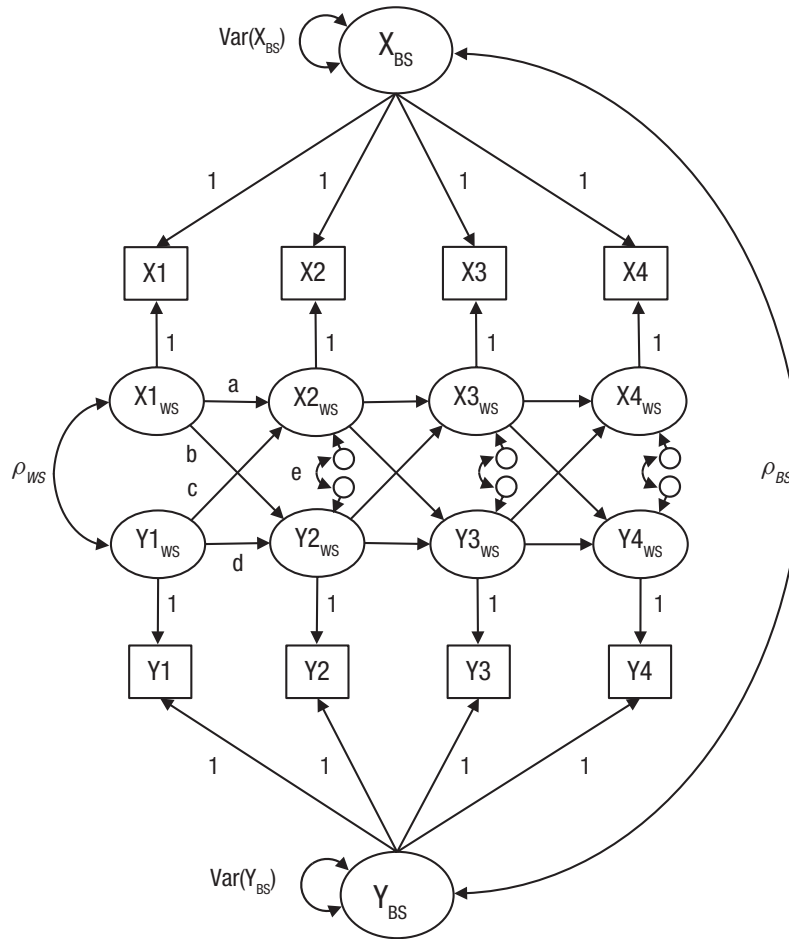
Next, we conducted the simulation using suboptimal PVSP values:  $\lambda_X = \lambda_M = \lambda_Y = .4$  and path  $c = 0$ . Under these circumstances, power dropped precipitously to .42 for path  $a$ , to .25 for path  $b$ , and to .06 for the  $ab$  indirect effect. See Row 4 of Table 2.

To understand better how changes in each PVSP affected power, we examined them in different combinations. In Table 2 (Rows 2 and 3), we see that changing the direct effect  $c'$  from .39 to 0 had essentially no effect on power to detect the  $ab$  indirect effect. However, reducing the factor loadings from .70 to .40 had a devastating effect, driving  $ab$  power down to .10. The impact of measurement fallibility in manifest-variable path-analytic mediation models has long been recognized (Bollen, 1989; Cole & Preacher, 2014; James et al., 1982; Kenny, 1979; Rigdon, 1994; Rubio & Gillespie, 1995). Ledgerwood and Shrout (2011) further described the adverse effect of low measurement reliability on the power to detect structural paths in latent variable models. Nevertheless, we were surprised that our highly plausible range of factor loadings would reveal such large differences in power for this (extremely common) mediation model. Even when using latent variables to test mediation, reliability of the manifest variables can profoundly affect power.

**Table 2.** Monte Carlo Power Estimates for Detecting Mediation in Structural-Equation-Modeling Approach to Mediation for Low and High PVSPs, Where  $a = .39, b = .39, ab = .15, \alpha = .05,$  and  $N = 350$

PVSPs		Power to detect		
Factor loadings	Direct effect (path $c'$ )	Path $a$	Path $b$	Indirect effect: $ab$
High: .70	High: .39	1.00	1.00	1.00
High: .70	Low: .00	1.00	1.00	1.00
Low: .40	High: .39	.48	.33	.10
Low: .40	Low: .00	.42	.25	.06
<b>After improving measurement</b>				
.60	.00	.99	.91	.90

Note: PVSPs = plausible values for secondary parameters.



**Fig. 3.** Random intercept cross-lagged panel model.

Thus, the PVSP approach again reveals ways that power can be enhanced by changing things other than sample size. In this case, analyses suggest that the researcher could concentrate on improving measurement reliability, perhaps by extending the duration of the assessment time, lengthening the questionnaire, or increasing the number of trials. If one could plausibly increase the lowest plausible factor loadings from .40 to .60 (as in the last row of Table 2), power would improve substantially, justifying a power-analysis paragraph like this:

To estimate the power of the tests of our primary hypotheses ( $a \neq 0$ ;  $b \neq 0$ ;  $ab \neq 0$ ), we used a PVSP Monte Carlo approach. Prior research has shown that factor loadings range from .4 to .7 (Muhtadie et al., 2013; Orcutt, 2006), and plausible values for path  $c'$  range from 0 to .39 for (MacKinnon et al., 2002). Under these circumstances and assuming the true effects were  $a = .39$ ,  $b = .39$ , and  $ab = .15$ , with  $\alpha = .05$  and  $N = 350$ , power for the test of the indirect effect ranges from .42 to  $\approx 1.00$  for path  $a$ , from .25 to  $\approx 1.00$  for path  $b$ , and from .06 to  $\approx 1.00$

for the  $ab$  indirect effect. However, after upgrading our measures, the lowest plausible factor loading would be .60, in which case power estimates for the three primary effects range from .90 to .99. (See bottom row of Table 2.)

### RI-CLPM

SEM approaches to cross-lagged panel analyses have been a mainstay of longitudinal research when studying the reciprocal prospective relations between two or more constructs. Recently, Hamaker et al. (2015) introduced an RI-CLPM that goes a step further, disentangling between-subjects and within-subjects aspects of the target relations, qualitatively changing how researchers approach CLPM data (Orth et al., 2021). For our demonstration, we used the RI-CLPM depicted in Figure 3. The model begins with longitudinal panel data on two variables measured at four waves: X1 to X4 and Y1 to Y4. Latent trait-like (or time-invariant) between-subjects variables  $X_{BS}$  and  $Y_{BS}$  are extracted from each series. Their correlation ( $\rho_{BS}$ ) represents that part of the relation

between the Xs and Ys that is entirely between subjects. The residual portions of these two time series (i.e., the  $X_{ws}$  series and the  $Y_{ws}$  series) are time-varying and relate to each other via a cross-lagged structure, estimating their prospective and reciprocal effects while controlling for  $X_{BS}$  and  $Y_{BS}$ . Paths  $a$  and  $d$  represent the autoregressive stabilities of  $X_{ws}$  variables and the  $Y_{ws}$  variables, respectively, and paths  $b$  and  $c$  represent the cross-lagged effects.

For the current application, two effects were of primary interest: the within-persons and the between-persons relation of X to Y (i.e., path  $b$  and  $\rho_{BS}$ , respectively). In complex analyses, the optimal PVSPs for one hypothesis may not be optimal for the other, as we show below. In the current example, we set path  $b = .30$  and  $\rho_{BS} = .50$ . (For heuristic purposes, we arbitrarily set the other cross-lag path  $c$  to 0; in some situations, researchers may want to consider high and low PVSPs for this parameter as well.) Secondary parameters were the Wave 1 correlation ( $\rho_{ws}$ ) and stability (paths  $a$  and  $d$ ). Because this was a longitudinal design, a key data characteristic was missingness because of attrition from the study over time. Based on prior studies (Grunewald et al., 2022; Kojima et al., 2021; Wu et al., 2018), we set low and high values for the Wave 1 correlation at  $\rho_{ws} = .30$  and  $.60$  and low and high stabilities at  $a = d = .30$  and  $.50$ . To demonstrate how Monte Carlo methods can also estimate the effect of attrition on power, we also set missingness at 0% and 10% attrition per wave (e.g., Fong-Jia et al., 2020; In-Albon et al., 2017) under the assumption that the missing data were missing completely at random.

### Simulations

Because we were uncertain how our secondary characteristics would affect power when testing more than one hypothesis, our first set of Monte Carlo simulations

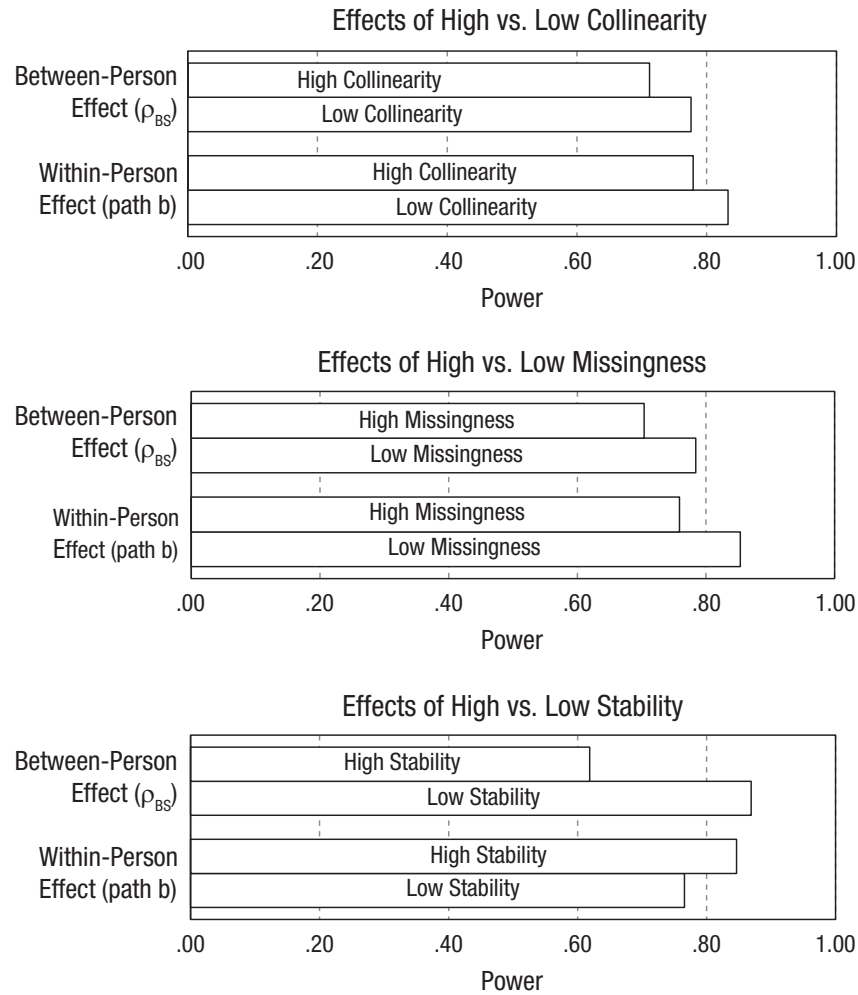
estimated power under all eight different combinations of high and low stability, Wave 1 correlation, and missingness (in a factorial design). In each simulation, path  $b = .30$ ,  $\rho_{BS} = .50$ , path  $c = 0$ ,  $var(X_{BS}) = var(Y_{BS}) = var(X_{ws}) = var(Y_{ws}) = 1.0$ ,  $\alpha = .05$ , and  $N = 200$ . We drew 10,000 samples from a multivariate normal population with the above characteristics. ML results are shown in Table 3. (For R code and Mplus code for this analysis, see the Supplemental Material. We did not include pwrSEM code for this example because it cannot take attrition into account.)

These changes in secondary characteristics had noteworthy effects, with power to detect path  $b$  varying from .63 to .90 and power to detect  $\rho_{BS}$  ranging from .58 to .93. Closer examination of Table 3 revealed three things. First, increasing the Wave 1 correlation reduced power slightly from .83 to .78 for the within-persons effect (path  $b$ ) and from .78 to .71 for the between-persons relation ( $\rho_{BS}$ ). Second, increasing missingness had a somewhat larger impact, reducing power from .85 to .76 for the within-persons effect and from .78 to .70 for the between-persons relation. Third, increasing stability paradoxically improved power to detect the within-persons effect from .77 to .85 but diminished power to detect the between-persons relation from .87 to .62. Figure 4 highlights each of these three effects, averaging across the other two conditions. The loss of power to detect  $\rho_{BS}$  happens in RI-CLPMs, trait-state models, and other approaches that combine autoregressive and latent-curve models (e.g., Curran & Bollen, 2001) because increases in the stability of the time-varying within-persons factors make it more difficult to distinguish them from the time-invariant between-persons factors. Even conceptually, to distinguish a completely stable “trait” from a highly stable “state” is difficult. In the current example, reducing the amount of missing data would improve the power to detect both  $\rho_{BS}$  and path  $b$ . At least three methods of

**Table 3.** Monte Carlo Power Estimates for Detecting Within-Subjects Effect (Path  $b$ ) and Between-Subjects Relation ( $\rho_{BS}$ ) in a Structural-Equation-Modeling Approach to RI-CLPM for Low and High PVSPs, Where Path  $b = .30$ , Path  $c = 0$ ,  $\rho_{BS} = .50$ ,  $\alpha = .05$ , and  $N = 200$

Values for PVSPs				
Wave 1 correlation $\rho_{ws}$	% Missingness/wave	Stability (path $a = d$ )	Average power for path $b$	Power for $\rho_{BS}$
Low: .30	Low: 0%	Low: .30	.86	<b>.94</b>
High: .60	Low: 0%	Low: .30	.72	.89
Low: .30	Low: 0%	High: .50	<b>.90</b>	.74
High: .60	Low: 0%	High: .50	.83	.63
Low: .30	High: 10%	Low: .30	.78	.87
High: .60	High: 10%	Low: .30	<b>.63</b>	.82
Low: .30	High: 10%	High: .50	.82	.66
High: .60	High: 10%	High: .50	.72	<b>.58</b>

Note: Highest and lowest power estimates are bold. PVSPs = plausible values for secondary parameters; RI-CLPM = random intercept cross-lagged panel model.



**Fig. 4.** How changes in secondary model parameters can differentially affect power to detect within-persons effects (e.g., path b) versus between-persons relations (e.g.,  $\rho_{BS}$ ) in random-intercept-cross-lagged-panel-analytic-model analyses. Each effect is computed while averaging over the other two effects.

accomplishing this are possible. One might be to commit more resources to participant retention. A second might be to use a sampling-replenishment strategy (e.g., Taylor et al., 2020). A third strategy (and the one implemented here) is simply to increase the sample size. This strategy revealed a need for 450 participants to achieve at least .80 power for each hypothesis under all PVSP conditions (results of these analyses are not shown).

Consequently, in a grant proposal, the power-analysis paragraph might look like this:

To estimate power for the tests of our primary hypotheses that  $b \neq 0$  and  $\rho_{BS} \neq 0$ , we used a PVSP Monte Carlo approach. The lowest and highest plausible values were .3 versus .6 for the Wave 1 correlation, .3 versus .5 for stability (e.g., Fong-Jia et al., 2020; In-Albon et al., 2017), and 0% versus 10% missingness per wave (Isaksson et al., 2020; Wouters et al., 2013). Assuming that the true effects for the

primary hypotheses were ( $b = .3$  and  $\rho_{BS} = .50$ ) and assuming  $\alpha = .05$  and  $N = 450$ , power estimates range from .95 to .99 for testing path  $b$  and from .80 to .99 for testing  $\rho_{BS}$ .

Note that using a Monte Carlo approach, one could also estimate the  $N$  needed to detect not just either but both hypotheses. We did not take this approach here, but strong arguments have been made that studies should be sufficiently powered for all hypotheses, not each (Maxwell, 2000, 2004). Monte Carlo methods for estimating power to test combined effects are enormously easier than analytic methods.

### Even More Complex Models

Needless to say, examining the effects of every combination of PVSPs in a full factorial design quickly becomes impractical as models increase in complexity and the

number of PVSPs grows. For example, the structural equation model in Figure 5 (where  $\beta_{YX}$  is the parameter of primary interest) contains 27 PVSPs and would require  $2^{27} = 134,217,728$  simulations to examine all possible combinations—completely impractical. Instead, we advocate the examination of broad types of PVSPs.

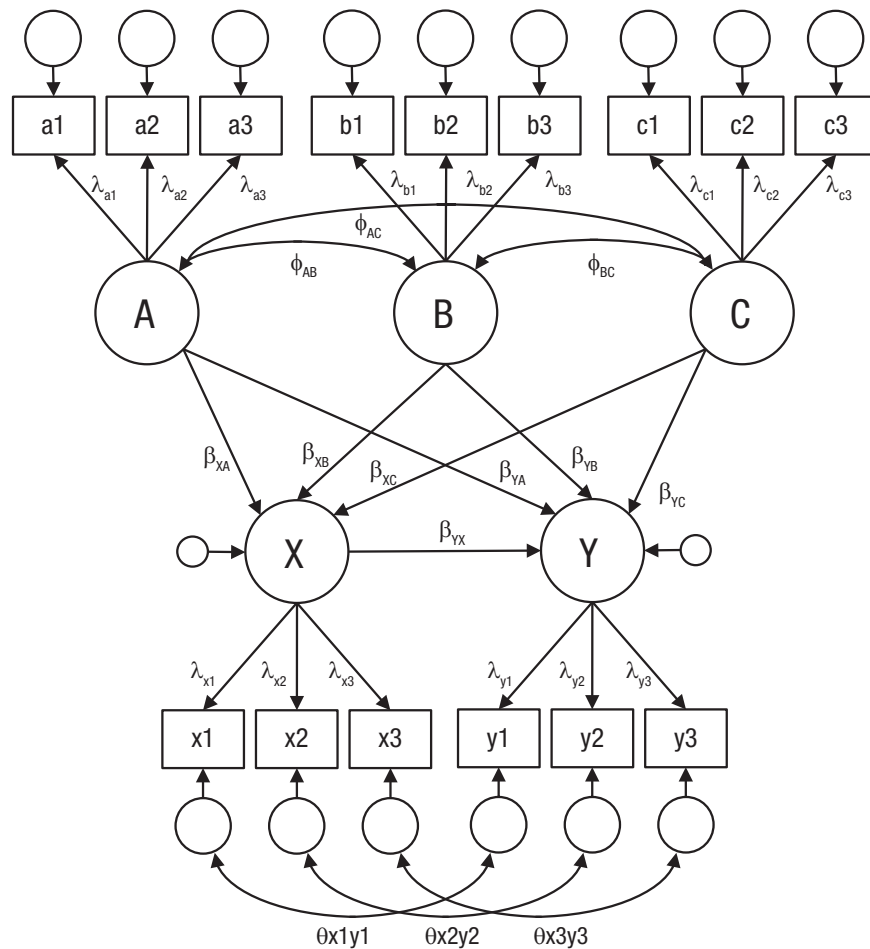
In the current example, one could combine or cluster these parameters into three broad types. The first pertains to measurement and consists of the 15 factor loadings, considered en masse. The second reflects the strength of the covariates, which can be conceptualized as the change in  $R^2$  for X (and for Y) that is due to A, B, and C. The third contains the three disturbance-term covariances. Examining large and small values for these three sets of PVSPs would require a  $2 \times 2 \times 2$  design and only eight simulations.

For this example, Table 4 shows power estimates to reject  $H_0: \beta_{YX} = 0$  for all combinations of low and high values for all three types of PVSPs, assuming the actual  $\beta_{YX} = 0.30$ ,  $\alpha = .05$ , and  $N = 300$ . Clearly, the measurement parameters had the greatest impact. When factor loadings were small (.35, .40, and .45 for the three measures of

each latent variable), power estimates ranged from .04 to .33. In contrast, when factor loadings were large (.65, .70, and .75 for each latent variable), power estimates ranged from .89 to .96. This pattern suggests that improvements in measurement could dramatically improve power without necessarily increasing the sample size. Alternative parameter combinations and clusters (e.g., dismantling the  $R^2$ s into their component path coefficients) yielded the same fundamental conclusion, demonstrating that informed and judicious grouping of PVSPs can reduce the number of conditions, making feasible a factorial approach to power in complex models.

### Discussion

Our primary goal in the current article was to address some of the practical problems associated with estimating and reporting a priori power when one's research hypotheses are embedded in complex statistical models. Based on simulations involving four increasingly complex statistical models, four key recommendations emerge.



**Fig. 5.** Complex causal model for latent  $X \rightarrow$  latent  $Y$  with three latent covariates (A, B, C) and correlated residuals among measures of X and Y.

**Table 4.** Monte Carlo Power Estimates for Detecting  $\beta_{YX}$  in the Structural Equation Model Depicted in Figure 5 When  $\beta_{YX} = 0.30$ ,  $\alpha = .05$ , and  $N = 300$

Values for PVSPs			
Factor loadings	Strength of covariates ( $\Delta R^2$ )	Covariances among residuals	Power to detect $\beta_{YX}$
Low (.35-.45)	Low (.03)	Low (.10)	.19
Low (.35-.45)	Low (.03)	High (.60)	.33
Low (.35-.45)	High (.27)	Low (.10)	.04
Low (.35-.45)	High (.27)	High (.60)	.28
High (.65-.75)	Low (.03)	Low (.10)	.96
High (.65-.75)	Low (.03)	High (.60)	.95
High (.65-.75)	High (.27)	Low (.10)	.93
High (.65-.75)	High (.27)	High (.60)	.89

Note: PVSP = plausible values for secondary parameter.

First, we strongly recommend Monte Carlo approaches for the estimation of power in the context of complex statistical models (e.g., Schoemann et al., 2017). Fortunately, Monte Carlo utilities are available in many of the same statistical packages that are often used to test such models (e.g., R, Mplus, Stata, SAS). Monte Carlo methods are especially important in situations in which closed-form power calculations are difficult or impossible or when researchers want to take into consideration such data characteristics as missingness, nonnormality, reliability, and so on. However, Monte Carlo power estimation can also be illuminating even when analytic approaches to power are available. In addition, several tutorials and applications exist that can facilitate these kinds of power analyses for certain kinds of statistical models (see Spybrook et al., 2016; van Breukelen & Candel, 2012; Wang & Rhemtulla, 2021). For example, in multiple regression, power calculations simply based on a medium  $f^2$  (Cohen, 1988) can obscure the impact of individual model parameters on the power to detect a particular slope.

Second, in our experience, many grant proposals cite previous studies to justify why the primary effect will be as large as the researchers claim it will be in their power analyses. Ironically, when the primary effect size can be well justified by previous studies, the need for the current grant proposal may be diminished. Why do (or fund) a study if how big the effect size will be is already known? However, using prior research to justify the size of the secondary model parameters is critical. When the project involves complex statistical models, the power to detect the primary effect will depend on the values of many secondary model parameters. In our experience, the types of parameters that have the greatest impact on power will vary from model to model and will depend on the range of PVSPs under consideration.

Often, however, simulations reveal that power ultimately hinges on one or two broad types of PVSPs. Monte Carlo approaches for power analysis require the investigator to assign values to these secondary characteristics. We recommend that these values (a) be reported and (b) should be well justified by prior research. Such prior research need not resemble the proposed study as long as it provides the information of interest. For example, in SEM, identifying lowest and highest plausible values for factor loadings could be based on cross-sectional research even though the study of interest is longitudinal.

Third, we suggest a six-step strategy for Monte Carlo power estimation: (1) Set the primary parameter to be the smallest effect size of interest or to be equal to an effect size predicted by theory (Anderson, 2019; Lakens, 2014). This is based on the premise that the researcher wants enough power to detect a value that is deemed to be important. (2) Then decide on the lowest and the highest plausible values for each secondary parameter.<sup>5</sup> Ideally, these values will be justified by previous research. (3) Conduct Monte Carlo power analyses for various combinations of high and low values for these PVSPs (in a full factorial design, if practical), bearing in mind that power might well be different for different hypotheses embedded in the same model. (4) When possible, use these results to inform the reallocation of resources or the modification of the study design so that power is at least adequate under the worst PVSP conditions for all hypotheses. (5) Report both the highest and lowest estimates of power to detect the primary parameter(s). These estimates constitute the best- and worst-case scenarios (i.e., the upper and lower bounds for one's power). Such a PVSP approach makes explicit the PVSP values on which power estimation is based (even when their actual values are uncertain), thus enabling the reader to evaluate the underlying assumptions. Despite the amount of work that goes into such analyses, the results can be summarized quite efficiently. (6) Finally, we recommend that authors include their simulation code in appendices to their articles and grant proposals. Doing so will (a) improve the transparency of power analyses, (b) enable reviewers to insert and test the effects of their own PVSPs, and (c) lower the threshold for novice users of code to start writing their own simulations.

Fourth, The PVSP approach can also be adapted to situations in which the research goal is to obtain a precise estimate of one or more parameters (e.g., Maxwell et al., 2008). The simulation design remains the same, but instead of focusing on sample size and power, the goal switches to obtaining sufficiently narrow confidence intervals for key parameters. Of course, this necessitates a judgment as to what constitutes "sufficiently narrow," but otherwise, the statistical aspects of

this approach follow the same essential logic as when the objective is to estimate power.

Fifth, some models are used to test multiple hypotheses. In some studies, one's power to detect all primary parameters may be similarly affected by changes in the secondary parameters. Such was the case in our mediation example. However, this will not be true for all multihypothesis models, as demonstrated in our RI-CLPM example. In such cases, researchers would do well to take the necessary a priori steps to ensure that their research design generates sufficient power to test all hypotheses.

Finally, we want to emphasize that this method can reveal ways to increase power by methods other than increasing sample size. Informed by power analyses that examine the effect of different PVSPs, researchers might change their research designs in other (sometimes more cost-effective) ways to improve power. Such analyses can inform resource allocation. For some complex analyses, increasing sample size might be the most effective way to improve power; for others, improving measurement might be a wiser course. The PVSP with the greatest impact on power will depend on the model, the hypothesis in question, and the values of sometimes multiple secondary model parameters. For instance, in some models, improving measurement may be a more efficient and cost-effective method of increasing power than increasing sample size.

### **Limitations, Future Directions, and Conclusions**

In the current article, we proposed a method for estimating and reporting power to test hypotheses that are embedded in complex statistical models. Given the frequent use of complex models in grant proposals and published research, such methods are critical. Nevertheless, several limitations of the current article suggest avenues for future research. First, we did not consider all types of complex models, not even close. Indeed, that was not our goal. Instead, we demonstrated our approach by applying it to four specific (and increasingly complex) models. Applications to qualitatively different models may require consideration of different secondary parameters.

Second, we did not consider all possible ways that secondary model parameters might be clustered when taking a factorial, empirical approach to power. For example, we treated factor loadings en masse and did not consider the possibility that improved measurement in some parts of a model might have a stronger effect on power than it would in other parts. We regard questions such as these to be important domains for future research.

Third, knowing how the value of a particular secondary parameter affects power is sometimes difficult. Sometimes we relied on previous work (e.g., research on covariates summarized by Maxwell et al., 2017; research on collinearity among predictors by Mason & Perreault, 1991). Other times, we empirically examined the power-related effect of each secondary parameter in the specific model under investigation. To understand more completely the power-related effects of secondary model parameters is an important direction for future research.

Fourth, we advocated for an empirical (i.e., Monte Carlo) approach that examines power as a function of various types of secondary model parameters (Muthén & Muthén, 2002). Although the tools for such analyses are available via commonly used statistical programs (e.g., R, Mplus, Stata, SAS, and pwrSEM), systematically examining the effects of multiple secondary parameters can be repetitive. Developing programs in which a factorial approach to power is highly automated would be a valuable addition to the applied researcher's toolbox.

Fifth, we did not consider the complicating possibility that varying a particular secondary model parameter could (in real studies) be associated with changes in other parameters. For example, greater collinearity of the primary predictor with covariates might be associated with smaller effect sizes for the primary predictor. In a related vein, selecting PVSPs from diverse studies could combine to increase one's power to detect spurious effects, as can other kinds of fallacious within-studies assumptions (e.g., Cole & Preacher, 2014). Meta-analytic examination of correlations among model parameters is an important area in need of future research.

Sixth, some well-intentioned efforts to enhance power can have iatrogenic effects. One example is when increasing reliability by lengthening a measure accidentally induces fatigue, biased responding, or attrition. Another example occurs when researchers select methodologically similar measures of a construct to improve factor loadings but inadvertently infuse the underlying latent variable with systematic method variance. Care must be taken when enhancing power by changing the research design, lest the changes compromise the integrity of the original study.

Finally, some readers might be concerned that the programming required for PVSP Monte Carlo power analyses will be beyond the reach of many research teams. We would argue, however, that conducting this kind of power analysis for hypotheses embedded in complex statistical models is not all that different from conducting the data analyses themselves. In some programs, a few additional lines of code can often convert the data-analytic program to a Monte Carlo power-analytic program (as in Mplus). In other cases, recently constructed and easily implemented apps are available (e.g.,

pwrSEM). If the research team has the expertise to do one, it almost certainly has the expertise to do the other.

In summary, we describe the use of highest and lowest PVSPs in Monte Carlo estimates of power to test hypotheses embedded in complex statistical models. Using this method can reveal the most effective ways to improve power in the context of a particular research design beyond simply increasing sample size. This method can also reduce the reporting of ambiguous and unjustified power estimates. If implemented, our PVSP approach may prevent the funding of underpowered studies and consequently, help to eliminate one reason for the current replication crisis.

## Transparency

*Action Editor:* Rogier Kievit

*Editor:* David A. Sbarra

*Author Contributions*

**David A. Cole:** Conceptualization; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision; Validation; Writing – original draft; Writing – review & editing.

**George Abitante:** Conceptualization; Formal analysis; Methodology; Writing – review & editing.

**Hoi Kan:** Conceptualization; Data curation.

**Qimin Liu:** Conceptualization; Formal analysis; Validation.

**Kristopher J. Preacher:** Conceptualization; Methodology; Writing – review & editing.

**Scott E. Maxwell:** Conceptualization; Methodology; Writing – review & editing.

*Declaration of Conflicting Interests*

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

*Funding*

George Abitante was supported by NIH T32 MH018921 during completion of this work.

## ORCID iDs

David A. Cole  <https://orcid.org/0000-0002-7535-1230>

George Abitante  <https://orcid.org/0000-0002-6422-8974>

## Acknowledgments

Hoi Kan is now affiliated with Princeton University. We are grateful to Patricia and Rodes Hart for the generous endowment that supported this research.

## Supplement Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/supp10.1177/25152459241302300>.

## Notes

1. G\*Power (<http://www.gpower.hhu.de/>), PowerUp! (<https://www.causalevaluation.org/>), WebPower (<http://webpower.psychstat.org>), SampSize (<https://www.epigenesys.org.uk/portfolio/>

<https://www.power-analysis.com>), Power Analysis and Sample Size (PASS: <https://www.ncss.com/software/pass>), nQUERY (<https://www.statsols.com/nquery/sample-size-software-options>), General Linear Mixed Model Power and Sample Size (GLIMPSE: <https://glimpse.samplesizeshop.org>), SAS PROC POWER, and PROC GLMPower. See also Schoemann et al. (2017).

2. Of course, various study-specific characteristics (e.g., attrition, measurement reliability) can also affect power and can also be examined using Monte Carlo methods; however, in the current article, we primarily focus on the primary and secondary parameters of the researcher's statistical model.

3. In real applications, high collinearity may also reduce the unique effect of the target predictor because more of its contribution to the dependent variable will be accounted for by the covariate. This would reduce power even more by reducing the primary effect size. In the current example, however, we hold  $\beta_1$  constant.

4. We elected not to use Mplus or pwrSEM for these analyses because they rely on Wald tests of the indirect effect, which rely on the incorrect assumption that  $ab$  is normally distributed and often lead to the underestimate of power (Fritz & MacKinnon, 2007; MacKinnon et al., 2004).

5. In practice, not all combinations of PVSPs will be possible. For example, the combination of several large slopes with large correlations among the predictors could produce  $R^2$  values that exceed 1.0. In such circumstances, researchers will have to make judgments about the plausibility of various combinations of study characteristics.

## References

- Anderson, S. F. (2019). Best (but oft forgotten) practices: Sample size planning for powerful studies. *The American Journal of Clinical Nutrition*, *110*, 280–295.
- Anderson, S. F., & Maxwell, S. E. (2017). Addressing the “replication crisis”: Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, *52*, 305–324. <https://doi.org/10.1080/00273171.2017.1289361>
- Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.
- Clarkson-Smith, L., & Hartley, A. A. (1990). Structural equation models of relationships between exercise and cognitive abilities. *Psychology and Aging*, *5*, 437–446. <https://doi.org/10.1037/0882-7974.5.3.437>
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). Lawrence Erlbaum Associates, Inc.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203771587>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*, 300–315. <https://doi.org/10.1037/a0033805>
- Collins, L. M., Dziak, J. J., & Li, R. (2009). Design of experiments with multiple independent variables: A resource

- management perspective on complete and reduced factorial designs. *Psychological Methods*, 14, 202–224. <https://doi.org/10.1037/a0015826>
- Curran, P. J., & Bollen, K. A. (2001). The best of both worlds: Combining autoregressive and latent curve models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 105–136). American Psychological Association.
- Du, H., Liu, F., & Wang, L. (2017). A Bayesian fill-in method for correcting for publication bias in meta-analysis. *Psychological Methods*, 22, 799–817. <https://doi.org/10.1037/met0000164>
- Du, H., & Wang, L. (2016). A Bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivariate Behavioral Research*, 51, 589–605. <https://doi.org/10.1080/00273171.2016.1191324>
- Fong-Jia, W., Chih-Fu, C., Mei-Yen, C., & Kim-Wai, R. (2020). Temporal precedence of physical literacy and basic psychological needs satisfaction: A cross-lagged longitudinal analysis of university students. *International Journal of Environmental Research and Public Health*, 17, Article 4615. <https://doi.org/10.3390/ijerph17124615>
- Fritz, M. S., Cox, M. G., & MacKinnon, D. P. (2015). Increasing statistical power in mediation models without increasing sample size. *Evaluation & the Health Professions*, 38(3), 343–366.
- Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18, 233–239.
- Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641–651. <https://doi.org/10.1177/1745691614551642>
- Grunewald, W., Troop-Gordon, W., & Smith, A. R. (2022). Relationships between eating disorder symptoms, muscle dysmorphia symptoms, and suicidal ideation: A random intercepts cross-lagged panel approach. *International Journal of Eating Disorders*, 55(12), 1733–1743. <https://doi.org/10.1002/eat.23819>
- Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). McGraw-Hill.
- Gustafsson, H. C., Cox, M. J., & Blair, C. (2012). Maternal parenting as a mediator of the relationship between intimate partner violence and effortful control. *Journal of Family Psychology*, 26, 115–123. <https://doi.org/10.1037/a0026283>
- Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, 20, 102–116.
- Hu, J., & Ai, H. (2016). Self-esteem mediates the effect of the parent-adolescent relationship on depression. *Journal of Health Psychology*, 21, 897–904. <https://doi.org/10.1177/1359105314541315>
- In-Albon, T., Meyer, A. H., Metzke, C. W., & Steinhausen, H. (2017). A cross-lag panel analysis of low self-esteem as a predictor of adolescent internalizing symptoms in a prospective longitudinal study. *Child Psychiatry and Human Development*, 48, 411–422. <https://doi.org/10.1007/s10578-016-0668-x>
- Isaksson, J., Sjöblom, S., Schwab-Stone, M., Stickley, A., & Ruchkin, V. (2020). Risk factors associated with alcohol use in early adolescence among American inner-city youth: A longitudinal study. *Substance Use & Misuse*, 55, 358–366. <https://doi.org/10.1080/10826084.2019.1671867>
- James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Sage.
- Ju, S., & Lee, Y. (2018). Developmental trajectories and longitudinal mediation effects of self-esteem, peer attachment, child maltreatment and depression on early adolescents. *Child Abuse & Neglect*, 76, 353–363. <https://doi.org/10.1016/j.chiabu.2017.11.015>
- Kenny, D. A. (1979). *Correlation and causality*. John Wiley & Sons.
- Kojima, R., Shinohara, R., Akiyama, Y., Yokomichi, H., & Yamagata, Z. (2021). Temporal directional relationship between problematic internet use and depressive symptoms among Japanese adolescents: A random intercept, cross-lagged panel model. *Addictive Behaviors*, 120, Article 106989. <https://doi.org/10.1016/j.addbeh.2021.106989>
- Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7), 701–710.
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- 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. <https://doi.org/10.1037/a0024776>
- Liu, X., & Wang, L. (2019). Sample size planning for detecting mediation effects: A power analysis procedure considering uncertainty in effect size estimates. *Multivariate Behavioral Research*, 54, 822–839. <https://doi.org/10.1080/00273171.2019.1593814>
- Lucas-Thompson, R., & Hostinar, C. E. (2013). Family income and appraisals of parental conflict as predictors of psychological adjustment and diurnal cortisol in emerging adulthood. *Journal of Family Psychology*, 27, 784–794. <https://doi.org/10.1037/a0034373>
- 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. <https://doi.org/10.1037/1082-989X.7.1.83>
- 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. [https://doi.org/10.1207/s15327906mbr3901\\_4](https://doi.org/10.1207/s15327906mbr3901_4)
- Mason, C. H., & Perreault, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research*, 28, 268–280. <https://doi.org/10.2307/3172863>
- Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods*, 5, 434–458. <https://doi.org/10.1037/1082-989X.5.4.434>
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9(2), 147–163. <https://doi.org/10.1037/1082-989X.9.2.147>

- Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). Routledge. <https://doi.org/10.4324/9781315642956>
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. <https://doi.org/10.1146/annurev.psych.59.103006.093735>
- McShane, B. B., & Böckenholt, U. (2015). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods*, *21*, 47–60. <https://doi.org/10.1037/met0000036>
- Mega, C., Ronconi, L., & De Beni, R. (2014). What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of Educational Psychology*, *106*, 121–131. <https://doi.org/10.1037/a0033546>
- Muhtadie, L., Zhou, Q., Eisenberg, N., & Wang, Y. (2013). Predicting internalizing problems in Chinese children: The unique and interactive effects of parenting and child temperament. *Development and Psychopathology*, *25*, 653–667. <https://doi.org/10.1017/S0954579413000084>
- Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (4th ed.). Routledge/Taylor & Francis Group.
- Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599–620. [https://doi.org/10.1207/S15328007SEM0904\\_8](https://doi.org/10.1207/S15328007SEM0904_8)
- Nguyen, D. T., Wright, E. P., Dedding, C., Pham, T. T., & Bunders, J. (2019). Low self-esteem and its association with anxiety, depression, and suicidal ideation in Vietnamese secondary school students: A cross-sectional study. *Frontiers in Psychiatry*, *10*, Article 698. <https://doi.org/10.3389/fpsy.2019.00698>
- O'Brien, R. G., & Muller, K. E. (1993). Unified power analysis for t-tests through multivariate hypotheses. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 297–344). Marcel Dekker.
- Orcutt, H. K. (2006). The prospective relationship of interpersonal forgiveness and psychological distress symptoms among college women. *Journal of Counseling Psychology*, *53*, 350–361. <https://doi.org/10.1037/0022-0167.53.3.350>
- Orth, U., Clark, D. A., Donnellan, M. B., & Robins, R. W. (2021). Testing prospective effects in longitudinal research: Comparing seven competing cross-lagged models. *Journal of Personality and Social Psychology*, *120*, 1013–1034. <https://doi.org/10.1037/pspp0000358>
- Pek, J., & Park, J. (2019). Complexities in power analysis: Quantifying uncertainties with a Bayesian-classical hybrid approach. *Psychological Methods*, *24*, 590–605. <https://doi.org/10.1037/met0000208>
- Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*, 319–332. <https://doi.org/10.1177/1745691614528519>
- Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*, 77–98. <https://doi.org/10.1080/00273171.2012.715557>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Reed-Fitzke, K., Withers, M. C., & Watters, E. R. (2021). Longitudinal connections of self-esteem and depression among adult children and their parents. *Journal of Adult Development*, *28*, 237–250. <https://doi.org/10.1007/s10804-021-09371-7>
- Rigdon, E. E. (1994). Demonstrating the effects of unmodeled random measurement error. *Structural Equation Modeling*, *1*, 375–380. <https://doi.org/10.1080/10705519409539986>
- Rogosa, D. R. (1980). A critique of cross-lagged correlation. *Psychological Bulletin*, *88*, 245–258. <https://doi.org/10.1037/0033-2909.88.2.245>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Rubio, D. M., & Gillespie, G. F. (1995). Problems with error in structural equation models. *Structural Equation Modeling*, *2*, 367–378. <https://doi.org/10.1080/10705519509540020>
- Saint-Georges, Z., & Vaillancourt, T. (2020). The temporal sequence of depressive symptoms, peer victimization, and self-esteem across adolescence: Evidence for an integrated self-perception driven model. *Development and Psychopathology*, *32*, 975–984. <https://doi.org/10.1017/S0954579419000865>
- Sakaluk, J. K. (2016). Exploring small, confirming big: An alternative system to the new statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology*, *66*, 47–54. <https://doi.org/10.1016/j.jesp.2015.09.013>
- 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*, 379–386. <https://doi.org/10.1177/1948550617715068>
- Sick, K., Pila, E., Nesbitt, A., & Sabiston, C. M. (2020). Does self-compassion buffer the detrimental effect of body shame on depressive symptoms? *Body Image*, *34*, 175–183. <https://doi.org/10.1016/j.bodyim.2020.05.012>
- Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. *International Journal of Research & Method in Education*, *39*, 255–267. <https://doi.org/10.1080/1743727X.2016.1150454>
- Takagishi, Y., Sakata, M., & Kitamura, T. (2011). Effects of self-esteem on state and trait components of interpersonal dependency and depression in the workplace. *Journal of Clinical Psychology*, *67*, 918–926. <https://doi.org/10.1002/jclp.20815>
- Taylor, L. K., Tong, X., & Maxwell, S. E. (2020). Evaluating supplemental samples in longitudinal research: Replacement and refreshment approaches. *Multivariate Behavioral Research*, *55*, 277–299. <https://doi.org/10.1080/00273171.2019.1628694>

- van Breukelen, G. J. P., & Candel, M. J. J. M. (2012). Calculating sample sizes for cluster randomized trials: We can keep it simple and efficient! *Journal of Clinical Epidemiology*, *65*, 1212–1218. <https://doi.org/10.1016/j.jclinepi.2012.06.002>
- Von Korff, L., & Grotevant, H. D. (2011). Contact in adoption and adoptive identity formation: The mediating role of family conversation. *Journal of Family Psychology*, *25*, 393–401. <https://doi.org/10.1037/a0023388>
- Wang, Y. A., & Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*, *4*. <https://doi.org/10.1177/2515245920918253>
- Wei, M., Shaffer, P. A., Young, S. K., & Zakalik, R. A. (2005). Adult attachment, shame, depression, and loneliness: The mediation role of basic psychological needs satisfaction. *Journal of Counseling Psychology*, *52*, 591–601. <https://doi.org/10.1037/0022-0167.52.4.591>
- Wouters, S., Duriez, B., Luyckx, K., Klimstra, T., Colpin, H., Soenens, B., & Verschuere, K. (2013). Corrigendum to “Depressive symptoms in university freshmen: Longitudinal relations with contingent self-esteem and level of self-esteem.” *Journal of Research in Personality*, *47*, 952. <https://doi.org/10.1016/j.jrp.2013.10.001>
- Wu, W., Carroll, I. A., & Chen, P. Y. (2018). A single-level random-effects cross-lagged panel model for longitudinal mediation analysis. *Behavior Research Methods*, *50*(5), 2111–2124. <https://doi.org/10.3758/s13428-017-0979-2>
- Zhang, Z., Mai, Y., Yang, M., & Zhang, M. Z. (2018). Package ‘WebPower.’ *Basic and advanced statistical power analysis version*, *72*, 555.
- Zhang, Z., & Yuan, K.-H. (2018). *Practical statistical power analysis using WebPower and R*. ISDSA Press.