

Psychological Methods

nmax and the Quest to Restore Caution, Integrity, and Practicality to the Sample Size Planning Process

Gregory R. Hancock and Yi Feng

Online First Publication, August 11, 2025. <https://dx.doi.org/10.1037/met0000776>

CITATION

Hancock, G. R., & Feng, Y. (2025). nmax and the quest to restore caution, integrity, and practicality to the sample size planning process. *Psychological Methods*. Advance online publication. <https://dx.doi.org/10.1037/met0000776>

©American Psychological Association, [2025]. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. The final article is available, upon publication, at: [10.1037/met0000776](https://dx.doi.org/10.1037/met0000776)

n_{\max} and the Quest to Restore Caution, Integrity, and Practicality to the Sample Size Planning Process

Gregory R. Hancock¹ and Yi Feng²

¹ Department of Human Development and Quantitative Methodology, University of Maryland

² Department of Psychology, University of California, Los Angeles

Abstract

In a time when the alarms of research replicability are sounding louder than ever, mapping out studies with statistical and inferential integrity is of paramount importance. Indeed, funding agencies almost always require grant applicants to present compelling *a priori* power analyses to justify proposed sample sizes, as a critical part of the information considered collectively to ensure a sound investment. Unfortunately, even researchers' most sincere attempts at sample size planning are fraught with the fundamental challenge of setting numerical values not just for the focal parameters for which statistical tests are planned, but for each of the model's other, more peripheral or contextual parameters as well. As we plainly demonstrate, regarding the latter parameters, even in very simple models, any slight deviation in well-intentioned numerical guesses can undermine power for the assessment of the more focal parameters that are of key theoretical interest. Toward remedying this all-too-common but seemingly underestimated problem in power analysis, we adopt a hope-for-the-best-but-plan-for-the-worst mindset and present new methods that attempt to (a) restore appropriate conservatism and robustness, and in turn credibility, to the sample size planning process, and (b) greatly simplify that process. Derivations and suggestions for practice are presented using the framework of measured variable path analysis models as they subsume many of the types of models (e.g., multiple linear regression, analysis of variance) for which sample size planning is of interest.

Translational Abstract


Among the many critical decisions when planning a study is the number of subjects to be sampled. From a statistical perspective the basis for such a decision is to achieve adequate power, that is, to gather enough data to have an acceptably high probability of detecting the variable effects or relations of primary theoretical interest. A key challenge in doing so, however, is accurately anticipating the context in which those exist. Focal variables' relations with covariates, for example, while not necessarily of keen interest, if set inaccurately can result in a study with inadequate sample size, and in turn little power to detect those effects or relations that are of interest. And yet it is virtually impossible to have sufficiently pertinent prior information with which to foretell that context, leaving researchers to fill gaps in their knowledge with little more than wishful thinking. The current work seeks to address this long-standing challenge, proposing an insurance policy of sorts in which contextual relations are numerically consolidated and then set conservatively, thereby helping researchers to plan for sample sizes that are robust to unanticipated and unfavorable contextual conditions, and ensuring statistical power to detect the focal effects and relations of theoretical importance.

Keywords: power analysis, sample size planning, measured variable path analysis, structural equation modeling, multiple linear regression

Power analysis has lost its way. To be specific, we are not speaking of the post hoc variety whereby statistical tests already conducted (and which usually failed to yield statistical significance) are used

to estimate their own (inadequate) statistical power. Such a "post-mortem examination" as Fisher (1938) famously referred to it, and many others have since echoed (e.g., Hoenig & Heisey, 2001;

Samantha F. Anderson served as action editor.

Gregory R. Hancock  <https://orcid.org/0000-0002-6313-006X>

Versions of this work have been presented at the Annual Meeting of the American Educational Research Association, the Modern Modeling Methods Conference, the Meeting of the Society of Multivariate Experimental Psychology, and by invitation at numerous universities nationally and internationally. We are most grateful to Fred Oswald, Roy Levy, Ethan McCormick, and Dan McNeish for their helpful comments on earlier versions of this work. Patrick Curran also provided comments. The authors have no known conflicts

of interest to disclose.

Gregory R. Hancock served as lead for conceptualization and writing—original draft. Yi Feng served in a supporting role for conceptualization and writing—original draft. Gregory R. Hancock and Yi Feng contributed equally to methodology.

Correspondence concerning this article should be addressed to Gregory R. Hancock, Department of Human Development and Quantitative Methodology, University of Maryland, 1230D Benjamin Building, 3942 Campus Drive, College Park, MD 20742-1115, United States. Email: ghancock@umd.edu

Maxwell, 2004; Yuan & Maxwell, 2005), is well known to be fraught with problems, not the least of which is quite simply that the study is over—the patient has already died. Rather, here we use the term *power analysis* in reference to its more useful a priori form, also known as *sample size planning*, in which a researcher attempts to estimate ahead of the study what sample size should be adequate in order to have some desired level of power for the statistical tests aimed at addressing the research questions of key theoretical interest. It is this a priori power analysis, this sample size planning, that we believe has lost its way.

In making this as-yet unjustified statement, we in no way wish to diminish the decades of methodological work that have extended the process of sample size planning to models and analyses of increasing complexity (see, e.g., Chattopadhyay et al., 2025; Donnelly et al., 2023; Feng & Hancock, 2021, 2023; Hancock, 2001; Hedges & Pigott, 2001, 2004; Mathieu et al., 2012; Moerbeek, 2022; Thoemmes et al., 2010; Tu et al., 2004; Wolf et al., 2013; Zhang, 2014) and that have attempted to make that process more accessible to applied researchers through dedicated software (e.g., G*Power, Faul et al., 2007; *semPower*, Moshagen & Bader, 2024; for a more comprehensive list see Feng & Hancock, 2023). All of this work is critically important methodologically and practically, and we hope it continues to flourish and reach the widest audience possible.

In order to understand our concern with power analysis, let us start by turning the clock back, say, 40 years, to a time when being versed in power analysis meant that one was armed with a well-worn copy of Cohen's classic treatise (e.g., the 1977 revised edition or the 1988 second edition). If you wished to conduct sample size planning for an independent samples *t* test, a χ^2 test of independence, or a host of other statistical tests with no broader variable-related or data-related context (e.g., covariates, complex samples), all you had to do was choose your test's intended Type I error rate (e.g., $\alpha = .05$), the desired level of power (e.g., $\pi = .80$), and the target effect size (e.g., Cohen's $d = 0.20$), and then turn to the appropriate table to reveal the necessary n (under standard distributional assumptions). And this was important not just statistically, but ideologically: it embodied the beliefs that science worth doing is worth planning for carefully, and that sample size is an investment—an insurance policy, so to speak—to guard against a potentially wasteful research endeavor. But alas, as foundational and grounding as this work was, when you step off the bus today clutching your copy of Cohen and gaze wide-eyed at the statistical skyscrapers all around you, it is clear that life is no longer so simple.

Consider as a generic starting example a fairly typical multiple linear regression model with a continuous outcome Y and four continuous predictors, where X is a predictor whose partial slope with Y is of focal research interest, and the remaining three predictors/covariates C_1 , C_2 , and C_3 are control variables that set the context for addressing the primary research question (e.g., demographics such as family income and parents' education level).¹ Expressed in standardized form for simplicity, the linear model is $Y = \gamma X + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \varepsilon$, where for this example the parameter for which sample size planning is desired is γ . After setting the intended α level for the statistical test of γ (e.g., $.05$), as well as the target power level π (e.g., $.80$), the task turns to the variable relations.

The relation of primary importance, that for which the study wishes to have sufficient power, is γ . Let us assume that the researcher, by whatever process, believes the standardized value $\gamma = .20$ to represent the *smallest effect size of interest* (SESOI;

Lakens, 2022; Lakens et al., 2018; also referred to as the *critical effect size* by Kraemer & Blasey, 2015, p. 11).² Given this information, along with the chosen levels for α and π , one cannot merely consult a table or, more modernly, insert those values into a software application to get the necessary sample size. This is because that sample size also depends on (a) the relations among the C variables (ρ_{21} , ρ_{31} , and ρ_{32}), (b) the relations that the C variables have with X (ρ_{1X} , ρ_{2X} , and ρ_{3X}), and (c) the partial slopes relating the C variables to the outcome Y above and beyond X (β_1 , β_2 , and β_3). And here is where things become more tenuous (as any statistical consultant who has tried to guide an applied researcher through this process will attest). In this example, there are nine (standardized) parameters that are peripheral to the researcher's theoretical interest, and thus for which power is not of primary concern, but that are in fact essential context for determining the sample size needed to test the parameter that is focal, γ . And realistically speaking, knowing the precise population values for all nine of these contextual parameters a priori is simply incredibly unlikely (as we will expand upon further later).³

So, what do researchers do then for these contextual parameters? They make guesses—thoughtful, educated guesses we hope, but guesses nonetheless. If, for example,⁴ a researcher were to choose $\rho_{21} = \rho_{31} = \rho_{32} = .35$, $\rho_{1X} = \rho_{2X} = \rho_{3X} = .30$, and $\beta_1 = \beta_2 = \beta_3 = .25$, employing one of the approaches reviewed in the next section yields an estimated necessary sample size of $n = 134$ for a maximum likelihood (ML) based test of the SESOI $\gamma = .20$ (assuming $\alpha = .05$ and $\pi = .80$). Of course, the precision of the sample size estimate is only as good as the assumptions upon which it rests, including, but not limited to, the accuracy of the values of the contextual parameters. If accurate (and if all other assumptions hold), this n is exactly the insurance policy needed to address the research question addressed by γ . If inaccurate, however, the resulting power would likely deviate from the target level π , possibly substantially so. On the one hand, under such circumstances, this n could possibly yield more power than planned, thus still serving as an effective insurance policy for testing the focal research question. Alternatively, and much more worrisome for our purposes, inaccurate guesses about the contextual parameters could yield less power when testing γ . We simply do not know, because we do not know the state of the population (indeed, if we did, the research would be unnecessary). Even the most educated guesses about parameters, focal or in this case contextual, can be misinformed due to publication bias, sampling variability, model (mis)

¹ A researcher could have theoretical interest in the partial slopes of more than one predictor of Y within a given model; for this motivating example we will focus on just one predictor.

² That is to say, and we emphasize here, that we are not asking here or anywhere in this article that the researcher approximate what is believed to be the "true" effect size, which for countless reasons can be misinformed and, especially problematically for planning purposes, upwardly biased. Rather, we are asking the researcher to set the smallest justifiable value having practical meaning for their constituent communities and being worthy of investigation.

³ If there are multiple focal parameters in a model, as is common, sample size planning should be conducted for each focal parameter. In such cases, a specific parameter will be focal in one analysis but contextual in another. Consequently, as will be addressed and illustrated in this article, when serving focally a parameter's worst-case scenario will be governed by its SESOI; when shifting to a contextual role that parameter's uncertainty, particularly in a problematic direction for planning purposes, must be accounted for.

⁴ Values were chosen to be equal merely for the simplicity of this example; researchers would be free to choose values differing within and across subsets (as long as their implied correlation matrix was positive definite [PD]).

specification, and so forth (see, e.g., Anderson et al., 2017; McShane & Böckenholt, 2016; Pek & Park, 2019; Pek et al., 2024; Perugini et al., 2014). Furthermore, the sample size of $n = 134$ might be an unexpectedly costly pill for the researcher to swallow, provoking initial responses such as: “Are you sure?”, “Do you have any idea how hard it is to get subjects from this population?!” and “We absolutely cannot afford to get that many.” But this reaction is typically short-lived, quickly giving way to negotiation, or the so-called “sample size samba” (Schulz & Grimes, 2005): “Well, what if we make that correlation a little smaller?”, “Can we make that β a bit bigger?”, “Let’s take out that third control variable and see what happens.”, and so on.

Now, to be clear, we absolutely believe that power analysis can be a place for principled exploration (e.g., Judd et al., 2017). It should be a process in which one evaluates the consequences of different values for contextual parameters (e.g., Cole et al., 2025), rather than a quest for some single sacred n to be revealed. In our view, however, such exploration should be for the purpose of probing concerning scenarios against which to be insured, not prospecting for a less expensive n and then cobbling together a story to convince oneself (and grant reviewers) that this has been some contextual truth all along. This latter, and in our experience extremely common, variation of the sample size samba is completely at odds with the original spirit of power analysis. What was intended to be an honest, playful insurance policy for our scientific endeavors has essentially devolved into a statistical yard sale, haggling to slink away as cheaply as possible. It is for this reason that we believe power analysis has lost its way.

The work we present here aims to confront this problematic mindset and the practices it precipitates, offering new approaches in the quest to (a) restore appropriate conservatism and robustness, and in turn credibility, to the sample size planning process, and (b) greatly simplify that process by reducing the amount of speculation required of the researcher. We provide derivations, illustrative simulations, and suggestions for practice that have broad applications to multiple linear regression and the larger analytic framework of measured variable path analysis, of which many methods are special cases. A discussion of the benefits, limitations, and opportunities for further extensions then follows.

Power for the Most Basic (Two-Predictor) Multiple Linear Regression Model

We will start with the most basic multiple linear regression model, a simplified version of the one presented above, having a continuous outcome Y , one continuous focal predictor X , and only one continuous contextual covariate C .⁵ This scenario is certainly of practical utility in its own right, but even more importantly, will serve as the basis for larger models discussed at length in subsequent sections of this article. Again, we will express the model in standardized form: $Y = \gamma X + \beta C + \varepsilon$, where the parameter for which sample size planning is desired is γ . After setting the intended α level for the test of γ (e.g., .05), as well as the target power π (e.g., .80), the researcher must choose the SESOI for the focal parameter γ : we will select the standardized value $\gamma = .30$ for this example. As for the contextual parameters, there are two in this scenario: (a) the relation between C and X , which for this example we will assume the researcher has selected as $\rho = .20$, and (b) the partial slope relating C variable to Y , which we will assume the researcher had a strong rationale for setting to $\beta = .40$ (and which we will address more

broadly later). These values lead to a model-implied correlation matrix (e.g., through path tracing; Wright, 1934) for X , C , and Y , respectively, of:

$$\begin{bmatrix} 1 & .20 & .38 \\ .20 & 1 & .46 \\ .38 & .46 & 1 \end{bmatrix}. \quad (1)$$

In order to assess the sample size needed for the current circumstance, several well-known and well-understood asymptotically equivalent approaches exist within the structural equation modeling (SEM) literature. One popular method is simulation-based (Muthén & Muthén, 2002), in which a large number of random samples are drawn from the population as specified (in our example, from a standardized population with the above correlation matrix), fitting the model (using ML) each time. Sample size is then adjusted iteratively until the proportion of α -level focal parameter tests achieving statistical significance is at least π .

In addition to the simulation strategy, there are two population analysis methods we will mention here (see, e.g., Feng & Hancock, 2023; Hancock & French, 2013). In the first, which is based on the focal parameter’s Wald test (i.e., its squared z value), the model is fitted directly to the population moments implied by the chosen parameter values (using ML), assuming an arbitrary value of n to start. Sample size is then adjusted until the z value (i.e., the square root of the Wald statistic) for the focal parameter test is as close as possible to, but no smaller than, the noncentrality parameter λ_z for a noncentral normal distribution corresponding to power π for α -level z tests (e.g., for $\pi = .80$ and $\alpha = .05$, $\lambda_z \approx 2.802$ to three decimals). In the second population analysis method, which utilizes the model-level likelihood ratio (LR) test, the model is fitted (using ML) to the population moments implied by the chosen parameter values, assuming an arbitrary (and typically large) value of n to start, but constraining the focal parameter γ to 0. The resulting degree of misfit associated solely with the focal parameter’s absence is reflected in the value of the ML fit function, F_{ML} , which is known (see, e.g., Satorra & Saris, 1985) to convert to a 1 df χ^2 noncentrality parameter as:

$$\lambda = (n - 1)F_{ML}. \quad (2)$$

Noncentrality parameters necessary to achieve power π for α -level 1 df χ^2 tests may be looked up (e.g., Johnson et al., 1995) or computed (e.g., using the `chi2ncp` function in the `gtm` R package; Johnson, 2020); for $\alpha = .05$ and $\pi = .80$ the value to three decimal places is $\lambda \approx 7.849$.⁶ Substituting this target noncentrality parameter and

⁵ As alluded to previously, the researcher could also view both predictors as focal, with neither as covariate per se. However, as we are interested in the power associated with testing each individual focal predictor’s partial slope, each predictor in turn may be viewed as the other predictor’s pro tem covariate. Furthermore, the predictor and/or the covariate could be dichotomous, as in group coding; for our purposes, without loss of generality, we will assume they are continuous (and standardized) for ease of presentation and interpretability.

⁶ With $\alpha = .05$, $\pi = .80$, and $df = 1$, the `chi2ncp` function yields $\lambda = 7.848861$. This six-decimal value will be used in all calculations in this article. Furthermore, it may be noted that a 1 df χ^2 test is equivalent to the square of a z test; as such, the noncentrality parameter for the former is the square of the noncentrality parameter for the latter (e.g., $7.848861 = 2.801582^2$ for $\alpha = .05$ and $\pi = .80$).

rearranging the formula for n yields:

$$n = (\lambda/F_{ML}) + 1 = (7.849/F_{ML}) + 1. \quad (3)$$

Following this second population analysis approach and constraining $\gamma = 0$ in the above example, the resulting fit function⁷ is $F_{ML} = 0.116$. Substituting this value into Equation 3 yields $n = (7.849/0.116) + 1 = 68.66$, which, when rounded up to ensure sufficient power, suggests an estimated sample size of $n = 69$ to achieve $\pi = .80$ power for detecting $\gamma = .30$ using an $\alpha = .05$ level test, given standard distributional assumptions and contextual parameters of $\rho = .20$ and $\beta = .40$.

Although it may seem fairly straightforward to get to the sample size of $n = 69$ in this simple example, doing so brings up two reasonable questions: (a) how likely is it that researchers get the contextual parameters correct (ρ and β), and (b) how much does it matter if they do not? On the first issue, we would argue that the chances are almost always essentially zero. For the researcher who cites past research in choosing their numerical values, we ask (rather rhetorically, and possibly annoyingly) the following questions:

- Did the prior work draw from the exact same population as the planned study?
- Was the same model estimated, for example, with the same other variables (focal and contextual)?
- Were the same scales/measures used, and did they have comparable reliability and thus the same expected relations as in the planned study?
- If the prior contextual parameter values were based on the researcher's own preliminary pilot work rather than published studies, how far off could the estimates be given the typically much smaller pilot sample size (and hence much larger standard errors)?

As telegraphed above, it is extremely unlikely that any of these questions could be answered satisfactorily, a reality that is prominent in the ongoing replication crisis. Indeed, even when a planned study perfectly replicates previous published studies in terms of the target population, sampling method, measurement properties, analytical model, and estimation approach, other sources of uncertainty still likely bias a priori model parameter estimates. For instance, publication bias, inherent sampling variability in point estimates, model misspecification, and assumption violations have been discussed extensively by, for example, Anderson et al. (2017), Pek and Park (2019), and Perugini et al. (2014). Perhaps even more fundamentally problematic for our purposes is that, particularly in complex models, estimates of contextual parameters are commonly deemed background noise relative to the focal parameters constituting the main storyline, and thus are often omitted from research reports; as such subsequent researchers simply lack the necessary information with which to educate their guesses. All of this is to say that, however well intentioned the contextual parameter guesses may be, there are simply too many sources of mismatch between prior work and the planned study to completely trust those parameter values, and this issue only compounds as the model becomes increasingly complex.

This then leads to our second question: If researchers likely do not have accurate prior knowledge about the contextual parameters, how much does it matter when it comes to the test of the focal parameter γ ? As an initial foray into thinking about this question, we repeat the prior sample size estimation with the standardized SESOI $\gamma = .30$

Table 1
Required Sample Size to Detect $\gamma = .20$

ρ	β						
	-.6	-.4	-.2	0	.2	.4	.6
-.6	51	88	114	129	134	127	110
-.4	48	73	91	100	101	94	78
-.2	49	69	82	88	87	78	62
0	53	71	82	85	82	71	53
.2	62	78	87	88	82	69	49
.4	78	94	101	100	91	73	48
.6	110	127	134	129	114	88	51

Note. The bold value represents the true sample size needed under the assumed conditions. The gray square represents sample sizes when one or both of the contextual parameters are off by $\pm .20$.

but manipulating the (standardized) values of the contextual parameters ρ and β both from $-.60$ to $+.60$ in increments of $.20$. Table 1 shows the estimated sample size for an $\alpha = .05$ level test of γ to have $\pi = .80$ power (under assumed distributional conditions). At the intersection of the assumed targets of $\rho = .20$ and $\beta = .40$, we see the previously reported $n = 69$ in bold. Now, imagine that while planning for these specific values, the true contextual parameters could be off from those assumed values by a seemingly negligible $\pm .20$; this is represented in Table 1 by a light gray square circumscribed around the assumed $n = 69$. If $\beta = .40$ had been correct, we see that being off in ρ by $\pm .20$ (i.e., $\rho = 0$ or $\rho = .40$) would require a larger sample size (as indicated by the two gray cells above and below 69), albeit only by 4 or fewer subjects (i.e., $n = 69$ would technically lead to an underpowered study, but not severely). If β had been incorrect and its true value was actually higher by $+.20$ (i.e., $\beta = .60$), the planned sample size of $n = 69$ would be more subjects than actually needed for values of ρ within $\pm .20$ of the assumed $\rho = .20$ (as indicated by the three gray cells to the right of 69), thus safely providing power in excess of the target level (i.e., $\pi > .80$). On the other hand, if β had been incorrect and its true value was actually off by $-.20$ (i.e., $\beta = .20$), the planned sample size of $n = 69$ could fall well short of the needed sample sizes for values of ρ within $\pm .20$ of the assumed $\rho = .20$ (as indicated by the three gray cells to the left of 69). With $\beta = .20$ and $\rho = .40$, for example, the required sample size would be estimated to be $n = 91$ (i.e., 22 more subjects than the assumed $n = 69$). And that is not even close to the worst case shown in the table. In the values displayed, we see that when $\beta = -.20$ and $\rho = .60$, the required sample size would be estimated to be $n = 134$, which is almost twice the assumed $n = 69$ based on the incorrect contextual parameter values of $\rho = .20$ and $\beta = .40$. As a result, using $n = 69$ under such conditions would yield power estimated (by the simulation method mentioned above) to be only $.54$ to detect the focal $\gamma = .30$ (with an $\alpha = .05$ level test). One can hardly imagine investing valuable time and resources into a study with barely better than a coin flip's chance of success.

Thus, given that the contextual parameters can be difficult (if not impossible, as argued above) to determine a priori with precision, and that the imprecision in contextual parameters can lead to a severely underpowered study that can thwart even the relatively

⁷ We used Mplus 8.9 (Muthén & Muthén, 1998–2023).

simple two-predictor multiple linear regression case, we can carry this further and ask a very practical question: Is there a sample size that would guarantee a minimum of π power for an α -level test of γ ? Such information would be valuable for understanding the worst-case scenario, or framed more proactively, for estimating the cost of the safest insurance policy. Fortunately, this is a question that can be answered analytically.

As detailed in Appendix A, given a fixed focal parameter γ , we can mathematically derive the contextual conditions under which the statistical power is minimized for testing its estimate, and more importantly, the corresponding theoretical lower bound of F_{ML} . Doing so in turn allows us to determine the largest necessary sample size, n_{max} , that is required to detect (with target power π) the noncentrality introduced by constraining the focal parameter γ to 0. This n_{max} will thus serve as the insurance policy that can cover us even under the least favorable contextual conditions. As shown in Appendix A, this worst-case scenario for testing a given γ occurs as $|\rho| \rightarrow 1$ and $\beta = -\gamma\rho$, yielding a fit function F_{ML} of:

$$F_{ML} = \ln \frac{1}{1 - \gamma^2(1 - \rho^2)}. \quad (4)$$

Recalling from Equation 1 that the $df\chi^2$ noncentrality parameter λ when setting focal parameter γ to 0 is $\lambda = (n - 1)F_{ML}$, we may substitute the expression for F_{ML} from Equation 4 into Equation 1, yielding:

$$\lambda = (n - 1) \ln \frac{1}{1 - \gamma^2(1 - \rho^2)}. \quad (5)$$

Equation 5 may, in turn, be rearranged to solve for n_{max} as:

$$n_{max} = \frac{\lambda}{-\ln[1 - \gamma^2(1 - \rho^2)]} + 1. \quad (6)$$

This implies that, under standard assumed conditions of conditional multivariate normality and independence of observations, for a given level of collinearity ρ (where $|\rho| < 1$) the sample size n_{max} ensures power of at least π all the way to the conditional *pessimism*, that is, to the least favorable contextual conditions for that focal γ and the contextual ρ (and which do not likely correspond precisely to the select conditions shown in Table 1).

Although not technically necessary, we can use a table displaying n_{max} values for select levels of γ and a range of discrete $|\rho|$ values to facilitate a more intuitive understanding of how these aspects are interrelated. Imagine, for example, that a researcher considers $\gamma = .30$ to be the smallest effect worth detecting for X , and that the collinearity of X with covariate C would not be expected to exceed $|\rho| = .70$. Table 2, which draws directly from Equation 6, shows a corresponding value of $n_{max} = 169$. That is, under standard assumed conditions, a sample size of $n = 169$ would be expected to yield at least $\pi = .80$ power to detect $\gamma = .30$ with an $\alpha = .05$ level test (under assumed distributional conditions), as long as collinearity did not exceed $|\rho| = .70$. Thus, the researcher was not required to specify exact values for ρ and β , but rather allowed n_{max} essentially to “plan for the worst” in order to ensure power coverage under all conditions deemed reasonably possible, even if unlikely.

Now at this point, for the two-predictor case addressed here one might argue (a) that the method above offers only a slight simplification to the previous sample size planning processes, given that it still requires a researcher-supplied upper threshold of ρ , and (b)

that the absolution of responsibility over the now hard-wired and maximally pessimistic β likely brings with it the cost of increased sample size. To the first point we agree: we cannot, and perhaps should not, fully relieve the researcher of responsibility for the collinearity decision. However, we prefer to think of it differently: rather than the common “taking an educated guess” framing, we see this as the researcher choosing to adopt a conservative mindset and taking responsibility for setting a realistic upper limit on the degree of expected collinearity (an issue we will return to later in the article). And in response to the second point about cost, indeed uncertainty about contextual relations has a price; however, as part of the quest to restore the integrity of power analysis as an insurance policy against just such uncertainty, we believe that planning for the worst while hoping for the best is a reasonable governing ideology when faced with the pervasive tenuousness of a priori contextual parameter value selection. We will return to this point with further discussion at the end of the article.

Power for Multiple Linear Regression Models With More Than Two Predictors

As we will see in this section, the previous two-predictor derivation has highly useful applicability when generalized to models that increase in complexity. To start, we can extend this simplified power analysis process to a general linear model with any number of predictors. For our purposes, let us again think of one predictor X as focal and p covariates C_1 through C_p as contextual. Earlier, we described a standardized example with a focal predictor X plus $p = 3$ covariates, yielding a standardized focal slope parameter γ for X plus nine standardized contextual parameters: three correlations among the covariates, three correlations of X with the covariates, and three partial slopes relating the covariates to the outcome Y . More generally for p covariates, staying in the standardized realm for simplicity we would expect a total of $p(p + 3)/2$ contextual parameters: $p(p - 1)/2$ correlations among the covariates, p correlations of X with the covariates, and p partial slopes relating the covariates to the outcome Y . This means that in a model with, say, $p + 1 = 6$ exogenous variables in total (one focal X variable plus five C covariates), treating any of the variables as focal with partial slope parameter γ would have $p(p + 3)/2 = 5(5 + 3)/2 = 20$ contextual parameters. As argued previously in the case of $p + 1 = 2$ exogenous variables (i.e., the simplest possible multiple linear regression model), it is already infeasible to provide accurate guesses about such contextual parameter values. Of course, the challenge only grows as more contextual predictors are introduced, and the consequences for failing to provide accurate population values can be considerable in terms of sample size planning and power.

Fortunately, the principles and methods developed in the previous section for multiple linear regression models with one predictor and $p = 1$ covariate can be applied to scenarios with any number p of covariates. The reason stems from what we refer to here as *collapsibility*, that is, that the p covariates themselves can be thought of as being collapsed into, and ultimately represented by, a single composite covariate. To elaborate, within the familiar unstandardized multiple linear regression expression $Y = b_0 + gX + b_1C_1 + \dots + b_pC_p + e$, the p covariates could be thought of as forming a linear composite, which we will label C_* , that is separate from the focal predictor X (i.e., $C_* = b_1C_1 + \dots + b_pC_p$, and thus $Y = b_0 + gX + C_* + e$). The covariance structure of this traditional model may be

Table 2*n_{max} Values as a Function of $|\gamma|$ and $|\rho|$, for an $\alpha = .05$ Level Test to Ensure $\pi = .80$ Power*

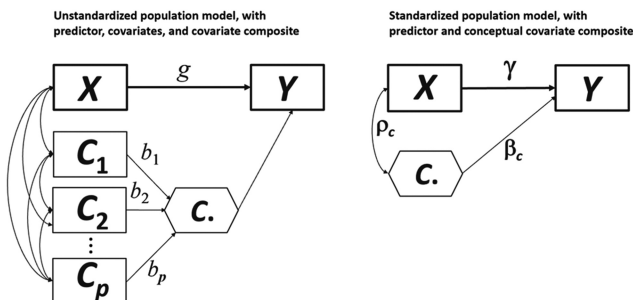
$ \gamma $	$ \rho $									
	$ \rho \leq 0.9$	$ \rho \leq 0.8$	$ \rho \leq 0.7$	$ \rho \leq 0.6$	$ \rho \leq 0.5$	$ \rho \leq 0.4$	$ \rho \leq 0.3$	$ \rho \leq 0.2$	$ \rho \leq 0.1$	$\rho = 0$
0.1	4,129	2,178	1,537	1,224	1,044	932	860	815	790	783
0.2	1,030	543	382	304	259	231	213	202	196	194
0.3	457	240	169	134	114	101	93	88	86	85
0.4	256	134	94	74	63	56	51	49	47	47
0.5	163	85	59	47	39	35	32	30	29	29
0.6	112	58	40	31	26	23	21	20	19	19
0.7	82	42	29	22	19	16	15	14	13	13
0.8	62	31	21	16	14	12	10	10	9	9
0.9	48	24	16	12	10	8	7	7	6	6

expressed within an unstandardized path analytic framework, as seen on the left in Figure 1, with a technically unnecessary but nonetheless illustrative hexagon-enclosed composite C , represented as an endogenous phantom factor.⁸ This model is equivalent in all respects to the original model without the extraneous C , most importantly with regard to power and sample size associated with all focal and contextual parameters.

Now consider the model on the right, represented in standardized form, depicting the composite covariate C , and the focal predictor X . In this model, C , need not literally be computed externally from the original p covariates, but rather stands symbolically as a proxy for all relevant contextual information contained in C_1 through C_p . Here, the parameter ρ_C is the Pearson correlation between X and C , while β_C is the standardized partial slope relating the composite C to the outcome Y above and beyond X .⁹ These parameters are herein referred to as contextual *metaparameters* (specifically, the *metacollinearity* parameter ρ_C and the *metaslope* parameter β_C). Also worth noting is that, while ρ_C represents a multivariate relation between X and the covariates C_1 through C_p , it is not the same as their population multiple correlation $P_{X.1 \dots p}$. In fact, because the latter is the Pearson correlation between X and a composite of C_1 through C_p optimized to predict X , whereas C is a composite of C_1 through C_p optimized to predict Y (above and beyond X), the magnitude of the metacollinearity cannot exceed that of the multiple correlation:

$$|\rho_C| \leq P_{X.1 \dots p}. \quad (7)$$

We will return to this point later.

Figure 1*Multiple Linear Regression Model With Covariate Composite*

To reiterate the issue above, the collapsed model is not intended to be an analytical model per se; rather, it will serve as a vehicle for sample size planning with respect to the standardized focal parameter γ . As derived in Appendix B, with ML estimation, the LR test of the focal parameter γ in the original (uncollapsed) model with covariates C_1 through C_p is strictly equivalent to that within the collapsed model with C as the proxy covariate. The important and unique implication of this equivalence is this: if we wish to conduct sample size planning for testing the focal parameter γ using the methods developed here, we only need values for the two contextual metaparameters (ρ_C and β_C) rather than for the $p(p+3)/2$ (standardized) contextual parameters in the uncollapsed model. In fact, as follows from the two-predictor case, the latter β_C will likewise automatically assume its conditional pessimum value, and as such will not require setting by the researcher. Thus, our contextual attention will be focused on the metacollinearity ρ_C .

Of course, as we should well expect by now, any misspecification in the contextual parameters, including in this meta form, can potentially lead to severely underpowered designs. Fortunately, given the equivalence of the collapsed form with respect to testing γ , simplified sample size planning for a general linear regression model can be obtained directly from the prior derivation as a relabeled version of the previous Equation 6 for the collapsed model:

$$n_{\max} = \frac{\lambda}{-\ln[1 - \gamma^2(1 - \rho_C^2)]} + 1. \quad (8)$$

Procedurally, this means that for testing the standardized parameter γ in the presence of p covariates (using an α -level test with target power π), the researcher (a) chooses the value for the SESOI γ and (b) sets a cautiously conservative threshold for the metacollinearity parameter ρ_C (while the value of β_C automatically assumes its conditional pessimum given γ and ρ_C).

From a practical standpoint, setting ρ_C may precipitate discomfort given that, instead of being the simple correlation between X and a single contextual covariate C , it is actually the Pearson correlation between X and the composite C of the p contextual covariates C_1 through C_p that has been optimized to predict Y . Note again that we

⁸ Although not shown in Figure 1, this would be parameterized in an unstandardized model with zero disturbance variance for C , and a path from C to Y set to 1.

⁹ Although we have no computational need for doing so here, both ρ_C and β_C can be derived algebraically or using standard path tracing rules.

are not, however, asking the researcher to make an educated guess as to the value of this more complex metaparameter; rather, as before, we are asking the researcher to set a cautiously conservative threshold (i.e., an upper limit) that one would not likely exceed. Still, unlike the two-variable case in which one needed merely to think in terms of a simple correlation between X and a single covariate C , the involvement of the composite C may now be leading us beyond the realm of reasonable intuition. Fortunately, to help ground us, recall that Equation 7 stated that the magnitude of the metacollinearity ρ_C is bounded by the multicollinearity between X and covariates C_1 through C_p , $P_{X.1 \dots p}$. This means that, although adding another layer of conservatism, setting an upper bound for the latter may be a useful practical guide to setting an upper bound for the former.

Continuing from the above, then, we could ask the following conservative but practical question: if X were to be predicted by the covariates C_1 through C_p , what is the largest proportion of explained variance that the researcher could reasonably expect to see in reality for these specific variables (i.e., that the researcher would be extremely surprised if exceeded)? The square root of this proportion is $P_{X.1 \dots p}$, which could be used as a conservative value to which to set the metacollinearity ρ_C . So, for example, for a researcher interested in predicting sixth-grade math achievement scores (Y) from fifth-grade math achievement scores (X), while controlling for $p = 5$ socioeconomic status (SES) measures (e.g., C_1 through C_5 include total family income, mother's educational level, father's educational level, mother's occupational prestige, and father's occupational prestige), we believe that researchers should be able to formulate and defend a statement of the sort, "We would expect fifth-grade math achievement scores to have a multiple correlation with the collection of SES measures no higher than .60." Once such a justifiable (and again, cautiously conservative) statement is made, sample size planning immediately becomes straightforward, and familiar.

For example, imagine that the above researcher decided that the standardized SESOI is $\gamma = .30$ (again, assuming this is a sufficiently conservative and reasonably justified value, and that this is a meaningful practical effect size that is worth investigation; see more details in the Discussion section), and that the worst-case metacollinearity is $\rho_C = .60$ (i.e., as informed by the multiple correlation per the sample statement above). For an $\alpha = .05$ level test to achieve a minimum of $\pi = .80$ power (which has noncentrality parameter $\lambda \approx 7.849$), under standard assumptions of conditional multivariate normality and independence of observations, it follows from Equation 8 that $n_{\max} = 134$ (technically, 133.30 rounded up). Thus, a sample size of 134 would yield at least $\pi = .80$ power to detect $\gamma = .30$ using an $\alpha = .05$ level test, as long as metacollinearity does not exceed $|\rho_C| = .60$. And in practice, given that this process assumes a worst-case scenario for ρ_C , and that β_C is at a corresponding pessimism, then the researcher would likely have more than .80 power with a sample size of n_{\max} given that the actual values of the contextual metaparameters are likely less pessimistic than planned for. As such, n_{\max} effectively serves as the insurance policy to guarantee statistical power.

In fact, we may think about this result even more generally. In the above example, there was one predictor X and $p = 5$ covariates. But any, or even all, of the covariates could have partial slope parameters relating to the outcome Y that are of theoretical interest for the purposes of sample size planning. In such a case, we can imagine all such focal predictors rotating through the position of X while all other predictors become pro tem covariates. This means that we

could ask a researcher for the SESOI γ across all focal parameters in the model, and a conservative metacollinearity threshold ρ_C expected to hold across all corresponding sets of contextual covariates. We would thus expect n_{\max} to provide at least π power for all α -level focal parameter tests (under standard assumed conditions), within a general linear model with any number of predictors, as long as no test's contextual metacollinearity exceeds the conservative researcher-specified level ρ_C .

Power for Measured Variable Path Analysis Models

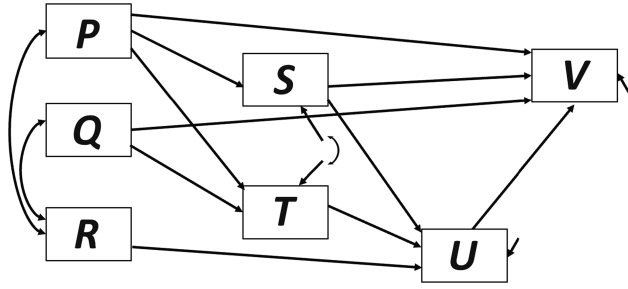
In the above section, we expanded to allow for any number of exogenous variables; in this section, we do the same for endogenous variables. This includes not just multivariate multiple regression models, where the same exogenous variables serve as predictors of each of multiple endogenous variables; here we extend the power analysis principles and practice to the general measured variable path analysis setting in which endogenous variables may be dependent upon both exogenous variables and other endogenous variables. Consider a generic such model in Figure 2, in which we assume for simplicity that all variables (labeled P through V) are continuous.¹⁰ In this model, variables P , Q , and R are exogenous, while variables S , T , U , and V are endogenous. To elaborate on the latter, $S = f_S(P)$ (i.e., S is a linear function of only one direct input variable, P), $T = f_T(P, Q)$, $U = f_U(R, S, T)$, and $V = f_V(P, Q, S, U)$. From the perspective of each endogenous variable, it is part of a microsystem within the larger model that consists of one or more predictors that may or may not covary, with their interrelations being a model-implied function of all legitimate path traces from one predictor to another (Wright, 1934). Said differently, for q endogenous variables, the larger measured variable path model can be thought of as a collection of q simple/multiple linear regression models spliced together. Indeed, in the early-to-mid-1900s, path models' parameters were often estimated in just such a parsed manner, using ordinary least squares methods (see, e.g., Pedhazur, 1982); since the advent of ML estimation, however, all model parameters are estimated concurrently as part of the model-fitting process.

In terms of sample size planning for measured variable path analysis models as a whole, the common SEM-based approaches as previously described involve the following steps: (1) assign values to all parameters, focal and contextual, to define the population; (2) set the α level for significance tests (e.g., .05) and the target power level π (e.g., .80); (3) estimate the n needed for testing each individual focal parameter by simulation or population analysis methods; (4) choose the largest such n needed across all focal parameters. Within this general process, Steps (2) through (4) are straightforward; Step (1), however, as we have argued previously, is essentially impossible with any degree of certainty. The broader implication here is that, because model fitting and estimation are approached at the model level with ML, improperly assumed parameter values in one location can (depending on the model structure) have repercussions for sample size planning for focal parameters elsewhere in the model.

Fortunately, for the purposes of sample size planning in the manner advocated in this article, we may return to the classical

¹⁰ Exogenous variables could also be dichotomous; for simplicity here, and to make the standardized example more easily interpreted, we are assuming all variables are continuous.

Figure 2
Generic Measured Variable Path Model

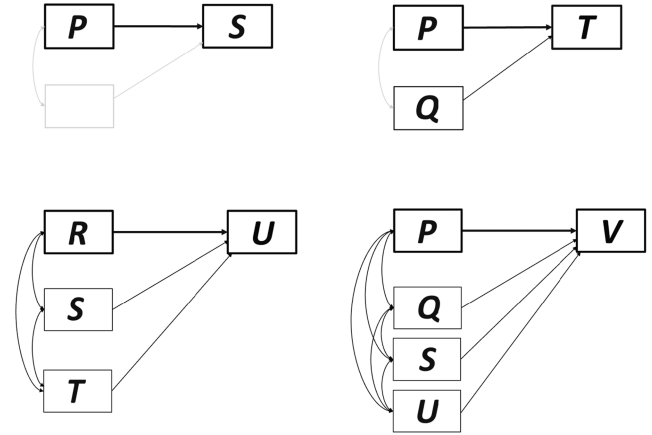


perspective in which we think of the measured variable path model with q endogenous variables as being parsable into q separate linear regression submodels. Now, in general, with ML estimation, these submodels and the original (unparsed) model do not necessarily yield the same ML results with respect to tests of parameters. This is because the fit function optimization occurs across the space spanned by all variables in a given analysis, whereas the submodels require optimization across a space spanned by fewer variables than the full model. However, the predictors in each of the q submodels can be considered locally exogenous, with their covariation being a function of all legally traceable relations between them within and beyond their submodel. As such, regardless of the specific values yielded by such tracing, the population covariation among these locally exogenous predictors cannot be worse (from the perspective of statistical power for testing γ) than the previously derived pessimism. Therefore, the conditional pessimism within each of the q submodels insulates against the other $q - 1$ submodels, as long as the larger unparsed model is correct. This model *parsability* (along with some practical points of elaboration addressed below) is what enables application of the sample size planning methods here to these more general measured variable path analysis models.

Imagine in Figure 2, for example, that a researcher considered the path from R to U to be focal. This path occurs within the submodel for endogenous variable U , which would have related predictors R , S , and T , and is shown in the bottom left of Figure 3. Thus, variables S and T constitute the *pro tem* context, with relevant contextual parameters including (in a standardized metric) the correlation between S and T , both of their correlations with R , and both of their partial slopes relating to U .¹¹ Also shown in Figure 3 are the remaining submodels for endogenous variables S , T , and V , for which similar focal/contextual parameter examples could be presented. In fact, every possible focal parameter from the original model in Figure 2 exists within a submodel in Figure 3, complete with predictor and outcome and contextual variables as relevant (i.e., none in the model for S , one in the model for T , two in the model for U , and three in the model for V). In each submodel, one of the possible focal parameters is represented at the top of the model with a thick, horizontal arrow.¹²

Now that the theoretical model has been conceptually parsed into its q submodels, for the purposes of sample size planning, we may draw upon the previous notion of collapsibility. Specifically, the submodels for endogenous variables U and V may have their contextual aspects collapsed further, as shown in Figure 4 (along with submodels for endogenous variables S and T). In each case, an exemplar standardized focal parameter γ and contextual metaparameters ρ_C and

Figure 3
Submodels From Parsing the Model in Figure 2



β_C are shown (grayed out as appropriate for endogenous variables S and T). Within each of these submodels, we know from the previous section of the article that for the SESOI γ across all focal parameters, and for a metacollinearity threshold ρ_C expected to hold across all corresponding sets of contextual covariates, there exists an n_{\max} value that provides at least π power for α -level tests of all focal parameters in that submodel (under standard assumed conditions).

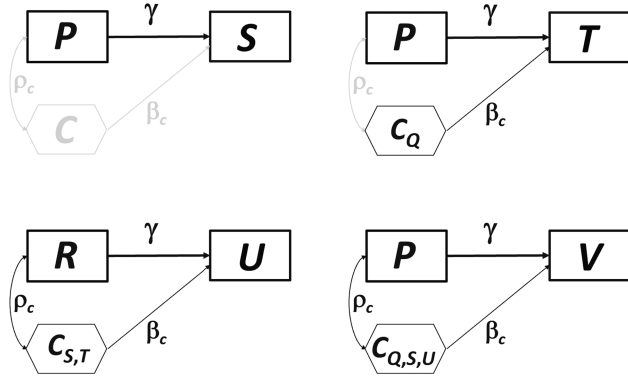
Based on the above, a researcher could therefore proceed in a piecemeal manner, determining an n_{\max} value for each submodel and then planning the study based on the largest such value across all submodels. However, in a measured variable path analysis model, a researcher might proceed even more generally. For sample size planning purposes, every structural focal parameter and its submodel context can be conceptually collapsed down to the same generic proxy model (as seen in the identical structures within Figure 4). This means that a researcher could go through the following steps at the level of the entire model (which we will illustrate in a tutorial example in the next section):

1. Choose the smallest (standardized) effect γ of interest for any focal structural path anywhere in the model.
2. Set the α level for the significance tests (e.g., .05) and the target power level π (e.g., .80), and find the corresponding noncentrality parameter (e.g., $\lambda \approx 7.849$ for $\alpha = .05$, $\pi = .80$).
3. Choose and justify a worst-case value for the contextual metacollinearity ρ_C across all submodel contexts.
4. Insert the values of γ , λ , and ρ_C into Equation 8 and round up to get n_{\max} .

¹¹ And of course, as a reminder, the roles of predictor and contextual variables could change within a submodel, depending how many of its parameters are considered focal.

¹² In Figure 3, the model with endogenous variable S in the top left corner shows a grayed-out contextual variable because no additional predictor exists in the original model in Figure 2. In the model for endogenous variable T in the top right corner, the correlation between P and Q is grayed out as no source of relation exists between P and Q in the original model in Figure 2. These special cases do not limit the generalizability of the proposed methods to come, but rather were chosen deliberately to illustrate the methods' versatility.

Figure 4
Collapsed Submodels From Parsing the Model in Figure 2



As stated previously, given that this process in practice takes a conservative worst-case scenario approach to setting ρ_c , and that β_c is automatically set at a corresponding pessimism, the researcher would likely have more than the target level of power with a sample size of n_{\max} , given that the actual values of the contextual meta-parameters are likely less pessimistic (e.g., the largest metacollinearity across all the submodels is likely less than the upper limit specified by the researchers). In fact—and this is the real punch line here—under standard assumed conditions, n_{\max} would be expected to provide at least π -level power for α -level testing of any structural parameter within any measured variable path model (i.e., with any number of variables, and any relations among them), as long as the test's associated metacollinearity does not exceed the conservative, researcher-specified level. The above expectation, of course, follows from the derivation that is based on asymptotic behavior; a demonstration of the behavior of n_{\max} in practice, as well as some important practical caveats, follows in the next section.

Illustrative Example

To illustrate how to obtain the n_{\max} estimate when planning for a measured variable path analysis, we use the context of the generic model in Figure 2, with four structural paths being of primary focal interest: γ_{VP} (i.e., $P \rightarrow V$), γ_{TQ} (i.e., $Q \rightarrow T$), γ_{UT} (i.e., $T \rightarrow U$), and γ_{VU} (i.e., $U \rightarrow V$).

Step 1: One could elicit standardized SESOIs for each of these, say:

$$\gamma_{VP} = .40, \gamma_{TQ} = .30, \gamma_{UT} = -.30, \text{ and } \gamma_{VU} = -.20, \quad (9)$$

again assuming each value to be, as per the spirit of SESOIs, sufficiently conservative, reasonably justified, and practically important. In this case, however, only the smallest SESOI must be formally articulated. In fact, its specific location within the model need not even be designated, merely that the SESOI anywhere is of absolute magnitude γ . So, imagine a researcher chooses $|\gamma| = .20$ as the (standardized) SESOI for the entire model.

Step 2: Now suppose the researcher plans to conduct the focal parameter statistical test at the $\alpha = .05$ level and targeting

$\pi = .80$ power. The corresponding noncentrality parameter for the $1 \text{ df } \chi^2$ test may be looked up in an existing table (e.g., Johnson et al., 1995) or computed using, for example, the `chi2ncp` function in the `gtx` R package (Johnson, 2020). With either approach, for $\alpha = .05$ and $\pi = .80$, the noncentrality parameter $\lambda \approx 7.849$.

Step 3: Next, a conservative metacollinearity level for the model (specifically, the highest upper threshold across all focal parameters' submodel contexts) is selected. In this example, assume the researcher sets an expected upper bound to metacollinearity of $\rho_c = .70$ (more information on such values will be provided later in the article).

Step 4: Using Equation 8 and making standard assumptions of conditional multivariate normality and independence of observations, we can obtain:

$$n_{\max} = \frac{7.849}{-\ln[1 - 0.2^2(1 - 0.7^2)]} + 1 = 381.8169. \quad (10)$$

Therefore, rounding n_{\max} up to the next integer to ensure sufficient power, the necessary sample size planned for this analysis would be $n_{\max} = 382$. Thus, based on all that has been presented so far, under standard assumed conditions, a sample size of 382 would be expected to yield at least $\pi = .80$ power to detect $|\gamma| = .20$ anywhere within this model, as long as the associated metacollinearity does not exceed $|\rho_c| = .70$.

Some Fine Print for the Case of Measured Variable Path Analysis Models

The metacollinearity parameter ρ_c upon which the proposed methods are based relies, in part, on the bivariate relations between a focal predictor X and each covariate (C_1, \dots, C_p) as well as among the covariates themselves. In the two-predictor and then general multiple regression cases addressed in Appendices A and B, these bivariate relations are each only a function of themselves. As such, the power and sample size for testing the focal parameter γ , which depends in part on the expected sampling behavior of ρ_c , more foundationally depends on the expected sampling variability in those exogenous variables' bivariate relations. In sample size planning for focal parameters within a measured variable path analysis model, however, where a model with q endogenous variables is treated herein as parsed into q simple/multiple regression submodels, bivariate relations between X and each covariate and among the covariates may have different model-implied origins: some might have been bivariate relations in the original unparsed model, but others might be a (multiplicative and/or additive) function of one or more directed and undirected relations from within that original model.¹³ Indeed, in Figure 2, consider the relation γ_{VU} (i.e., $U \rightarrow V$) as focal, in which case variables P , Q , and S become contextual covariates in the collapsed submodel upon which sample size planning is based with the methods in this article. In this example (and assuming a standardized model for simplicity), the bivariate relation between P and S in the collapsed submodel is

¹³ For example, in the path model shown in Figure 2, the model-implied total correlation between variables T and R is the sum of two products, involving four model parameters in all: $\beta_{TP}\rho_{PR} + \beta_{TQ}\rho_{QR}$.

merely a function of β_{SP} in the original model. On the other hand, the bivariate relation between U and P in the collapsed submodel is implied in the original model to be a more complex function of six free elements: $\beta_{SP}\beta_{US} + \beta_{TP}\beta_{UT} + \rho_{PR}\beta_{UR}$. This means that while the sampling behavior of the bivariate relation between P and S would be expected to be largely the same in the original model and any relevant submodels involving only P and S as the contextual covariates, with a finite sample, the behavior for the bivariate relation between U and P would be expected to have some divergence between the collapsed submodel and the original path model. Specifically, with a finite sample, one might expect the sampling behavior of the assumed simple bivariate relation between U and P to potentially underestimate the actual variability implied and propagated from the sampling behavior of the bivariate relations' six constituent elements within the original uncollapsed model. In short, this could translate into a power shortfall. Further examination of this issue follows in the next section.

Monte Carlo Simulation Illustration

To get a practical sense of the potential cost associated with the propagation issue described above, we conducted a modest simulation using the measured variable path analysis model in Figure 2 for purposes of illustration. Note that, given the mathematical derivations upon which the proposed methods rest, we did not intend this article to be a simulation study per se, nor did we feel a large study of that type, commonly crossing many conditions factorially, would be necessary. Indeed, given all the points of conservatism built into the methods proposed in this article, we expected the propagation issue to amount to very little, if anything, in practice, with whatever effects may exist likely being felt more for testing focal parameters further "downstream" in the model (i.e., where bivariate focal predictor and covariate relations are expected to be more complex functions of original model parameters).

In this simulation illustration, the structure of the data-generating model, which was the path model shown in Figure 2, was fixed across all simulation conditions. The factors that were systematically manipulated were: (a) designation of the focal parameter in the model (10 different structural paths), (b) the (standardized) SESOI for the focal parameter (three values), (c) the upper bound of the contextual metacollinearity ρ_C (seven levels), and (d) the (standardized) contextual parameter values (1,000 different sets), resulting in a maximum of $10 \times 3 \times 7 \times 1,000 = 210,000$ different possible scenarios for assessing the empirical power with n_{\max} . Within each permissible condition (i.e., where the combination of focal and contextual parameters yielded a PD covariance matrix), 1,000 sample replications were randomly generated and used for testing the focal parameter. The empirical power for each condition was calculated as the proportion of α -level statistically significant test results across all converged replications.

To elaborate, each of the 10 direct paths in this model in turn served as the focal parameter for hypothesis testing. The focal parameter γ was set to be .20, .30, and .40 in turn, and metacollinearity ρ_C values were examined ranging from .50 to .80 in increments of .05. Contextual parameters were randomly generated from uniform distributions (with the only exception being residual variances, which were fixed deterministically based on path coefficients values in order to yield unit variances for all variables). The simulation results for an upper bound $\rho_C = .70$ will be presented in more detail below.¹⁴

As n_{\max} only depends on ρ_C and γ (Equation 8), the corresponding planned sample size n_{\max} can be determined to be 382, 169, and 94, respectively, for each standardized SESOI focal value of γ , for an $\alpha = .05$ level test to provide $\pi = .80$ power (under standard assumed conditions). These sample sizes will thus be used to assess the empirical power through simulations across a wide range of randomly varying contextual conditions. More specifically, for each focal parameter and a chosen γ value, 1,000 sets of contextual parameter values were randomly generated, thus defining 30,000 unique population conditions in total given the upper bound $\rho_C = .70$: 10 possible focal parameters \times 3 γ values \times 1,000 contextual parameter conditions. Of these, a total of 24,322 produced permissible values and PD population covariance matrices, with the resulting actual population metacollinearity ρ_C ranging from $-.84$ to $.87$ (i.e., it was allowed to exceed, that is, be even worse than, the planned ρ_C level as resulting population parameter conditions dictated). Within each of these unique permissible population model conditions, empirical power for testing the focal parameter with n_{\max} was assessed across 1,000 replications of random data generation from that population condition with the original model in Figure 2 fitted each time. Empirical power was obtained by calculating the proportion of statistically significant ($p < .05$) test results from the convergent models among the 1,000 replications; this value was derived for all 24,322 permissible population conditions. It is important to note that it is possible for an empirical power estimate to dip below π (in this case, .80) for four reasons: (a) random Monte Carlo error, (b) finite sampling behavior deviating from the asymptotics assumed in the mathematical derivations upon which n_{\max} was based, (c) selected values of ρ_C being further from the asymptotic worst-case scenario where $|\rho_C|$ approaches 1.00 (see Appendix A), and (d) randomly generated population conditions where the resulting metacollinearity was allowed to be in excess of the chosen level (thus yielding no necessary expectation of sufficient power).

For reasons described below, simulation results specifically for the case of $\rho_C = .70$ are summarized in Table 3, which focus on two outcomes for each condition: the percentage of relevant population conditions with an empirical power of 0.80 or larger,¹⁵ and the first percentile point (P_{01}) in the distribution of that condition's empirical power estimates (to accommodate Monte Carlo error). The latter, in other words, is the empirical power value that is lower than 99% of the other empirical power estimates for testing the same focal parameter under different randomly drawn contextual parameter values. The results are summarized from three perspectives separately: (a) all simulated population conditions (with permissible values and a PD population covariance matrix), (b) population conditions in which the model-implied metacollinearity ρ_C does not exceed the chosen upper bound of .70, and (c) population conditions in which the multiple correlation between the focal predictor and the contextual

¹⁴ Full results for ρ_C ranging from 0.50 to 0.80 are available from the authors upon request.

¹⁵ This is analogous to the "assurance level" discussed by Du and Wang (2016), which was defined as the probability of achieving or exceeding a target power level given the sample size. The difference here is that the parameter values in our simulation were not drawn from a posterior distribution to form a posterior power distribution. The contextual parameter values were drawn from predefined uniform distributions in order to cover as many possible scenarios as possible to assess the robustness of n_{\max} under a wide range of conditions.

Table 3
Distribution of Empirical Power With n_{\max}

γ	n_{\max}	Focal par.	All PD		$ \rho_C \leq .70$		$ \rho_{X.1 \dots p} \leq .70$	
			% $\geq .80$	P_{01}	% $\geq .80$	P_{01}	% $\geq .80$	P_{01}
0.2	382	γ_{SP}	100.00%	0.966	100.00%	0.966	100.00%	0.966
		γ_{TP}	100.00%	0.970	100.00%	0.970	100.00%	0.970
		γ_{TQ}	100.00%	0.971	100.00%	0.971	100.00%	0.971
		γ_{UR}	100.00%	0.949	100.00%	0.949	100.00%	0.954
		γ_{US}	100.00%	0.950	100.00%	0.950	100.00%	0.950
		γ_{UT}	100.00%	0.937	100.00%	0.937	100.00%	0.945
		γ_{VP}	99.89%	0.858	100.00%	0.864	100.00%	0.865
		γ_{VQ}	100.00%	0.960	100.00%	0.960	100.00%	0.961
		γ_{VS}	99.64%	0.834	99.64%	0.834	100.00%	0.856
		γ_{VU}	99.88%	0.881	99.88%	0.881	100.00%	0.912
		γ_{SP}	100.00%	0.970	100.00%	0.970	100.00%	0.970
		γ_{TP}	100.00%	0.975	100.00%	0.975	100.00%	0.975
0.3	169	γ_{TQ}	100.00%	0.975	100.00%	0.975	100.00%	0.975
		γ_{UR}	100.00%	0.949	100.00%	0.949	100.00%	0.955
		γ_{US}	100.00%	0.948	100.00%	0.948	100.00%	0.948
		γ_{UT}	100.00%	0.933	100.00%	0.933	100.00%	0.950
		γ_{VP}	100.00%	0.882	100.00%	0.884	100.00%	0.886
		γ_{VQ}	100.00%	0.969	100.00%	0.969	100.00%	0.970
		γ_{VS}	99.40%	0.837	99.52%	0.845	100.00%	0.887
		γ_{VU}	99.36%	0.863	99.36%	0.863	100.00%	0.916
		γ_{SP}	100.00%	0.975	100.00%	0.975	100.00%	0.975
		γ_{TP}	100.00%	0.977	100.00%	0.977	100.00%	0.977
		γ_{TQ}	100.00%	0.978	100.00%	0.978	100.00%	0.978
		γ_{UR}	100.00%	0.952	100.00%	0.953	100.00%	0.960
0.4	94	γ_{US}	100.00%	0.965	100.00%	0.965	100.00%	0.965
		γ_{UT}	100.00%	0.928	100.00%	0.941	100.00%	0.947
		γ_{VP}	100.00%	0.874	100.00%	0.880	100.00%	0.895
		γ_{VQ}	100.00%	0.966	100.00%	0.971	100.00%	0.972
		γ_{VS}	99.74%	0.859	99.74%	0.859	100.00%	0.889
		γ_{VU}	99.12%	0.820	99.11%	0.819	100.00%	0.911

Note. Focal Par. = focal parameter; PD = positive definite; P_{01} = first percentile.

covariates ($P_{X.1 \dots p}$) does not exceed the chosen upper bound for ρ_C of .70 (and thus neither does ρ_C).

Results are both as expected and reassuring. To start, even allowing multicollinearity ρ_C to exceed the planned level of .70 (i.e., the “all PD” columns), as some randomly generated parameter conditions did, most scenarios still yield empirical power in excess of .80 in 100% of conditions; indeed, even the worst scenario had more than 99% of the cases with empirical power exceeding .80. Furthermore, all of the corresponding P_{01} values exceeded 0.80, with the lowest being 0.820. Lowest performance, as expected, was consistently associated with parameters involving V, that is, parameters involving this ultimate downstream variable, whose contextual variable bivariate relations are compound functions of multiple parameters. To reiterate, however, performance was generally outstanding, even with population parameter conditions whose implied multicollinearity ρ_C values were allowed to exceed the planned level of .70.

Restricting cases only to those with multicollinearity ρ_C at or below .70 (i.e., the middle pair of columns in Table 3), power performance stayed the same or improved slightly, as expected. Also interesting from a practical standpoint is what happens when we filter by cases whose population conditions have multicollinearity $P_{X.1 \dots p}$ at or below .70. Recall that this serves as a generally more conservative restriction than multicollinearity $\rho_C \leq .70$, and this conservatism is reflected in the performance in the final two columns of Table 3 as every single condition had empirical power at or above .80. In other

words, if a researcher had set the target multicollinearity using an expected worst-case level of multicollinearity of contextual covariates with a predictor, power would be ensured in all cases.

To reiterate what was stated at the start of this section, this is by no means intended to be an exhaustive simulation; however, we do believe it nicely illustrates the viability of the proposed n_{\max} approach. First, although only a single measured variable path analysis model was utilized, it was designed to accommodate parameters in a wide variety of structural and contextual locations. Focal parameter values were chosen in the small to medium range (γ from .20 to .40), with n_{\max} values from 94 to 382, and with contextual parameters spanning the full permissible range. Perhaps the most challenging practical issue is the selection of the multicollinearity level ρ_C ; this was varied from .50 to .80, finding that for the current illustration ρ_C values examined below .70 yielded power that occasionally dipped below .80, and ρ_C values examined exceeding .70 generally provided excessive power. While additional power is not a bad thing from a statistical standpoint, it can be unnecessarily costly in terms of sample size. For the circumstances examined here, we found $\rho_C = .70$ to be a surprisingly robust choice for multicollinearity in terms of providing the target level of power for parameters throughout the model, even under scenarios in which the chosen level of multicollinearity is violated. This means that a predictor X and the optimally predictive linear composite C , of the contextual variables C_1, \dots, C_p share around half of their variance (i.e., $0.70^2 = 0.49$). And while we do not wish to endorse $\rho_C = .70$ as a universal multicollinearity recommendation,

seeing the behavior of n_{\max} under this and neighboring conditions, in a structurally diverse measured variable path analysis model, appears to offer some useful numerical grounding. Furthermore, these results also suggest that, if a researcher preferred to conduct sample size planning from a multicollinearity rather than metacollinearity perspective, setting ρ_C based on a belief about the more conservative $P_{X.1 \dots p}$ appears to ensure adequate power, as long as $P_{X.1 \dots p}$ indeed does not exceed the stated level set for ρ_C .

Discussion

Planning quality research requires extensive a priori reflection regarding, for example, sufficient sample size and statistical power, the study's potential for contribution to its field, the research design, methods of sampling, variables to be included, the nature and quality of those variables' measurement, proper specification of models, and expectations regarding the parameters within those models. The current study has focused on sample size planning, which is at the core of a priori power analysis, with particular attention to model parameters' role within this process. To summarize, the models analyzed in research endeavors include not just the focal parameters that are directly tied to the primary research questions, but also the contextual parameters that surround them; sample size planning has historically required a priori information from the researcher about both. While the methodological literature offers many guidelines and strategies for setting focal parameters for sample size planning, especially within the traditional Cohen framework, setting values for the surrounding contextual parameters, particularly for more complex models, is far more difficult. Quite simply, as this article and others before it have argued, our models typically differ from prior studies in too many ways to be able to trust the values of their contextual parameters with any reasonable certainty (if such values were even reported in those prior studies at all). This means that when choosing such values, researchers are routinely forced to fill the gaps in their contextual knowledge with educated guesses, and maybe, however unintentionally, a bit of wishful thinking. Unfortunately, by whatever means those contextual values are arrived upon, the cost of their inevitable inaccuracy can be a woefully underpowered and possibly completely wasted research endeavor.

Thus, rather than relying on such tenuous contextual speculation and hoping for the best, the methods offered in the current work encourage planning for the worst. Toward that end, we have developed a simple, practical approach for conducting sample size planning for methods that fall within the scope of the measured variable path analysis framework. This domain explicitly subsumes a number of familiar analyses, including multiple linear regression and its derivatives (e.g., independent samples *t* test, analysis of variance), and should straightforwardly extend to a variety of specific measured variable modeling scenarios (e.g., SEM approaches to conditional process modeling; Hayes & Preacher, 2013).

This approach to sample size planning, as we believe is worth reiterating, has been built upon both statistical and ideological foundations. Statistically, and in reverse order from their appearance in the article, we encountered the notion of parsability. If you page through just about any applied journal nowadays, the models being presented and analyzed are increasingly sophisticated. In the social and behavioral sciences, and well beyond as well, we seldom ask questions that are answered by the test of a single parameter, but instead tend to

focus on larger systems where research questions are addressed by the interplay of many such parameters. Given these models' complexity, the case was made here that it can be useful to consider them in parts. Doing so is not meant to claim that the whole functions precisely as the sum of its parts, but rather that planning for each of those parts—especially the weakest—can be a very useful strategy when it comes to making sample size preparations for the model as a whole.

In the measured variable path analysis models addressed here, those parts are general linear regression submodels, each having a single endogenous outcome and as many predictors as variables leading directly into that outcome. These submodels can differ in the number of predictors as well as the number of focal paths, seemingly undermining parsing's promise of simplification. But here we may invoke the second statistical concept encountered in the article, that of collapsibility: for each focal parameter within a submodel, its submodel collapses into the predictor, the outcome, and a linear composite of any and all remaining predictors to serve as a contextual proxy covariate. To be clear, we don't mean to imply that all such collapsed submodels are the same; rather, from a planning perspective, if you can plan for the weakest such collapsed submodel, then you should be prepared for the entire model.

Sample size planning for such a collapsed submodel still requires inputs from the researcher. Our concern with parameter inputs, as echoed throughout this article, is that prior knowledge is almost always of questionable accuracy. And thus we invoked another important notion, the pessimism. This is the worst-case context surrounding a given effect size of interest, where that context depends on the metacollinearity parameter (ρ_C) as specified by the researcher, as well as the metaslope for the composite covariate (β_C), which is automatically assumed to be in the least favorable configuration.

Although the pessimism is derived mathematically and serves as one of the three statistical pillars of this sample size planning approach, it also directly embodies the ideological foundation embraced here, that of planning for the worst. Imagine how different, say, a grant proposal would be that explicitly adopted this perspective. Rather than fumbling through a narrative that is benign in its fiction under the best of circumstances, a researcher could instead—and with complete sincerity—include something as simple as the following:

The proposed investigation will utilize a sample size of $n = 382$, which (under standard assumed conditions) is expected to provide a minimum of .80 power (using .05-level tests) to detect standardized $X \rightarrow Y$ direct effects of absolute magnitude as small as .20 anywhere in the theoretical model, when other concurrent predictors of a given endogenous variable Y have metacollinearity with the given X variable of up to .70.¹⁶

The proposed methods make possible such simplicity.

Although the focus of this current study has been primarily on managing uncertainty in the most challenging area, the contextual parameters, the proposed methods still rest on the ability of the researcher to specify a meaningful value for the focal parameter embedded within that context. By meaningful value, however, we do not mean a true value. Attempts at accurate estimation of focal parameters, as much methodological research has discussed,

¹⁶ The researcher would still need to defend the specified level of metacollinearity.

frequently leads to underpowered studies (Anderson, 2019; Anderson et al., 2017; McShane & Böckenholt, 2016; Perugini et al., 2014), and both frequentist and Bayesian approaches have been offered for remediation and strategic conservatism in such estimation and planning (e.g., Anderson & Maxwell, 2017; Anderson et al., 2017; Du & Wang, 2016; McShane & Böckenholt, 2016; Pek & Park, 2019; Perugini et al., 2014; Spiegelhalter & Freedman, 1986). Rather, and importantly, with regard to focal parameter specification, we have assumed an SESOI perspective (e.g., Lakens, 2022; Lakens et al., 2018), which requires the researcher to speculate as to what constitutes “meaningful” for theoretical and applied stakeholders and then operationalize that speculation within the focal parameters themselves. In subscribing to this perspective, anything smaller than the specified SESOI is deemed too small to be of inferential interest, or at least too small to commit the necessary resources toward such inference, but this SESOI requires no less justification than other approaches to focal parameter specification (albeit possibly using different criteria for that justification). Irrespective of how one justifies their SESOI, objectively and/or subjectively (for more details, see Lakens, 2022; Lakens et al., 2018), the key contribution of the approach proposed in the current article is this: if the true focal parameter is at least as large as the SESOI justified by the researcher, n_{\max} ensures (under standard distributional assumptions) at least the target level of power for the statistical test of that parameter.

With regard to our proposed methods, we do note that in our simulation illustration, the empirical statistical power yielded by n_{\max} was higher than the target power level ($\pi = .80$) in most cases, suggesting it may still be able to accommodate some deviation in the focal parameter from its designated value. As such, this approach may still achieve the target power even when the SESOI selected and justified is somewhat higher than actual minimum thresholds of practical interest and importance. But to what extent our methods are robust under increasing deviations from the assumed population value remains for future investigation. Nevertheless, the conservatism inherent to n_{\max} is expected to at least partially offset potential loss of power when the conditions are less optimistic, whether by virtue of a smaller focal parameter or a higher metacollinearity than specified.

Returning to dealing with uncertainty in the contextual parameters, other approaches also exist. Cole et al. (2025), for example, recently proposed the plausible values for secondary parameters approach, where a plausible range for each nonfocal model parameter is explicitly defined. Researchers with solid prior knowledge, for instance, may determine that one of the contextual paths is no lower than 0.1 and no higher than 0.5, that one of the residual covariances is no less than 0.05 and no larger than 0.3, and so on, establishing lower and upper bounds for each contextual parameter. Based on this information, a full factorial Monte Carlo simulation is implemented by examining every combination of these boundary values, generating a collection of power estimates under varying population conditions, which in turn can inform sample size planning. Of course, this approach does rely on researchers providing reasonable upper and lower bounds for parameters within potentially complex systems, where the combinations of those bounds must themselves define feasible populations (e.g., with PD covariance matrices). Furthermore, as those systems become increasingly complex, the number of combinations can become rather prohibitive. In the relatively modest model in Figure 2, for example, and assuming a standardized model for simplicity, for a single focal parameter, there are 12 contextual structural

and nonstructural parameters for which to specify lower and upper bounds, leading to a sizeable $2^{12} = 4,096$ simulation conditions. And there is no guarantee that the worst-case scenario, whether in a simple or more complex model, will occur at the intersection of bounds rather than somewhere in between.

As another option for dealing with uncertainty, a sequential design could be used to help refine knowledge of focal and contextual parameters. Although, as noted previously, a standalone external pilot study has many potential limitations (see also Browne, 1995), an internal pilot study could be employed whereby researchers use parameter estimates from data collected in an initial pilot phase to update the nuisance parameters and inform sample size planning in subsequent phases. This approach could, however, introduce endogenous bias in more complex models that involve multiple focal parameters, given that each focal parameter also serves as a contextual parameter in testing a different focal parameter (see, e.g., Wittes & Brittain, 1990).

In the end, the methods we propose in this article resonate with prior studies’ concerns about uncertainty in setting model parameters for sample size planning, offering a framework guided by what we believe to be appropriate conservatism regarding focal parameters implicitly (through subscribing to an SESOI perspective) and contextual parameters explicitly (through pessimistic metaparameters). Certainly, more work (some relatively simple and some more complex) remains to be done, beyond the scope of the foundations we intended to lay in the current article. For example, although n_{\max} often suggests a sample size that could be considered “large,” Table 2 shows that this is not always the case. Indeed, some of the sample sizes shown could present challenges within SEM, including biased estimates, model nonconvergence, inadmissible solutions, and unstable parameter estimates and model fit indices, particularly when unconstrained traditional ML estimation is used (Bentler & Yuan, 1999; Ulitzsch et al., 2023; Wolf et al., 2013). Therefore, other key factors for sample size planning, including the stability and interpretability of the model solution, desirable distributional properties of the data, and an adequate estimation precision level (Pek et al., 2024; Wolf et al., 2013), may need to be considered as well.

In addition, sample size planning methods were not explicitly illustrated here for models in which the unstandardized parameter values are particularly important, nor for testing parameters that are explicit functions of other parameters (e.g., indirect and total effects). Furthermore, we did not address models with categorical outcomes, multiple groups, mean structures, latent variables, and many other variations and extensions within and beyond the SEM framework. We also presented these methods with the assumption of complete data and standard distributional conditions (as is quite customary in the methodological power analysis literature); understanding these methods’ performance and robustness under other conditions, including extensions to more complex (e.g., multilevel) data structures, would be both interesting and necessary.

Finally, having been a part of countless sample size planning dialogs ourselves, we are quite familiar with the retort, “But I can only afford ____ subjects.” Whatever number fills in the blank, it represents the reality of resource and budget constraints experienced all too commonly by applied researchers. Indeed, such constraints are often the primary cause of the haggling and sambaing to which we have also borne frequent witness. To this, we offer three responses. First, and admittedly least empathically, we start with a reminder that the truths of the universe are unmoved by haggling

and sambaing, however artful they may be. That is, the knobs one twiddles in setting parameter values have no wires leading to the population parameters themselves. This means that a genuinely earnest attempt at sample size planning that yields a value for n that one cannot afford should stop, or at least pause, a study from moving forward. Otherwise, one is proceeding with an investigation that has an often severely reduced probability of detecting the relations of key theoretical interest: 0.70, 0.60, 0.50, or worse. Does one really wish to conduct a study whose key conclusions are effectively the toss of a coin?

Second, perhaps a more compassionate perspective is one not framed as haggling or sambaing, but rather captured in another common response: “I only have ____ amount of money. What can I get for it?” Here the researcher still presumably wishes to maintain a target level of power, but is inquiring as to the various corresponding settings of the parameter knobs. For instance, we translate that question as, “If n_{\max} were limited to 200 subjects, to what values of the focal parameter (γ) and metacollinearity (ρ_C) would that correspond?” Assuming an $\alpha = .05$ level test and a target power of $\pi = .80$ (for which the noncentrality parameter is $\lambda \approx 7.849$), this is tantamount to articulating an isopower contour in which:

$$n_{\max} = 200 = \frac{7.849}{-\ln[1 - \gamma^2(1 - \rho_C^2)]} + 1, \quad (11)$$

which could be approximately satisfied by an infinite number of combinations, including ($\gamma = .20$, $\rho_C = .20$) or ($\gamma = .45$, $\rho_C = .90$). While perhaps informative, we do worry that this kind of reverse-engineering will reverse us right back to the kind of negotiations from which current sample size practices suffer, and which n_{\max} was aiming to help avoid. On the other hand, doing so could also serve as a bit of a sensitivity analysis, providing reasonable encouragement should the combinations of focal parameter and metacollinearity values be in defensibly meaningful ranges.

Third, and transcending issues regarding the current methods, other ways exist to maximize information on a fixed budget. Planned missing data designs, for example, can allow for an increased number of subjects overall by lowering the cost per individual participant through strategically limiting the number of variables gathered from each subject (see, e.g., Feng & Hancock, 2021; Rhemtulla & Hancock, 2016; Rhemtulla et al., 2016). On the other hand, perhaps power does not even constitute the proper target; in specific settings, for example, maybe one should be determining sample size for the purpose of achieving a defensible level of precision in the estimation of the focal parameter (e.g., Kelley & Maxwell, 2003; Lai & Kelley, 2011). How these issues, and those already mentioned above within this section, can work hand in hand with n_{\max} remains to be explored. And thus we consider the current work to be a useful start, helping to ensure that researchers’ sample size planning is rooted in caution (through strategic pessimism), integrity (through the formal inculcation of that pessimism into n_{\max} to help avoid researchers’ motivated manipulations), and practicality (in the simplicity of the methods we propose). And in utilizing these methods, then, researchers are putting into action the beliefs that (a) studies worth doing are worth erring on the side of caution in planning, and that (b) admitting uncertainty in our prior knowledge, and being willing to pay to ensure against its potentially severe consequences, is a reasonable investment in a better and more replicable science.

References

- Anderson, S. F. (2019). Best (but oft forgotten) practices: Sample size planning for powerful studies. *The American Journal of Clinical Nutrition*, 110(2), 280–295. <https://doi.org/10.1093/ajcn/nqz058>
- Anderson, S. F., & Maxwell, S. E. (2017). Addressing the “replication crisis”: Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, 52(3), 305–324. <https://doi.org/10.1080/00273171.2017.1289361>
- Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, 28(11), 1547–1562. <https://doi.org/10.1177/0956797617723724>
- Bentler, P. M., & Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, 34(2), 181–197. <https://doi.org/10.1207/S15327906Mb340203>
- Browne, R. H. (1995). On the use of a pilot sample for sample size determination. *Statistics in Medicine*, 14(17), 1933–1940. <https://doi.org/10.1002/sim.4780141709>
- Chattopadhyay, B., Bandyopadhyay, T., Kelley, K., & Padalunkal, J. J. (2025). A sequential approach for noninferiority or equivalence of a linear contrast under cost constraints. *Psychological Methods*, 30(2), 425–439. <https://doi.org/10.1037/met0000570>
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (revised ed.). Academic Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.
- Cole, D. A., Abitante, G., Kan, H., Liu, Q., Preacher, K. J., & Maxwell, S. E. (2025). Practical problems estimating and reporting power when hypotheses are embedded in complex statistical models. *Advances in Methods and Practices in Psychological Science*, 8(1), Article 25152459241302300. <https://doi.org/10.1177/25152459241302300>
- Donnelly, S., Jorgensen, T. D., & Rudolph, C. W. (2023). Power analysis for conditional indirect effects: A tutorial for conducting Monte Carlo simulations with categorical exogenous variables. *Behavior Research Methods*, 55(7), 3892–3909. <https://doi.org/10.3758/s13428-022-01996-0>
- Du, H., & Wang, L. (2016). A Bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivariate Behavioral Research*, 51(5), 589–605. <https://doi.org/10.1080/00273171.2016.1191324>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Feng, Y., & Hancock, G. R. (2021). Oh no! They cut my funding! Using “post hoc” planned missing data designs to salvage longitudinal research. *Child Development*, 92(3), 1199–1216. <https://doi.org/10.1111/cdev.13501>
- Feng, Y., & Hancock, G. R. (2023). SEM as a framework for power analysis. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed., pp. 163–183). Guilford Press.
- Fisher, R. A. (1938). Presidential address to the First Indian Statistical Congress. *Sankhya*, 4(1), 14–17. <https://www.jstor.org/stable/40383882>
- Graybill, F. A. (1983). *Matrices with applications in statistics*. Wadsworth Publishing Company.
- Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and mimic approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, 66(3), 373–388. <https://doi.org/10.1007/BF02294440>
- Hancock, G. R., & French, B. F. (2013). Power analysis in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 117–159). IAP Information Age Publishing.
- Hayes, A. F., & Preacher, K. J. (2013). Conditional process modeling: Using structural equation modeling to examine contingent causal processes.

- In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 219–266). IAP Information Age Publishing.
- Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, 6(3), 203–217. <https://doi.org/10.1037/1082-989X.6.3.203>
- Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, 9(4), 426–445. <https://doi.org/10.1037/1082-989X.9.4.426>
- Hoening, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1), 19–24. <https://doi.org/10.1198/000313001300339897>
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (Vol. 2, 2nd ed.). Wiley.
- Johnson, T. (2020). *gtx: R package for genetic data analysis* (Version 2.1.6) [Computer software]. <https://github.com/tobyjohnson/gtx>
- Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, ANOVA, and beyond* (3rd ed.). Routledge. <https://doi.org/10.4324/9781315744131>
- Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, 8(3), 305–321. <https://doi.org/10.1037/1082-989X.8.3.305>
- Kraemer, H. C., & Blasey, C. (2015). *How many subjects?: Statistical power analysis in research* (2nd ed.). Sage Publications.
- Lai, K., & Kelley, K. (2011). Accuracy in parameter estimation for targeted effects in structural equation modeling: Sample size planning for narrow confidence intervals. *Psychological Methods*, 16(2), 127–148. <https://doi.org/10.1037/a0021764>
- Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, 8(1), Article 33267. <https://doi.org/10.1525/collabra.33267>
- 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>
- Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology*, 97(5), 951–966. <https://doi.org/10.1037/a0028380>
- 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>
- McShane, B. B., & Böckenholt, U. (2016). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods*, 21(1), 47–60. <https://doi.org/10.1037/met0000036>
- Moerbeek, M. (2022). Power analysis of longitudinal studies with piecewise linear growth and attrition. *Behavior Research Methods*, 54(6), 2939–2948. <https://doi.org/10.3758/s13428-022-01791-x>
- Moshagen, M., & Bader, M. (2024). Sempower: General power analysis for structural equation models. *Behavior Research Methods*, 56(4), 2901–2922. <https://doi.org/10.3758/s13428-023-02254-7>
- Muthén, L. K., & Muthén, B. O. (1998–2023). *Mplus user's guide* (8th ed.). 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: A Multidisciplinary Journal*, 9(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8
- Noble, B. (1969). *Applied linear algebra*. Prentice-Hall.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). Holt, Rinehart, & Winston.
- Pek, J., & Park, J. (2019). Complexities in power analysis: Quantifying uncertainties with a Bayesian-classical hybrid approach. *Psychological Methods*, 24(5), 590–605. <https://doi.org/10.1037/met0000208>
- Pek, J., Pitt, M. A., & Wegener, D. T. (2024). Uncertainty limits the use of power analysis. *Journal of Experimental Psychology: General*, 153(4), 1139–1151. <https://doi.org/10.1037/xge0001273>
- Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, 9(3), 319–332. <https://doi.org/10.1177/1745691614528519>
- Rhemtulla, M., & Hancock, G. R. (2016). Planned missing data designs in educational psychology research. *Educational Psychologist*, 51(3–4), 305–316. <https://doi.org/10.1080/00461520.2016.1208094>
- Rhemtulla, M., Savalei, V., & Little, T. D. (2016). On the asymptotic relative efficiency of planned missingness designs. *Psychometrika*, 81(1), 60–89. <https://doi.org/10.1007/s11336-014-9422-0>
- Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, 50(1), 83–90. <https://doi.org/10.1007/BF02294150>
- Schulz, K. F., & Grimes, D. A. (2005). Sample size calculations in randomised trials: Mandatory and mystical. *The Lancet*, 365(9467), 1348–1353. [https://doi.org/10.1016/S0140-6736\(05\)61034-3](https://doi.org/10.1016/S0140-6736(05)61034-3)
- Spiegelhalter, D. J., & Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, 5(1), 1–13. <https://doi.org/10.1002/sim.4780050103>
- Thoemmes, F., MacKinnon, D. P., & Reiser, M. R. (2010). Power analysis for complex mediational designs using Monte Carlo methods. *Structural Equation Modeling: A Multidisciplinary Journal*, 17(3), 510–534. <https://doi.org/10.1080/10705511.2010.489379>
- Tu, X. M., Kowalski, J., Zhang, J., Lynch, K. G., & Crits-Christoph, P. (2004). Power analyses for longitudinal trials and other clustered designs. *Statistics in Medicine*, 23(18), 2799–2815. <https://doi.org/10.1002/sim.1869>
- Ulitzsch, E., Lüdtke, O., & Robitzsch, A. (2023). Alleviating estimation problems in small sample structural equation modeling—A comparison of constrained maximum likelihood, Bayesian estimation, and fixed reliability approaches. *Psychological Methods*, 28(3), 527–557. <https://doi.org/10.1037/met0000435>
- Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9(1–2), 65–72. <https://doi.org/10.1002/sim.4780090113>
- Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, 73(6), 913–934. <https://doi.org/10.1177/0013164413495237>
- Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3), 161–215. <https://doi.org/10.1214/aoms/1177732676>
- Yuan, K. H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, 30(2), 141–167. <https://doi.org/10.3102/10769986030002141>
- Zhang, Z. (2014). Monte Carlo based statistical power analysis for mediation models: Methods and software. *Behavior Research Methods*, 46(4), 1184–1198. <https://doi.org/10.3758/s13428-013-0424-0>

(Appendices follow)

Appendix A

Deriving the Conditional Pessimism

In this appendix, we derive the global minimum of the maximum likelihood (ML) fit function F_{ML} . We start with the proof that the power for testing the focal parameter in multiple linear regression is minimized with perfect collinearity $|\rho| \rightarrow 1$.

Suppose we have the population regression model (with mean-centered variables¹⁷ to eliminate the need for a mean structure here, for simplicity):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\Gamma} + \boldsymbol{\varepsilon}, \quad (\text{A1})$$

with population parameters denoted by $\boldsymbol{\Theta} = (\boldsymbol{\Gamma}', \sigma_{\varepsilon}^2)'$. The predictors \mathbf{X} contain both the focal predictor $\mathbf{X}_{\text{focal}}$ and p contextual predictors \mathbf{C} , $\mathbf{X} = (\mathbf{X}_{\text{focal}} \mid \mathbf{C})$. The covariance matrix of the predictors \mathbf{X} is denoted as:

$$\text{cov}(\mathbf{X}) = \frac{1}{n-1}(\mathbf{X}\mathbf{X}') = \boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_{XX} & \boldsymbol{\Phi}_{XC} \\ \boldsymbol{\Phi}_{CX} & \boldsymbol{\Phi}_{CC} \end{bmatrix}. \quad (\text{A2})$$

The coefficients $\boldsymbol{\Gamma}$ can thus be partitioned into focal parameter γ and peripheral parameters $\boldsymbol{\beta}$:

$$\boldsymbol{\Gamma} = [\gamma \mid \boldsymbol{\beta}']'. \quad (\text{A3})$$

We can also partition the parameters $\boldsymbol{\Theta}$ into $\boldsymbol{\Theta} = (\boldsymbol{\Theta}'_{\text{focal}} \mid \boldsymbol{\Theta}'_{\text{peripheral}})' = (\gamma \mid \boldsymbol{\beta}', \sigma_{\varepsilon}^2)'$.

Assuming normally distributed errors, the likelihood function is computed as:

$$L(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{X}) = P(\mathbf{y}|\boldsymbol{\Theta}, \mathbf{X}) = (2\pi\sigma_{\varepsilon}^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2} \sum \frac{(y_i - \mathbf{x}_i\boldsymbol{\Gamma})^2}{\sigma_{\varepsilon}^2}\right), \quad (\text{A4})$$

and the corresponding log likelihood is computed as:

$$LL(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{X}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma_{\varepsilon}^2) - \frac{1}{2\sigma_{\varepsilon}^2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\Gamma})'(\mathbf{y} - \mathbf{X}\boldsymbol{\Gamma})]. \quad (\text{A5})$$

It is well known that the maximum likelihood estimation (MLE) estimator of $\boldsymbol{\Theta}$ (i.e., $\hat{\boldsymbol{\Theta}}$) can be obtained as follows:

$$\begin{aligned} \hat{\boldsymbol{\Gamma}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ \hat{\sigma}_{\varepsilon}^2 &= \frac{1}{n} \sum (y_i - \mathbf{x}_i\hat{\boldsymbol{\Gamma}})^2. \end{aligned} \quad (\text{A6})$$

The MLE estimator $\hat{\boldsymbol{\Theta}}$ is also asymptotically normally distributed, with asymptotic variance computed from the inverse of Fisher information:

$$I(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{X})|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}}^{-1} = \left(-\frac{1}{n} \left[\frac{\partial^2 LL(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{X})}{\partial \boldsymbol{\Theta}^2} \right]_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}} \right)^{-1}. \quad (\text{A7})$$

The observed information can be organized as a block matrix:

$$\begin{aligned} &I(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{X})|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}} \\ &= \frac{1}{n} \begin{bmatrix} -\left[\frac{\partial^2 LL(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{X})}{\partial \boldsymbol{\Gamma}^2} \right]_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}} & -\left[\frac{\partial}{\partial \sigma_{\varepsilon}^2} \frac{\partial LL(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{X})}{\partial \boldsymbol{\Gamma}} \right]_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}} \\ -\left[\frac{\partial}{\partial \boldsymbol{\Gamma}} \frac{\partial LL(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{X})}{\partial \sigma_{\varepsilon}^2} \right]_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}} & -\left[\frac{\partial^2 LL(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{X})}{\partial \sigma_{\varepsilon}^2} \right]_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}} \end{bmatrix}. \end{aligned} \quad (\text{A8})$$

¹⁷ The proof of the conditional pessimism is derived more generally on mean-centered data, with standardized data being a special case. As shown later, conclusions regarding the conditional pessimism remain the same.

What we care about for the problem at hand is the first block, because the standard error of $\hat{\Gamma}$ is computed as the square root of its first element:

$$\begin{aligned} -\frac{1}{n} \left[\frac{\partial^2 LL(\Theta; \mathbf{y}, \mathbf{X})}{\partial \Gamma^2} \right]_{\Theta=\hat{\Theta}} &= -\frac{1}{n} \left[\frac{\partial}{\partial \Gamma} \frac{\partial}{\partial \Gamma} \left(-\frac{1}{2\sigma_\epsilon^2} [\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\Gamma - \Gamma'\mathbf{X}'\mathbf{y} + \Gamma'\mathbf{X}'\mathbf{X}\Gamma] \right) \right]_{\Theta=\hat{\Theta}} \\ &= -\frac{1}{n} \left[\frac{\partial}{\partial \Gamma} \left(-\frac{1}{2\sigma_\epsilon^2} [(\mathbf{y}'\mathbf{X})' - \mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\Gamma] \right) \right]_{\Theta=\hat{\Theta}} = \frac{1}{n} \frac{\mathbf{X}'\mathbf{X}}{\hat{\sigma}_\epsilon^2}. \end{aligned} \quad (\text{A9})$$

On the other hand, it can also be shown that the other two blocks contain only zeroes:

$$\begin{aligned} -\frac{1}{n} \left[\frac{\partial}{\partial \sigma_\epsilon^2} \frac{\partial LL(\Theta; \mathbf{y}, \mathbf{X})}{\partial \Gamma} \right]_{\Theta=\hat{\Theta}} &= -\frac{1}{n} \left[\frac{\partial}{\partial \sigma_\epsilon^2} \left(\frac{\mathbf{X}'(\mathbf{y} - \mathbf{X}\Gamma)}{\sigma_\epsilon^2} \right) \right]_{\Theta=\hat{\Theta}} \\ &= -\frac{1}{n} \left[-\frac{\mathbf{X}'(\mathbf{y} - \mathbf{X}\Gamma)}{(\sigma_\epsilon^2)^2} \right]_{\Theta=\hat{\Theta}} \\ &= \frac{1}{n} \frac{[\mathbf{X}'(\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y})]}{(\hat{\sigma}_\epsilon^2)^2} \\ &= \frac{1}{n} \frac{[\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]}{(\hat{\sigma}_\epsilon^2)^2} \\ &= \frac{1}{n} \frac{[\mathbf{X}'\mathbf{y} - \mathbf{I}\mathbf{X}'\mathbf{y}]}{(\hat{\sigma}_\epsilon^2)^2} = \mathbf{0}. \end{aligned} \quad (\text{A10})$$

Therefore, the information can be simplified as:

$$I(\Theta; \mathbf{y}, \mathbf{X})|_{\Theta=\hat{\Theta}} = \begin{bmatrix} \frac{1}{n} \frac{\mathbf{X}'\mathbf{X}}{\hat{\sigma}_\epsilon^2} & \mathbf{0} \\ \mathbf{0} & -\frac{1}{n} \left[\frac{\partial^2 LL(\Theta; \mathbf{y}, \mathbf{X})}{\partial \sigma_\epsilon^2{}^2} \right]_{\Theta=\hat{\Theta}} \end{bmatrix}, \quad (\text{A11})$$

and the asymptotic covariance matrix can thus be simplified as:

$$(I(\Theta; \mathbf{y}, \mathbf{X})|_{\Theta=\hat{\Theta}})^{-1} = \begin{bmatrix} \left(\frac{1}{n} \frac{\mathbf{X}'\mathbf{X}}{\hat{\sigma}_\epsilon^2} \right)^{-1} & \mathbf{0} \\ \mathbf{0} & -\frac{1}{n} \left[\frac{\partial^2 LL(\Theta; \mathbf{y}, \mathbf{X})}{\partial \sigma_\epsilon^2{}^2} \right]_{\Theta=\hat{\Theta}}^{-1} \end{bmatrix}. \quad (\text{A12})$$

The first block of the asymptotic covariance matrix can be further written as a partitioned matrix:

$$\begin{aligned}
 \left[\frac{1}{n} \frac{(\mathbf{X}'\mathbf{X})}{\hat{\sigma}_\epsilon^2} \right]^{-1} &= \left(\frac{1}{n\hat{\sigma}_\epsilon^2} \begin{bmatrix} \mathbf{X}'_{\text{focal}} \\ \mathbf{C}' \end{bmatrix} \begin{bmatrix} \mathbf{X}_{\text{focal}} & \mathbf{C} \end{bmatrix} \right)^{-1} \\
 &= \left(\frac{1}{n\hat{\sigma}_\epsilon^2} \begin{bmatrix} \mathbf{X}'_{\text{focal}}\mathbf{X}_{\text{focal}} & \mathbf{X}'_{\text{focal}}\mathbf{C} \\ \mathbf{C}'\mathbf{X}_{\text{focal}} & \mathbf{C}'\mathbf{C} \end{bmatrix} \right)^{-1} \\
 &= \begin{bmatrix} \frac{1}{n\hat{\sigma}_\epsilon^2} (\mathbf{X}'_{\text{focal}}\mathbf{X}_{\text{focal}}) & \frac{1}{n\hat{\sigma}_\epsilon^2} (\mathbf{X}'_{\text{focal}}\mathbf{C}) \\ \frac{1}{n\hat{\sigma}_\epsilon^2} (\mathbf{C}'\mathbf{X}_{\text{focal}}) & \frac{1}{n\hat{\sigma}_\epsilon^2} (\mathbf{C}'\mathbf{C}) \end{bmatrix}^{-1} \\
 &= \frac{n\hat{\sigma}_\epsilon^2}{n-1} \begin{bmatrix} \Phi_{XX} & \Phi_{XC} \\ \Phi_{CX} & \Phi_{CC} \end{bmatrix}^{-1} \\
 &= \frac{n\hat{\sigma}_\epsilon^2}{n-1} \begin{bmatrix} (\Phi_{XX} - \Phi_{XC}\Phi_{CC}^{-1}\Phi_{CX})^{-1} & -\Phi_{XX}^{-1}\Phi_{XC}(\Phi_{CC} - \Phi_{CX}\Phi_{XX}^{-1}\Phi_{XC})^{-1} \\ -\Phi_{CC}^{-1}\Phi_{CX}(\Phi_{XX} - \Phi_{XC}\Phi_{CC}^{-1}\Phi_{CX})^{-1} & (\Phi_{CC} - \Phi_{CX}\Phi_{XX}^{-1}\Phi_{XC})^{-1} \end{bmatrix}^{-1}.
 \end{aligned} \tag{A13}$$

The first block in this partitioned matrix is by definition a scalar, and thus can be further simplified as:

$$\begin{aligned}
 \frac{n\hat{\sigma}_\epsilon^2}{n-1} (\Phi_{XX} - \Phi_{XC}\Phi_{CC}^{-1}\Phi_{CX})^{-1} &= \frac{n\hat{\sigma}_\epsilon^2}{n-1} \times \frac{1/\sigma_X^2}{1 - \frac{\Phi_{XC}}{\sqrt{\sigma_X^2}} \Phi_{CC}^{-1} \frac{\Phi_{CX}}{\sqrt{\sigma_X^2}}} \\
 &= \frac{n\hat{\sigma}_\epsilon^2}{n-1} \times \frac{1/\sigma_X^2}{1 - \frac{\Phi_{XC}}{\sqrt{\sigma_X^2}} [\text{diag}(\Phi_{CC})]^{-\frac{1}{2}} [\text{diag}(\Phi_{CC})]^{\frac{1}{2}} \Phi_{CC}^{-1} [\text{diag}(\Phi_{CC})]^{\frac{1}{2}} [\text{diag}(\Phi_{CC})]^{-\frac{1}{2}} \frac{\Phi_{CX}}{\sqrt{\sigma_X^2}}} \\
 &= \frac{n\hat{\sigma}_\epsilon^2}{n-1} \times \frac{1/\sigma_X^2}{1 - \mathbf{P}_{XC} [\text{diag}(\Phi_{CC})]^{\frac{1}{2}} \Phi_{CC}^{-1} [\text{diag}(\Phi_{CC})]^{\frac{1}{2}} \mathbf{P}_{XC}} \\
 &= \frac{n\hat{\sigma}_\epsilon^2}{n-1} \times \frac{1/\sigma_X^2}{1 - \mathbf{P}_{XC} \left([\text{diag}(\Phi_{CC})]^{-\frac{1}{2}} \Phi_{CC} [\text{diag}(\Phi_{CC})]^{\frac{1}{2}} \right)^{-1} \mathbf{P}_{XC}} \\
 &= \frac{n\hat{\sigma}_\epsilon^2}{n-1} \times \frac{1/\sigma_X^2}{1 - \mathbf{P}_{XC} (\mathbf{P}_{CC})^{-1} \mathbf{P}_{XC}} \\
 &= \frac{n\hat{\sigma}_\epsilon^2}{n-1} \times \frac{1/\sigma_X^2}{1 - \rho_{XC}^2},
 \end{aligned} \tag{A14}$$

where ρ_{XC}^2 is the coefficient of determination for regressing the focal predictor X on contextual predictors \mathbf{C} .

Therefore, the standard error for $\hat{\gamma}$ is:¹⁸

$$\sqrt{\frac{1}{n(n-1)(1-\rho_{XC}^2)}} = \sqrt{\frac{\hat{\sigma}_\epsilon^2/\sigma_X^2}{(n-1)(1-\rho_{XC}^2)}}. \tag{A15}$$

\therefore when $\rho_{XC}^2 \rightarrow 1$, the standard error for testing $\hat{\gamma} \rightarrow \infty$, and thus the corresponding statistical power approaches its global minimum. ■

Next, we derive the theoretical lower bound of the statistical power for testing the focal parameter γ under perfect collinearity $\rho_{XC}^2 \rightarrow 1$.

The population covariance matrix is defined based on a full (f) model:

$$\Sigma_f(\Theta_f) = \begin{bmatrix} \Phi & \Phi\Gamma \\ \Gamma'\Phi & \Gamma'\Phi\Gamma + \Psi \end{bmatrix}. \tag{A16}$$

¹⁸ Again, this standard error formula still assumes mean-centered variables for generality, with standardized data being a special case. If standardized variables are assumed here, the derivation of this equation would become more straightforward, but arriving at the same conclusion that when $\rho_{XC}^2 \rightarrow 1$, the statistical power approaches its global minimum. Note that although the standard error will be different depending on the scaling, the difference is only in the numerator of the equation; the pessimum, however, is determined solely based on the denominator.

The population parameter can be partitioned into the focal parameter γ and peripheral contextual parameters: $\Theta_f = (\gamma, \Theta'_{\text{peripheral}})'$. We then need to fit a reduced (r) model to the population, where the focal path γ is constrained to zero, $\Theta_r = (\gamma = 0, \Theta'_{\text{peripheral}})'$. The reduced model implied covariance matrix can be expressed as:

$$\Sigma_r(\Theta_r) = \begin{bmatrix} \Phi_r & \Phi_r \Gamma_r \\ \Gamma_r' \Phi_r & \Gamma_r' \Phi_r \Gamma_r + \Psi_r \end{bmatrix}. \quad (\text{A17})$$

To obtain the parameter estimates for the reduced model, we need to find the parameter estimates $\hat{\Theta}_r$ that minimizes the fit function:

$$F_r = \ln |\Sigma_r| + \text{tr}(\Sigma_f \Sigma_r^{-1}) - \ln |\Sigma_f| - p. \quad (\text{A18})$$

For our current problem, we can consider the simple case with one focal predictor X and one contextual predictor C , further assuming standardized form from this point onward in deriving the theoretical power lower bound. Therefore, the population values of the model parameters can be written as:

$$\begin{aligned} \Phi &= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \\ \Gamma &= [\gamma \quad \beta]', \\ \Psi &= \psi_\epsilon, \end{aligned} \quad (\text{A19})$$

with the known constraint that:

$$\Gamma' \Phi \Gamma + \Psi = \gamma^2 + \beta^2 + 2\beta\gamma\rho + \psi_\epsilon = 1. \quad (\text{A20})$$

The population covariance matrix is thus:

$$\Sigma_f(\Theta_f) = \begin{bmatrix} 1 & \rho & \gamma + \beta\rho \\ \rho & 1 & \beta + \gamma\rho \\ \gamma + \beta\rho & \beta + \gamma\rho & \gamma^2 + \beta^2 + 2\beta\gamma\rho + \psi_\epsilon \end{bmatrix}. \quad (\text{A21})$$

Correspondingly, the reduced model parameters are denoted as:

$$\begin{aligned} \Phi_r &= \begin{bmatrix} c_1 & r \\ r & c_2 \end{bmatrix}, \\ \Gamma_r &= [0 \quad b], \\ \Psi_r &= c_3, \end{aligned} \quad (\text{A22})$$

producing the reduced model implied covariance matrix:

$$\Sigma_r(\Theta_r) = \begin{bmatrix} c_1 & r & br \\ r & c_2 & b \\ br & b & b^2 c_2 + c_3 \end{bmatrix}. \quad (\text{A23})$$

When $\rho \rightarrow 1$ (or -1 ; the conclusion will be the same), the reduced model parameter estimates are the following:

$$\left\{ \begin{array}{l} \hat{c}_1 = 1 \\ \hat{c}_2 = 1 \\ \hat{r} = \rho \\ \hat{b} = \gamma\rho + \beta \\ \hat{c}_3 = 1 - \gamma^2\rho^2 - \beta^2 - 2\gamma\rho\beta \end{array} \right., \quad (\text{A24})$$

which minimizes the discrepancy between $\Sigma_f(\Theta_f)$ and $\Sigma_r(\hat{\Theta}_r)$.

Therefore, under the condition $|\rho| \rightarrow 1$, the estimated reduced model parameters are theoretically the following:

$$\begin{aligned} \hat{\Phi}_r &= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \\ \hat{\Gamma}_r &= [0 \quad \gamma\rho + \beta], \\ \hat{\Psi}_r &= 1 - \gamma^2\rho^2 - \beta^2 - 2\gamma\rho\beta. \end{aligned} \quad (\text{A25})$$

(Note, however, when ρ deviates increasingly from 1, the estimated model parameters will also deviate more from this theoretical result.) ■

(Appendices continue)

To derive the theoretical lower bound of power for testing $\hat{\gamma}$, we follow the Satorra and Saris (1985) approach. The statistical power is minimized when the following target fit function is minimized:

$$F_r = \ln |\hat{\Sigma}_r| + \text{tr}(\Sigma_f \hat{\Sigma}_r^{-1}) - \ln |\Sigma_f| - 3. \quad (\text{A26})$$

Let us substitute the population parameter values and theoretical parameter estimates into each of the terms within the target fit function:

1. The first term: $\ln |\hat{\Sigma}_r|$

$$\begin{aligned} \ln |\hat{\Sigma}_r| &= \ln \begin{vmatrix} 1 & \rho & (\gamma\rho + \beta)\rho \\ \rho & 1 & \gamma\rho + \beta \\ (\gamma\rho + \beta)\rho & \gamma\rho + \beta & (\gamma\rho + \beta)^2 + 1 - \gamma^2\rho^2 - \beta^2 - 2\gamma\rho\beta \end{vmatrix} \\ &= \ln [(1 - \rho^2)[1 - (\beta + \gamma\rho)^2]]. \end{aligned} \quad (\text{A27})$$

2. The second term: $\text{tr}(\Sigma_f \Sigma_r^{-1})$

$$\begin{aligned} \Sigma_r^{-1} &= \begin{bmatrix} \frac{1}{1 - \rho^2} & \frac{\rho}{\rho^2 - 1} & 0 \\ \frac{\rho}{\rho^2 - 1} & \frac{-1 + \beta^2\rho^2 + 2\beta\gamma\rho^3 + \gamma^2\rho^4}{(1 - \rho^2)[1 - (\beta + \gamma\rho)^2]} & \frac{\beta + \gamma\rho}{-1 + (\beta + \gamma\rho)^2} \\ 0 & \frac{\beta + \gamma\rho}{-1 + (\beta + \gamma\rho)^2} & \frac{-1}{-1 + (\beta + \gamma\rho)^2} \end{bmatrix} \\ \Sigma_f \Sigma_r^{-1} &= \begin{bmatrix} 1 & \rho & \gamma + \beta\rho \\ \rho & 1 & \beta + \gamma\rho \\ \gamma + \beta\rho & \beta + \gamma\rho & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{1 - \rho^2} & \frac{\rho}{\rho^2 - 1} & 0 \\ \frac{\rho}{\rho^2 - 1} & \frac{-1 + \beta^2\rho^2 + 2\beta\gamma\rho^3 + \gamma^2\rho^4}{(1 - \rho^2)[1 - (\beta + \gamma\rho)^2]} & \frac{\beta + \gamma\rho}{-1 + (\beta + \gamma\rho)^2} \\ 0 & \frac{\beta + \gamma\rho}{-1 + (\beta + \gamma\rho)^2} & \frac{-1}{-1 + (\beta + \gamma\rho)^2} \end{bmatrix} \\ &= \begin{bmatrix} 1 & \frac{\gamma(\beta + \gamma\rho)(\rho^2 - 1)}{-1 + (\beta + \gamma\rho)^2} & \frac{\gamma(\rho^2 - 1)}{-1 + (\beta + \gamma\rho)^2} \\ 0 & 1 & 0 \\ \gamma & -\gamma\rho & 1 \end{bmatrix} \\ \therefore \text{tr}(\Sigma_f \Sigma_r^{-1}) &= 3. \end{aligned} \quad (\text{A28})$$

3. The third term: $\ln |\Sigma_f|$

$$\begin{aligned} \ln |\Sigma_f| &= \ln \begin{vmatrix} 1 & \rho & \gamma + \beta\rho \\ \rho & 1 & \beta + \gamma\rho \\ \gamma + \beta\rho & \beta + \gamma\rho & 1 \end{vmatrix} \\ &= \ln [(1 - \rho^2)[1 - \beta^2 - \gamma^2 - 2\beta\gamma\rho]]. \end{aligned} \quad (\text{A29})$$

Putting them together, we thus have the target fit function:

$$\begin{aligned} F_r &= \ln [(1 - \rho^2)[1 - (\beta + \gamma\rho)^2]] + 3 - \ln [(1 - \rho^2)[1 - \beta^2 - \gamma^2 - 2\beta\gamma\rho]] - 3 \\ &= \ln \frac{(1 - \rho^2)[1 - (\beta + \gamma\rho)^2]}{(1 - \rho^2)[1 - \beta^2 - \gamma^2 - 2\beta\gamma\rho]}. \end{aligned} \quad (\text{A30})$$

Note that for $\ln \frac{(1 - \rho^2)[1 - (\beta + \gamma\rho)^2]}{(1 - \rho^2)[1 - \beta^2 - \gamma^2 - 2\beta\gamma\rho]}$ to be mathematically legitimate, ρ^2 cannot be exactly equal to 1.

To make sure this is mathematically permissible, let us denote $\rho = 1 - \delta$, where δ is a sufficiently small positive value such that $0 < \delta < 1$. [Note. The same results can be obtained for the case $\rho \leq 0$. We can let $\rho = -1 + \delta$, \forall a sufficiently small positive value δ , $0 < \delta < 1$.]

(Appendices continue)

The target fit function can thus be expressed as:

$$\begin{aligned}
 F_r &= \ln \frac{1 - [\beta + \gamma(1 - \delta)]^2}{[1 - \beta^2 - \gamma^2 - 2\beta\gamma(1 - \delta)]} \\
 &= \ln \frac{1 - [(\beta + \gamma) - \gamma\delta]^2}{[1 - (\beta^2 + \gamma^2 + 2\beta\gamma) + 2\beta\gamma\delta]} \\
 &= \ln \frac{[1 - (\beta + \gamma)^2 + 2\gamma\delta\beta] + 2\gamma^2\delta - \gamma^2\delta^2}{[1 - (\beta + \gamma)^2 + 2\beta\gamma\delta]} \\
 &= \ln \left[1 + \frac{\gamma^2\delta(2 - \delta)}{1 - (\beta + \gamma)^2 + 2\beta\gamma\delta} \right].
 \end{aligned} \tag{A31}$$

Importantly, we are interested in the population condition that yields the lowest power given a fixed focal parameter γ . That is to say, we need to find the population value of contextual path parameter β that minimizes the target function F_r . Therefore, we take the derivative with respect to β :

$$\begin{aligned}
 \frac{\partial F_r}{\partial \beta} &= \frac{[1 - (\beta + \gamma)^2 + 2\beta\gamma\delta]}{[1 - (\beta + \gamma)^2 + 2\gamma\delta\beta] + 2\gamma^2\delta - \gamma^2\delta^2} \left(\frac{\partial}{\partial \beta} \left[1 + \frac{\gamma^2\delta(2 - \delta)}{1 - (\beta + \gamma)^2 + 2\beta\gamma\delta} \right] \right) \\
 &= \frac{[1 - (\beta + \gamma)^2 + 2\beta\gamma\delta]}{[1 - (\beta + \gamma)^2 + 2\gamma\delta\beta] + 2\gamma^2\delta - \gamma^2\delta^2} \left(\frac{\gamma^2\delta(2 - \delta)(-2\beta - 2\gamma + 2\gamma\delta)}{[1 - (\beta + \gamma)^2 + 2\beta\gamma\delta]^2} \right) \\
 &= \frac{1}{[1 - (\beta + \gamma)^2 + 2\gamma\delta\beta] + 2\gamma^2\delta - \gamma^2\delta^2} \left(\frac{\gamma^2\delta(2 - \delta)(-2\beta - 2\gamma + 2\gamma\delta)}{1 - (\beta + \gamma)^2 + 2\beta\gamma\delta} \right) \\
 &= \frac{-2\gamma^2\delta(2 - \delta)(\beta + \gamma - \gamma\delta)}{([1 - (\beta + \gamma)^2 + 2\gamma\delta\beta] + 2\gamma^2\delta - \gamma^2\delta^2)[1 - (\beta + \gamma)^2 + 2\beta\gamma\delta]}.
 \end{aligned} \tag{A32}$$

To have $\frac{\partial F_r}{\partial \beta} = 0$ given a sufficiently small $0 < \delta < 1$, we need to set $\beta + \gamma - \gamma\delta = 0$. Therefore, we have the following result:

$$\beta = \gamma(\delta - 1). \tag{A33}$$

\therefore when $\beta = -\gamma\rho$, F_r is minimized. ■

The theoretical lower bound of F_r can thus be computed as:

$$\begin{aligned}
 F_r &= \ln \left[1 + \frac{\gamma^2\delta(2 - \delta)}{1 - (\gamma(\delta - 1) + \gamma)^2 + 2\gamma(\delta - 1)\gamma\delta} \right] \\
 &= \ln \left[\frac{1 + \gamma^2\delta^2 - 2\gamma^2\delta + \gamma^2\delta(2 - \delta)}{1 + \gamma^2\delta^2 - 2\gamma^2\delta} \right] \\
 &= \ln \left[\frac{1}{1 + \gamma^2\delta^2 - 2\gamma^2\delta} \right] \\
 &= \ln \left[\frac{1}{1 - \gamma^2(1 - \rho^2)} \right].
 \end{aligned} \tag{A34}$$

In summary, in this appendix, we have proved the following: when $|\rho| \rightarrow 1$ and $\beta = -\gamma\rho$, the fit function F_r is minimized, and thus the statistical power for testing $\hat{\gamma}$ is minimized:

$$\lim_{|\rho| \rightarrow 1, \beta = -\gamma\rho} F_r = \ln \left[\frac{1}{1 - \gamma^2(1 - \rho^2)} \right]. \tag{A35}$$

Therefore, the theoretical n_{\max} is,

$$n_{\max} = \frac{\lambda}{\ln \left[\frac{1}{1 - \gamma^2(1 - \rho^2)} \right]} + 1 = \frac{\lambda}{-\ln [1 - \gamma^2(1 - \rho^2)]} + 1. \tag{A36}$$

Appendix B

Equivalence of Uncollapsed and Collapsed Multiple Regression Models for Testing γ

In this appendix, we prove the equivalence between the uncollapsed model and the collapsed model in terms of testing focal path γ .

We begin with the likelihood ratio test (LRT) under the uncollapsed model. The parameters (Θ) of the uncollapsed model are denoted as follows:

1. The path coefficients are: $\mathbf{\Gamma}_{1 \times (p+1)} = [\gamma, \beta_1, \beta_2, \dots, \beta_p]$,
and particularly, the contextual path coefficients are: $\mathbf{\beta}_{1 \times p} = [\beta_1, \beta_2, \dots, \beta_p]$,
2. The covariance matrix for exogenous predictors \mathbf{X} is:

$$\mathbf{\Phi} = \begin{bmatrix} c_{XX} & & & \\ c_{1X} & c_{11} & & \\ \vdots & \vdots & \ddots & \\ c_{pX} & c_{p1} & \dots & c_{pp} \end{bmatrix}, \quad (\text{B1})$$

3. The error variance of outcome Y is ψ .

The observed data are denoted as:

$$\mathbf{Q}_{(p+2) \times n} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \\ \mathbf{C} \end{bmatrix} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \\ \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_p \end{bmatrix}. \quad (\text{B2})$$

Therefore, the model-implied covariance matrix under the uncollapsed model can be written in the following block matrix:

$$\mathbf{\Sigma}(\Theta) = \left[\begin{array}{c|c} \mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}' + \psi & \mathbf{\Gamma}\mathbf{\Phi} \\ \hline \mathbf{\Phi}\mathbf{\Gamma}' & \mathbf{\Phi} \end{array} \right] = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C}^* & \mathbf{D} \end{bmatrix}, \quad (\text{B3})$$

$\begin{matrix} 1 \times 1 & 1 \times (p+1) \\ (p+1) \times 1 & (p+1) \times (p+1) \end{matrix}$

where:

$$\begin{aligned} \mathbf{A}_{1 \times 1} &= \mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}' + \psi, \\ \mathbf{B}_{1 \times (p+1)} &= \mathbf{\Gamma}\mathbf{\Phi}, \\ \mathbf{C}^*_{(p+1) \times 1} &= \mathbf{\Phi}\mathbf{\Gamma}', \\ \mathbf{D} &= \mathbf{\Phi}. \end{aligned} \quad (\text{B4})$$

Assuming multivariate normality and independence, the likelihood of Θ given the observed data are:

$$\begin{aligned} L(\Theta; \mathbf{Q}) &= \prod_{i=1}^n (2\pi)^{-\frac{p+2}{2}} \det[\mathbf{\Sigma}(\Theta)]^{-\frac{1}{2}} \exp \left[-\frac{1}{2} \mathbf{Q}_i' \mathbf{\Sigma}(\Theta)^{-1} \mathbf{Q}_i \right] \\ &= (2\pi)^{-\frac{(p+2)n}{2}} \det[\mathbf{\Sigma}(\Theta)]^{-\frac{n}{2}} \exp \left[-\frac{1}{2} \sum \mathbf{Q}_i' \mathbf{\Sigma}(\Theta)^{-1} \mathbf{Q}_i \right]. \end{aligned} \quad (\text{B5})$$

Setting the focal parameter $\gamma = 0$ and keeping the other parameters fixed, we will obtain the reduced (uncollapsed) model, whose model parameters (Θ_r) are correspondingly defined as the following:

1. The path coefficients are: $\mathbf{\Gamma}_r_{1 \times (p+1)} = [0, \beta_1, \beta_2, \dots, \beta_p]$
and particularly, the contextual path coefficients are: $\mathbf{\beta}_r = \mathbf{\beta}_{1 \times p}$
2. The covariance matrix for exogenous predictors \mathbf{X} is:

$$\mathbf{\Phi}_r = \mathbf{\Phi} \quad (\text{B6})$$

3. The error variance of outcome Y is $\psi_r = \psi$.

(Appendices continue)

Therefore, the model-implied covariance matrix under the reduced uncollapsed model can be written as the following block matrix:

$$\begin{aligned}\Sigma_r(\Theta) &= \begin{bmatrix} \Gamma_r \Phi_r \Gamma_r' + \Psi_r & \Gamma_r \Phi_r \\ \Phi_r \Gamma_r' & \Phi_r \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}_r & \mathbf{B}_r \\ \mathbf{C}_r & \mathbf{D}_r \end{bmatrix} = \begin{bmatrix} \mathbf{A}_r & \mathbf{B}_r \\ \mathbf{C}_r & \mathbf{D} \end{bmatrix},\end{aligned}\quad (\text{B7})$$

where:

$$\begin{aligned}\mathbf{A}_r &= \Gamma_r \Phi_r \Gamma_r' + \Psi_r = \Gamma_r \Phi \Gamma_r' + \Psi, \\ \mathbf{B}_r &= \Gamma_r \Phi, \\ \mathbf{C}_r &= \Phi \Gamma_r', \\ \mathbf{D}_r &= \mathbf{D} = \Phi.\end{aligned}\quad (\text{B8})$$

Similarly, assuming multivariate normality and independence, the likelihood of Θ_r given the observed data are:

$$\begin{aligned}L_r(\Theta_r; \mathbf{Q}) &= \prod_{i=1}^n (2\pi)^{-\frac{p+2}{2}} \det[\Sigma_r(\Theta)]^{-\frac{1}{2}} \exp\left[-\frac{1}{2} \mathbf{Q}_i' \Sigma_r(\Theta)^{-1} \mathbf{Q}_i\right] \\ &= (2\pi)^{-\frac{(p+2)n}{2}} \det[\Sigma_r(\Theta)]^{-\frac{n}{2}} \exp\left[-\frac{1}{2} \sum \mathbf{Q}_i' \Sigma_r(\Theta)^{-1} \mathbf{Q}_i\right].\end{aligned}\quad (\text{B9})$$

With Equations B5 and B9, the LRT for testing the focal parameter γ is thus defined as:

$$\begin{aligned}\frac{L(\Theta; \mathbf{Q})}{L_r(\Theta_r; \mathbf{Q})} &= \frac{(2\pi)^{-\frac{(p+2)n}{2}} \det[\Sigma(\Theta)]^{-\frac{n}{2}} \exp\left[-\frac{1}{2} \sum \mathbf{Q}_i' \Sigma(\Theta)^{-1} \mathbf{Q}_i\right]}{(2\pi)^{-\frac{(p+2)n}{2}} \det[\Sigma_r(\Theta)]^{-\frac{n}{2}} \exp\left[-\frac{1}{2} \sum \mathbf{Q}_i' \Sigma_r(\Theta)^{-1} \mathbf{Q}_i\right]} \\ &= \left[\frac{\det[\Sigma(\Theta)]}{\det[\Sigma_r(\Theta)]} \right]^{-\frac{n}{2}} \exp\left[-\frac{1}{2} \sum \mathbf{Q}_i' (\Sigma(\Theta)^{-1} - \Sigma_r(\Theta)^{-1}) \mathbf{Q}_i\right].\end{aligned}\quad (\text{B10})$$

Using Theorem 8.2.1 (2) in Graybill (1983, p. 184), the determinant of $\Sigma(\Theta)$ and $\Sigma_r(\Theta)$ can be written as:

$$\begin{aligned}\det[\Sigma(\Theta)] &= \det[\mathbf{D}] \det[\mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C}^*] \\ &= \det[\mathbf{D}] (\mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C}^*), \\ \det[\Sigma_r(\Theta)] &= \det[\mathbf{D}_r] \det[\mathbf{A}_r - \mathbf{B}_r \mathbf{D}_r^{-1} \mathbf{C}_r] \\ &= \det[\mathbf{D}] (\mathbf{A}_r - \mathbf{B}_r \mathbf{D}^{-1} \mathbf{C}_r).\end{aligned}\quad (\text{B11})$$

Therefore, the first term in Equation B10 can be computed as:

$$\begin{aligned}\left[\frac{\det[\Sigma(\Theta)]}{\det[\Sigma_r(\Theta)]} \right]^{-\frac{n}{2}} &= \left[\frac{\det[\mathbf{D}] (\mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C}^*)}{\det[\mathbf{D}] (\mathbf{A}_r - \mathbf{B}_r \mathbf{D}^{-1} \mathbf{C}_r)} \right]^{-\frac{n}{2}} = \left[\frac{(\mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C}^*)}{(\mathbf{A}_r - \mathbf{B}_r \mathbf{D}^{-1} \mathbf{C}_r)} \right]^{-\frac{n}{2}} \\ &= \left[\frac{(\Gamma \Phi \Gamma' + \Psi - \Gamma \Phi \Phi^{-1} \Phi \Gamma')}{(\Gamma_r \Phi \Gamma_r' + \Psi_r - \Gamma_r \Phi \Phi^{-1} \Phi \Gamma_r')} \right]^{-\frac{n}{2}} \\ &= \left[\frac{(\Gamma \Phi \Gamma' + \Psi - \Gamma \Phi \Gamma')}{(\Gamma_r \Phi \Gamma_r' + \Psi_r - \Gamma_r \Phi \Gamma_r')} \right]^{-\frac{n}{2}} \\ &= \left[\frac{\Psi}{\Psi_r} \right]^{-\frac{n}{2}} = \left[\frac{\Psi}{\Psi} \right]^{-\frac{n}{2}} = 1.\end{aligned}\quad (\text{B12})$$

(Appendices continue)

Next we further simplify the second term in Equation B10. Using Theorem 8.2.1 (1) in Graybill (1983, p. 184) and Equation 1.36 in Noble (1969, p. 25), the inverse of $\Sigma(\Theta)$ and $\Sigma_r(\Theta)$ can be written as:

$$\begin{aligned}\Sigma^{-1}(\Theta) &= \begin{bmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}^*)^{-1} & -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}^*)^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}^*(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}^*)^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}^*(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}^*)^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix}, \\ \Sigma_r(\Theta)^{-1} &= \begin{bmatrix} (\mathbf{A}_r - \mathbf{B}_r\mathbf{D}_r^{-1}\mathbf{C}_r)^{-1} & -(\mathbf{A}_r - \mathbf{B}_r\mathbf{D}_r^{-1}\mathbf{C}_r)^{-1}\mathbf{B}_r\mathbf{D}_r^{-1} \\ -\mathbf{D}_r^{-1}\mathbf{C}_r(\mathbf{A}_r - \mathbf{B}_r\mathbf{D}_r^{-1}\mathbf{C}_r)^{-1} & \mathbf{D}_r^{-1} + \mathbf{D}_r^{-1}\mathbf{C}_r(\mathbf{A}_r - \mathbf{B}_r\mathbf{D}_r^{-1}\mathbf{C}_r)^{-1}\mathbf{B}_r\mathbf{D}_r^{-1} \end{bmatrix}.\end{aligned}\tag{B13}$$

Therefore, the difference between $\Sigma(\Theta)$ and $\Sigma_r(\Theta)$ can be computed as the following:

$$\Sigma^{-1}(\Theta) - \Sigma_r(\Theta)^{-1} = \begin{bmatrix} \text{(I)} & \text{(II)} \\ \text{(III)} & \text{(IV)} \end{bmatrix},\tag{B14}$$

where (I), (II), (III), and (IV) each represents the corresponding element in this block matrix:

$$\begin{aligned}\text{(I)} &= (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}^*)^{-1} - (\mathbf{A}_r - \mathbf{B}_r\mathbf{D}_r^{-1}\mathbf{C}_r)^{-1} \\ &= \frac{1}{\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}^*} - \frac{1}{\mathbf{A}_r - \mathbf{B}_r\mathbf{D}_r^{-1}\mathbf{C}_r} \\ &= \frac{1}{\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}' + \Psi - \mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}'} - \frac{1}{\mathbf{\Gamma}_r\mathbf{\Phi}\mathbf{\Gamma}_r' + \Psi_r - \mathbf{\Gamma}_r\mathbf{\Phi}\mathbf{\Gamma}_r'} \\ &= \frac{1}{\Psi} - \frac{1}{\Psi_r} = \frac{1}{\Psi} - \frac{1}{\Psi} = \mathbf{0};\end{aligned}\tag{B15}$$

$$\begin{aligned}\text{(II)} &= -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}^*)^{-1}\mathbf{B}\mathbf{D}^{-1} + (\mathbf{A}_r - \mathbf{B}_r\mathbf{D}_r^{-1}\mathbf{C}_r)^{-1}\mathbf{B}_r\mathbf{D}_r^{-1} \\ &= -\frac{1}{\Psi}\mathbf{B}\mathbf{D}^{-1} + \frac{1}{\Psi_r}\mathbf{B}_r\mathbf{D}_r^{-1} \\ &= -\frac{1}{\Psi}(\mathbf{B} - \mathbf{B}_r)\mathbf{D}^{-1} \\ &= -\frac{1}{\Psi}(\mathbf{\Gamma}\mathbf{\Phi} - \mathbf{\Gamma}_r\mathbf{\Phi})\mathbf{\Phi}^{-1} \\ &= -\frac{1}{\Psi}(\mathbf{\Gamma} - \mathbf{\Gamma}_r) \\ &= -\frac{1}{\Psi}\left[\gamma \middle| \mathbf{0} \right]_{1 \times p};\end{aligned}\tag{B16}$$

$$\text{(III)} = -\mathbf{D}^{-1}\mathbf{C}^*(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}^*)^{-1} + \mathbf{D}_r^{-1}\mathbf{C}_r(\mathbf{A}_r - \mathbf{B}_r\mathbf{D}_r^{-1}\mathbf{C}_r)^{-1} = \begin{bmatrix} \frac{\gamma}{\Psi} \\ \mathbf{0} \end{bmatrix}_{p \times 1} \left(-\frac{1}{\Psi}\right);\tag{B17}$$

$$\begin{aligned}
 (\text{IV}) &= \mathbf{D}^{-1} + \mathbf{D}^{-1} \mathbf{C}^* (\mathbf{A} - \mathbf{B} \mathbf{D}^{-1} \mathbf{C}^*)^{-1} \mathbf{B} \mathbf{D}^{-1} - \mathbf{D}_r^{-1} - \mathbf{D}_r^{-1} \mathbf{C}_r (\mathbf{A} - \mathbf{B}_r \mathbf{D}_r^{-1} \mathbf{C}_r)^{-1} \mathbf{B}_r \mathbf{D}_r^{-1} \\
 &= \mathbf{D}^{-1} + \mathbf{D}^{-1} \mathbf{C}^* \left(\frac{1}{\psi} \right) \mathbf{B} \mathbf{D}^{-1} - \mathbf{D}_r^{-1} - \mathbf{D}_r^{-1} \mathbf{C}_r \left(\frac{1}{\psi_r} \right) \mathbf{B}_r \mathbf{D}_r^{-1} \\
 &= \mathbf{D}^{-1} + \mathbf{D}^{-1} \mathbf{C}^* \left(\frac{1}{\psi} \right) \mathbf{B} \mathbf{D}^{-1} - \mathbf{D}^{-1} - \mathbf{D}^{-1} \mathbf{C}_r \left(\frac{1}{\psi} \right) \mathbf{B}_r \mathbf{D}^{-1} \\
 &= \left(\frac{1}{\psi} \right) \mathbf{D}^{-1} (\mathbf{C}^* \mathbf{B} - \mathbf{C}_r \mathbf{B}_r) \mathbf{D}^{-1} \\
 &= \left(\frac{1}{\psi} \right) \mathbf{D}^{-1} (\Phi \Gamma' \Gamma \Phi - \Phi_r \Gamma_r' \Gamma_r \Phi_r) \mathbf{D}^{-1} \\
 &= \left(\frac{1}{\psi} \right) \Phi^{-1} (\Phi \Gamma' \Gamma \Phi - \Phi_r \Gamma_r' \Gamma_r \Phi_r) \Phi^{-1} \\
 &= \left(\frac{1}{\psi} \right) (\Gamma' \Gamma - \Gamma_r' \Gamma_r) \\
 &= \left(\frac{1}{\psi} \right) \left\{ \left[\begin{array}{c|c} \gamma^2 & \gamma \boldsymbol{\beta} \\ \hline \boldsymbol{\beta}' \gamma & \boldsymbol{\beta}' \boldsymbol{\beta} \end{array} \right] - \left[\begin{array}{c|c} 0 & \boldsymbol{\theta}' \\ \hline \mathbf{0} & \boldsymbol{\beta}' \boldsymbol{\beta} \end{array} \right] \right\} \\
 &= \left(\frac{1}{\psi} \right) \left[\begin{array}{c|c} \gamma^2 & \gamma \boldsymbol{\beta} \\ \hline \boldsymbol{\beta}' \gamma & \mathbf{0} \end{array} \right].
 \end{aligned} \tag{B18}$$

Therefore, $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta}) - \boldsymbol{\Sigma}_r(\boldsymbol{\Theta})^{-1}$ is computed as:

$$\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta}) - \boldsymbol{\Sigma}_r(\boldsymbol{\Theta})^{-1} = \left[\begin{array}{cc} 0 & -\frac{1}{\psi} \left[\begin{array}{c|c} \gamma & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{1} \times p \end{array} \right] \\ \left[\begin{array}{c} \gamma \\ \hline \mathbf{0} \\ p \times 1 \end{array} \right] \left(-\frac{1}{\psi} \right) & \left(\frac{1}{\psi} \right) \left[\begin{array}{c|c} \gamma^2 & \gamma \boldsymbol{\beta} \\ \hline \boldsymbol{\beta}' \gamma & \mathbf{0} \end{array} \right] \end{array} \right] = \left[\begin{array}{cc} 0 & -\frac{1}{\psi} \gamma \\ \gamma \left(-\frac{1}{\psi} \right) & \left(\frac{1}{\psi} \right) \gamma^2 \\ \mathbf{0} & \left(\frac{1}{\psi} \right) \boldsymbol{\beta}' \gamma \end{array} \right] \begin{array}{c} \boldsymbol{\theta}' \\ \left(\frac{1}{\psi} \right) \gamma \boldsymbol{\beta} \\ \mathbf{0} \end{array}. \tag{B19}$$

Therefore, the second term in Equation B10 is simplified as:

$$\begin{aligned}
& \exp \left[-\frac{1}{2} \sum \mathbf{Q}_i' (\boldsymbol{\Sigma}(\boldsymbol{\Theta})^{-1} - \boldsymbol{\Sigma}_r(\boldsymbol{\Theta})^{-1}) \mathbf{Q}_i \right] \\
&= \exp \left\{ -\frac{1}{2} \sum \mathbf{Q}_i' \begin{bmatrix} 0 & -\frac{1}{\Psi} \gamma & \mathbf{0}' \\ \gamma \left(-\frac{1}{\Psi} \right) & \left(\frac{1}{\Psi} \right) \gamma^2 & \left(\frac{1}{\Psi} \right) \gamma \boldsymbol{\beta} \\ \mathbf{0} & \left(\frac{1}{\Psi} \right) \boldsymbol{\beta}' \gamma & \mathbf{0} \end{bmatrix} \mathbf{Q}_i \right\} \\
&= \exp \left\{ -\frac{1}{2} \sum \begin{bmatrix} Y_i & X_i & \mathbf{C}_i' \end{bmatrix} \begin{bmatrix} 0 & -\frac{1}{\Psi} \gamma & \mathbf{0}' \\ \gamma \left(-\frac{1}{\Psi} \right) & \left(\frac{1}{\Psi} \right) \gamma^2 & \left(\frac{1}{\Psi} \right) \gamma \boldsymbol{\beta} \\ \mathbf{0} & \left(\frac{1}{\Psi} \right) \boldsymbol{\beta}' \gamma & \mathbf{0} \end{bmatrix} \begin{bmatrix} Y_i \\ X_i \\ \mathbf{C}_i \end{bmatrix} \right\} \tag{B20} \\
&= \exp \left\{ -\frac{1}{2} \sum \begin{bmatrix} X_i \gamma \left(-\frac{1}{\Psi} \right) & -\frac{1}{\Psi} \gamma Y_i + X_i \left(\frac{1}{\Psi} \right) \gamma^2 + \mathbf{C}_i' \left(\frac{1}{\Psi} \right) \boldsymbol{\beta}' \gamma & X_i \left(\frac{1}{\Psi} \right) \gamma \boldsymbol{\beta} \end{bmatrix} \begin{bmatrix} Y_i \\ X_i \\ \mathbf{C}_i \end{bmatrix} \right\} \\
&= \exp \left\{ -\frac{1}{2} \sum \left[X_i \gamma \left(-\frac{1}{\Psi} \right) Y_i - \frac{1}{\Psi} \gamma Y_i X_i + X_i^2 \left(\frac{1}{\Psi} \right) \gamma^2 + \mathbf{C}_i' \left(\frac{1}{\Psi} \right) \boldsymbol{\beta}' \gamma X_i + X_i \left(\frac{1}{\Psi} \right) \gamma \boldsymbol{\beta} \mathbf{C}_i \right] \right\} \\
&= \exp \left\{ -\frac{1}{2\Psi} \sum \left[-X_i \gamma Y_i - \gamma Y_i X_i + X_i^2 \gamma^2 + \mathbf{C}_i' \boldsymbol{\beta}' \gamma X_i + X_i \gamma \boldsymbol{\beta} \mathbf{C}_i \right] \right\} \\
&= \exp \left\{ -\frac{1}{2\Psi} \sum \left[-2\gamma X_i Y_i + X_i^2 \gamma^2 + 2 \sum_{j=1}^p (\beta_j C_{ij}) \gamma X_i \right] \right\}.
\end{aligned}$$

Piecing together Equations B12 and B20, Equation B10 can be simplified as the following:

$$\begin{aligned}
\frac{L(\boldsymbol{\Theta}; \mathbf{Q})}{L_r(\boldsymbol{\Theta}_r; \mathbf{Q})} &= \left[\frac{\det[\boldsymbol{\Sigma}(\boldsymbol{\Theta})]}{\det[\boldsymbol{\Sigma}_r(\boldsymbol{\Theta})]} \right]^{-\frac{n}{2}} \exp \left[-\frac{1}{2} \sum \mathbf{Q}_i' (\boldsymbol{\Sigma}(\boldsymbol{\Theta})^{-1} - \boldsymbol{\Sigma}_r(\boldsymbol{\Theta})^{-1}) \mathbf{Q}_i \right] \\
&= \exp \left\{ -\frac{1}{2\Psi} \sum \left[-2\gamma X_i Y_i + X_i^2 \gamma^2 + 2 \sum_{j=1}^p (\beta_j C_{ij}) \gamma X_i \right] \right\}. \tag{B21}
\end{aligned}$$

Assuming multivariate normality and independence, the likelihood ratio for the collapsed model can be similarly computed (which can be considered a special case of the uncollapsed model with $p = 1$):

$$\frac{L^{(C)}(\boldsymbol{\Theta}^{(C)}; \mathbf{Q}^{(C)})}{L_r^{(C)}(\boldsymbol{\Theta}_r^{(C)}; \mathbf{Q}^{(C)})} = \exp \left\{ -\frac{1}{2\Psi} \sum \left[-2\gamma X_i Y_i + X_i^2 \gamma^2 + 2(\beta_C C_{\bullet}) \gamma X_i \right] \right\}. \tag{B22}$$

Since we define the collapsed model with the known constraints: $C_{\bullet} = \sum_1^p \frac{\beta_j}{\beta_C} C_j$, therefore we can establish the equivalence between Equations B21 and B22:

$$\begin{aligned}
\frac{L^{(C)}(\boldsymbol{\Theta}^{(C)}; \mathbf{Q}^{(C)})}{L_r^{(C)}(\boldsymbol{\Theta}_r^{(C)}; \mathbf{Q}^{(C)})} &= \exp \left\{ -\frac{1}{2\Psi} \sum \left[-2\gamma X_i Y_i + X_i^2 \gamma^2 + 2(\beta_C C_{\bullet}) \gamma X_i \right] \right\} \\
&= \exp \left\{ -\frac{1}{2\Psi} \sum \left[-2\gamma X_i Y_i + X_i^2 \gamma^2 + 2 \sum_{j=1}^p (\beta_j C_{ij}) \gamma X_i \right] \right\} \tag{B23} \\
&\therefore \frac{L^{(C)}(\boldsymbol{\Theta}^{(C)}; \mathbf{Q}^{(C)})}{L_r^{(C)}(\boldsymbol{\Theta}_r^{(C)}; \mathbf{Q}^{(C)})} = \frac{L(\boldsymbol{\Theta}; \mathbf{Q})}{L_r(\boldsymbol{\Theta}_r; \mathbf{Q})}.
\end{aligned}$$

This establishes that the LRT statistic for testing the focal parameter γ is equivalent between the original uncollapsed model with p contextual predictors and the collapsed model with only one composite contextual covariate.

Received January 2, 2025
Revision received June 2, 2025
Accepted June 3, 2025 ■