Too Small to Succeed: Small Samples and the *p* Value Problem

Miguel I. Aguirre-Urreta
Florida International University

Mikko Rönkkö
University of Jyväskylä

Cameron N. McIntosh
Statistics Canada

Acknowledgments

Mikko Rönkkö acknowledges the Academy of Finland grant number 311309.

Abstract

Determining an appropriate sample size is a critical planning decision in quantitative empirical research. In recent years, there has been a growing concern that researchers have excessively focused on statistical significance in large sample studies to the detriment of effect sizes. This research focuses on a related concern at the other end of the spectrum. We argue that a combination of bias in significant estimates obtained from small samples (compared to their population values) and an editorial preference for the publication of significant results compound to produce marked bias in published small sample studies. We then present a simulation study covering a variety of statistical techniques commonly used to examine structural equation models with latent variables. Our results support our contention that significant results obtained from small samples are likely biased and should be considered with skepticism. We also argue for the need to provide a priori power analyses to understand the behavior of parameter estimates under the small sample conditions we examine.

Keywords: Statistical Significance; Small Samples; Partial Least Squares (PLS); Regression; Structural Equation Modeling; Simulation; Estimation Bias; Publication Bias.

Introduction

In recent years, there has been a renewed interest in sample size in Information Systems. On one end of the sample size spectrum, Lin, Lucas, and Shmueli (2013) cautioned that relying solely on significance tests (i.e., *p* values) when interpreting findings from large samples is inadvisable because even trivially small effects with no practical relevance whatsoever will be statistically significant. Instead, Lin et al. (2013) strongly recommended that results be interpreted based on effect size measures, confidence intervals, and effect plots. Similar calls for interpreting effect sizes have recently been made elsewhere in IS (Aguirre-Urreta & Rönkkö, 2018) and other disciplines (Bettis et al., 2016; Cumming, 2014; Kelley & Preacher, 2012).

On the other end of the spectrum, research focusing on small sample scenarios has compared various statistical techniques to determine which is best suited – or whether there are any noticeable differences – when the sample size is small, in the order of 100 observations or less. However, the more foundational issue of whether statistical techniques should be used when sample sizes are small has received less attention. Within the IS field, the use of small samples has primarily been studied in the context of structural equation models with latent variables, which are very common models in the discipline (Goodhue et al., 2012; Ringle et al., 2012). For these types of models, the Partial Least Squares (PLS) (Wold, 1982) has been traditionally used to address small sample scenarios in IS research (Goodhue et al., 2012)[1] because it is believed to be more suitable with small samples than other common estimators (e.g., maximum-likelihood estimators; Chin & Newsted, 1999; Gefen et al., 2011; Hair et al., 2019, Chapter 13, 2021; Peng & Lai, 2012; Reinartz et al., 2009). Goodhue et al. (2012) challenged this status quo by pointing out that the claims about small sample advantages lacked specificity. They then demonstrated that PLS was, in fact, more biased than structural equation modeling with maximum likelihood estimation (SEM) when the same model was estimated with both techniques. Consequently, some recent articles have started to argue against using a small sample size to justify using PLS (Benitez et al., 2020; Henseler et al., 2016; Rigdon, 2016; Rönkkö, McIntosh, Antonakis, et al., 2016)[2].

We contribute to this discussion by stepping back and asking a more fundamental research question: Can any valid inferences be made when very small samples are employed? Small samples lead to two different problems. First, estimates from small samples are highly imprecise (Brand & Bradley, 2016; Marszalek et al., 2011) and thus make the results of an individual study less trustworthy – and less publishable. Second, and perhaps more problematically, the fact that p values influence which studies are published results in a positive bias in published research (Antonakis, 2017b; Gerber & Malhotra, 2008). In turn, the magnitude of this bias, in terms of estimates compared to the population values of the parameters under consideration, is exacerbated in small sample studies (Gerber et al., 2001). Taken together, our research is concerned with the accuracy of the subset of significant estimates, which are the ones most likely to be published, obtained from research studies conducted with small samples. In particular, we focus on research that tests structural models and on the techniques which could be used to estimate those models. With this work, we seek to understand the performance of these techniques when applied to small-sample scenarios, inform authors, reviewers, and editors about the consequences of conducting research with underpowered designs, and provide recommendations to improve the quality of research practice moving forward.

We start our article by explaining why these two mechanisms lead to severe bias in published results obtained from small samples. After that, we present a series of simulations that provide evidence for this effect's existence using the statistical techniques commonly employed in IS research. We conclude our work with recommendations for

individual researchers, reviewers, and editors in our discipline.

Significant Results from Small Samples

Our main argument is that significant estimates obtained from small samples are likely to be biased when compared to their population values. In this section, we present the three key components of this argument. First, we explain the logic of null hypothesis significance testing (NHST). Second, we review multidisciplinary evidence for a strong editorial preference for the publication of significant results, which leads to the "file drawer" problem, also known as publication bias, the active suppression of non-significant findings (Franco et al., 2014). This, in turn, has been shown to lead to questionable research practices where researchers optimize their research designs and analyses to be as likely as possible to produce statistically significant support for the hypotheses (Banks et al., 2016). Third, we explain how these two mechanisms lead to bias in published results and how this bias is magnified in small samples.

The Logic of Null Hypothesis Significance Testing

NHST is the dominant approach for statistical inference in IS and the social sciences more broadly. The starting point of NHST is that we are interested in a well-defined population (e.g., "large corporations" or "employees in large business organizations") but, for practical reasons, we cannot observe the full population; instead, we use only a sample to calculate a statistic of interest (e.g., a path coefficient, a difference between groups, or a correlation). In a well-designed study with a sufficient sample size, the sample estimate should be close to the population value in which the researcher is interested. However, in small samples it is more likely to get a large estimate solely by chance, even if the effect was non-existent in the population. The purpose of NHST is to rule out chance as an explanation for the finding, and thus allow researchers to make claims about effects in the population.

The process of NHST starts with a null hypothesis (H_0). In the case of comparing means between groups or differences between the treatment and control conditions in experimental studies, the null hypothesis is typically that of no difference between the groups in the population (e.g., H_0: $m_1 - m_2 = 0$). In regression or correlational analyses, the typical null hypothesis is that a regression or correlation coefficient between two variables is zero in the population (e.g., H_0: $\beta = 0$). If the null hypothesis is true in the population, observing a large effect in the sample would still be possible, but highly unlikely. NHST quantifies this probability with the *p* value. The *p* value is calculated by first calculating a test statistic – for example, the ratio of a parameter estimate to its standard error referred to as *t* or *z* statistic depending on the content – that reflects how far the estimate is from zero on a standardized metric. Subsequently, the *p* value is calculated as the probability of obtaining the value of the test statistic, or a more extreme value of it, given the null hypothesis is true. If the *p* value falls below a pre-specified significance level, typically 5%, we conclude the sample estimate is *statistically significantly* different from zero, which in turn is interpreted as taken to signify rejection of the null hypothesis. The rejection of the null hypothesis is then interpreted as evidence for the existence of the effect in the population (for further details on NHST and *p* values, see Nickerson, 2000).

The formula of the commonly used *t* statistic reveals two important general features of NHST that are important for our argument. First, because the test statistic is proportional to the magnitude of the estimate, larger estimates are more likely to be statistically significant than smaller estimates. Second, because the standard error that is used as the denominator depends on sample size, reaching statistical significance is more common with larger samples. These two features interact so that when sample size becomes smaller, the estimates must be larger to reach statistical significance. This in itself would not be a major problem, unless statistical significance was also a factor when making publication decisions.

Publication of Significant Results

As explained above, the use of NHST with small samples does not, in itself, lead to bias in published results; the latter also requires a strong preference for significant results in editorial decisions about which research results are published. Two streams of literature provide clear evidence of the existence of this phenomenon. First, throughout the years, several studies have chronicled the existence of editorial preferences related to the significance of reported findings in a variety of fields (Dickersin, 1990; Dwan et al., 2008; Egger & Smith, 1998; Ferguson & Heene, 2012; Greenwald, 1975; Hubbard & Armstrong, 1997; Kepes et al., 2012; Kühberger et al., 2014; Rothstein et al., 2005; Sterling et al., 1995; Thornton & Lee, 2000). One particularly relevant piece of evidence for our argument is that studies with small samples and nonsignificant results are largely missing from the published literature (Chan et al., 2004; Dickersin, 2005; Greenwald, 1975; Ioannidis, 2005; McDaniel et al., 2006; Song, 2010); that is, only small sample studies which have a pattern of significant results make it through the publication process, or are written up and submitted in the first place, or a combination of both effects. The issue of the selectively publishing of studies

– only writing up and reporting studies producing significant findings – has also been addressed under the label of 'file drawer problem' in meta-analytical research (Adams et al., 2017; Bellefontaine & Lee, 2014). While we are not aware of any direct evidence of this problem in IS research, the fact that unpublished studies are routinely sought for in meta-analyses (e.g., Joseph et al., 2007; Sharma & Yetton, 2007; Wu & Lederer, 2009) suggests that there are at least some concerns about this in the field as well.

Second, there is evidence that authors of research studies may use questionable means to ensure that their articles contain small *p* values for their proposed hypotheses. This happens in two ways: The practice of *p*-hacking, which refers to trying out multiple different data manipulations and/or types of statistical analyses, and then only reporting those that yielded significant results (Head et al., 2015; Simonsohn et al., 2014), and the practice of 'HARKing'[3], or 'Hypothesizing After the Results are Known' (Kerr, 1998), which refers to "presenting a post hoc hypothesis in the introduction of the research report as if it were an a priori hypothesis" (p. 197). Examples of mechanisms by which *p*-hacking occurs include conducting analyses midway through a research study to decide whether to continue collecting data, recoding response variables and selectively reporting only some of those, making decisions on the inclusion or exclusion of outliers based on their impact on the findings, excluding, combining, or splitting experimental groups after experimental data have been collected, including or excluding covariates in regression models, or prematurely stopping data exploration upon reaching significant findings (Head et al., 2015). In practice, this means that authors first analyze the data and then develop hypotheses to match the relationships or differences that were found to be statistically significant. The practice of HARKing is a form of outcome reporting bias, where researchers report a subset of statistically significant outcomes, but omit non-significant ones (Copas & Shi, 2001; Hutton & Williamson, 2000; O'Boyle et al., 2017; Williamson et al., 2005).

In his discussion of the incentives behind HARKing, Kerr (1998) emphasizes the primacy of the publication process. In particular, research that both presents a priori hypotheses and then provides confirming evidence regarding those is most likely to fit with the ideal model of a well-conducted research study, and thus more likely to find its way into the published literature. The practice of HARKing can deliver on both counts: by knowing the pattern of results first and then tailoring the writing of the submission to deliver confirming evidence of hypotheses based on that pattern of results, prospective authors are able to produce a submission that improves their chances of being published. As noted by Kerr (1998, p. 205): "evaluative preferences by editors, reviewers and (ultimately) readers implicitly reward HARKing. Furthermore, professional authorities sometimes sanction or even insist upon HARKing. Editors and reviewers will sometime direct authors to HARK". There is also, unfortunately, a body of evidence indicating that researchers are vulnerable to these demands (Bedeian et al., 2010; Fanelli, 2010; Ferguson & Brannick, 2012; Kepes & McDaniel, 2013; Nosek et al., 2012) For similar discussions, see Leamer (1983), Selvin and Stuart (1966), Turner et al (2008), Starbuck (2016), Banks et al (2016), or Bosco et al., (2016).

Significance and Estimation Bias

We will now explain why the logic of null hypothesis significance testing and the preference for publishing significant results, together, lead to severely biased results, and why this problem is even more serious when small samples are employed. We start by considering the simple example of a correlation between two variables using simulated datasets where the population correlation ranges from 0 to .50 and sample sizes from 10 to 70, to focus on the range of what would be commonly considered small samples. Table 1 shows the mean correlation over 10,000 simulated samples for each combination of sample size and population correlation, the share of statistically significant correlation at $\alpha = .05$, and the mean of only those correlations that were significant. Given the known sampling distribution of the correlation coefficient, it is possible to calculate, for a given sample size, the minimum sample correlation that would need to be observed in order for the estimate to be considered significant[4]. For the sample size range employed here, and a significance level of 5%, these are (in absolute magnitude): .632 (N = 10), .444 (N = 20), .361 (N = 30), .312 (N = 40), .279 (N = 50), .254 (N = 60) and .235 (N = 70).

- - - Insert Table 1 about here - - -

The means of the correlations differ strikingly depending on whether all or just statistically significant correlations were considered. For example, when N = 10 and the population correlation is .10, the mean correlation is .102 – a virtually unbiased estimate – when all replications from the simulation are considered but is .396 (a 296% relative bias[5]) when only significant correlations are tallied[6]. While finding a significant correlation with such small sample is unlikely to happen statistical power is just 6%, it is nonetheless possible if researchers prioritize the reporting of significant findings; that is, when we only see published in the literature that small subset of studies where the results turned out to be significant. In addition, when we consider the possibility that researchers may either adjust their procedures to fish for smaller *p* values (*p*-hack) or run a larger set of correlations and then develop hypotheses for those that are significant (HARK), in practice the probability of publishing significant findings from such a small

sample is likely to be larger. Note that this demonstration effectively mimics the consequences of selective reporting or "cherry-picking" significant findings in real applications using small samples – out of a larger pool of conducted studies, one ends up with only a published subset of studies which, when combined for a meta-analysis, would yield a grossly distorted picture of the true effect size[7]. In addition, the simulation also speaks to the replicability crisis in psychology and other social sciences, as one cannot realistically expect to reproduce significant but unlikely results emerging from small samples in subsequent investigations (Stanley & Spence, 2014).

The selection process that leads to bias is demonstrated in Figure 1, which shows the distribution of significant correlations for four extreme conditions. The figure reveals that by only choosing to report significant correlations, we essentially censor all values that are close to zero. The figure reveals that the reason why the bias depends on sample size and population correlation is because the former makes the region of "unpublishable" correlations close to zero narrower and the latter moves the distribution further from this region. The magnitude of the relative bias is thus a function of statistical power. From Table 1, we can see that, as statistical power increases with increasing population correlation and sample size, the relative bias of the significant correlations decreases accordingly. In particular, as simulation conditions improve (that is, larger samples and stronger values for the population correlation), the mean of significant estimates becomes essentially identical to all estimates (e.g., in the case of a correlation of .50 and N = 70, 99.6% of all replications were significant, and therefore essentially no replications would have been omitted from reporting, resulting in an overall unbiased estimate of the true value of the parameter across all research studies in this domain).

- - - Insert Figure 1 about here - - -

As this simple example shows, the significance of an estimate is a function of both the population effect size and the inherent variability in the estimate, which is in turn dependent on sample size. Although the parameter is fixed but typically unknown in the population of interest, the sample size is, at least to some degree, under the control of the researcher. In this research, we are largely concerned with the presence of relative bias in significant estimates that have been obtained from small samples, specifically in the context of structural equation models with latent variables, which are one of the most common modeling approaches in the IS discipline. In particular, we examine here five different statistical approaches – Ordinary Least Squares (OLS, based on sum scores) regression, Disattenuated regression (DR), Partial Least Squares (PLS), Consistent Partial Least Squares (PLSc), and Maximum Likelihood Estimation for Latent Variable Models (SEM[8]) – that could be used to estimate such models. Our choice of modeling approaches was intended to cover the gamut of possible approaches of interest for the estimation of structural models with latent variables. First, we included both OLS regression and PLS, which differ in their approaches to weighting the items which measure each construct in the research model. Whereas equal-weight sum scores are used in OLS regression, in PLS each item is given a different weight in the creation of the composites used to represent the constructs. However, both of these approaches lead to biased estimates for the paths of interest due to the presence of measurement error in the items (Dijkstra & Henseler, 2015; Evermann & Rönkkö, 2021). As a result of this bias, approaches which seek to correct for the effects of attenuation due to measurement error have been developed; for example, Consistent PLS (PLSc) (Dijkstra & Henseler, 2015). With the goal of providing a counterpart to PLSc based on equal weights, we also included a disattenuated version of OLS regression in the group of approaches examined here. Finally, we included SEM as an estimation approach which can model each individual item (and its measurement error) separately and does not rely on a disattenuation (bias-correction) step as part of the process. Taken together, we believe these approaches capture the range of possibilities which may be of interest to IS researchers when seeking to estimate structural equation models with latent variables.

Simulation Design

We conducted a simulation study to examine the accuracy of significant estimates obtained from small samples under a variety of conditions and data analysis techniques. In particular, we employed the base model by Qureshi and Compeau (2009) with the correlation between the two exogenous latent variables set to a medium value of .3 (see Figure 2 for the structural portion of the model).

- - - Insert Figure 2 about here - - -

We otherwise designed the simulation conditions following previously published research. First, the number of indicators measuring each latent variable in the model was either 3, 6 or 9. Second, the loadings relating each indicator to its latent variable were either all 0.70, 0.80, 0.90 or mixed where one third of the loadings at 0.70, one third at 0.80, and the remaining third of the loadings in each condition at 0.90. Finally, sample sizes were 20, 40, 60, 80 or 100. All these were combined in a full factorial design, for a total of 60 different conditions. For each

combination of these conditions, 1,000 replications were generated; for a given replication, when an estimator failed to converge, the data were discarded, and a new random sample was drawn. Therefore, all results presented here are based on 1,000 replications where the estimator for a given technique successfully converged. Where bootstrapping was needed, 1,000 bootstrap resamples were generated for each replication. All data were drawn from a multivariate normal distribution.

All data generation and analyses were conducted in the *R Statistical Environment* (R Core Team, 2021, v. 4.1.0)[9]. The five techniques compared here were implemented as follows. OLS regression, PLS and PLSc were implemented with the *matrixpls* (Rönkkö, 2021, v. 1.0.11) package. Disattenuated regression (DR) was also implemented with the *matrixpls* package; in particular, the *minres* estimator was used to obtain the factor loadings used in the disattenuation of the correlation matrix[10]. Finally, all SEM analyses were conducted with the *lavaan* (Rosseel, 2012, v. 0.6-8) package.

For OLS and DR statistical significance was assessed based on *t* tests by taking the ratio of the estimate to its standard error and considering an estimate significantly different from zero when the absolute value of the ratio was larger than a critical value from the *t* distribution[11]; in both cases, the standard errors were calculated using the closed-form equations[12]. For PLS and PLSc, an estimate was considered significant when zero was not included in a 95% percentile confidence interval, following existing recommendations in this regard (Aguirre-Urreta & Rönkkö, 2018; Dijkstra & Henseler, 2015). The SEM estimates were tested with *z* tests. In all cases, data generation was done with the *simsem* (Pornprasertmanit et al., 2020, v. 0.5-15) package[13].

Results

Statistical Power

As would be expected, it is very unlikely to obtain significant results with small samples[14] (sample size being one of the key determinants of the likelihood to reach significance, the size of the effect under consideration being the other). Given that there are four non-zero effects in our population model, we first consider the percentage of replications, for each statistical approach and simulation condition, which resulted in the estimate of all four parameters being found significant. We then do the same with a more relaxed rule of any three out of a total of four effects being found significant (see Appendix A for the results of all analyses where all estimates where significant, and Appendix B for the same results where three or more estimates were significant). We should note that, since all four effects are non-zero in the population, a finding of four significant effects would be an ideal outcome for an empirical researcher, as any other outcome would result in an inference error (e.g., concluding an effect is not significantly different from zero when the true value of the effect is non-zero in the population, a Type II error).

As our results show, the likelihood of all four estimates being found significant is, overall, quite small, which in itself should be a major concern for researchers; that is, the overall statistical power of studies under these conditions is very low. Given that a major portion of the time and resources invested in conducting empirical research are allocated to the design and validation of the data collection instrument, collecting only a small amount of data makes it very unlikely that anything publishable will be obtained. The overall significance rate (all four effects significant) was 1.13% for OLS, 2.95% for DR, 1.96% for PLS, 2.19% for SEM, and 1.74% for PLSc (note that, for the cases of SEM and PLSc, the conditions were no replications converged have been removed from the base for the purpose of these calculations; these results, and those discussed next, are based solely on the number of conditions in which 1,000 replications were successfully obtained). These results, however, vary markedly based on levels of each of the simulation design conditions included in this research (see Table 2).

- - - Insert Table 2 about here - - -

Even so, statistical power rates were very low throughout all statistical approaches and conditions (and certainly much lower than the commonly recommended .8 level) (e.g., Cohen, 1988, 1992). Considering every particular combination of conditions in our simulation, the highest power for OLS was 2.9% (sample size of 100, 9 indicators, .9 loadings), for DR was 8.3% (sample size of 100, 3 indicators, .7 loadings), for PLS was 6.5% (sample size of 100, 6 indicators, mixed loadings), for SEM was 4.2% (sample size of 100, 6 indicators, mixed loadings), and for PLSc was 4.5% (sample size of 100, 9 indicators, .7 loadings). In summary, statistical power levels were very low for all approaches and simulation conditions, making it very unlikely that a researcher would reach the correct conclusion that all four effects were different from zero. In the small number of replications where that was the case, however, the resulting estimates were severely biased, which is the main result of interest in this research. We discuss these findings next.

Bias from Significant Results

We now turn to the main focus of this research, the extent to which significant results obtained from small samples are biased, when compared to the true population values of the parameters under consideration. To provide a more comprehensive picture of the issue, we also report on additional analyses that showcase the degree of relative bias as a function of number of indicators and strength of loadings. All the following results and discussions refer only to those replications where all four path estimates were considered significant, as described above, which is the main issue under consideration here.

- - - Insert Figure 3 about here - - -

Figure 3 shows average relative bias, compared to the true population values of each parameter, by statistical approach and over different levels of sample size. Estimation bias is clearly dependent on the population magnitude of the parameter under consideration. For the strongest of the paths examined here, the C → D relationship (population value .6), relative bias is small, but not negligible. For this path, relative bias is in the -10% to 20% range, depending on the particular combination of sample size and statistical approach under consideration. Bias is more noticeable for the second strongest path, capturing the A → D relationship (population value .35). As the corresponding panel in Figure 3 shows, particularly for the smaller sample sizes considered here, there is a marked positive relative bias. For the condition with N = 20, this relative bias is 54.5% for OLS, 29.7% for PLS, 40.74% for PLSc, 53.56% for DR, and 59.3% for SEM. Though estimates improve with increasing sample size, there is still noticeable bias with N = 40.

For the path B → C (population value .2), Figure 3 shows major positive relative bias, which does improve with increasing sample size, but remains very marked even at N = 100. For this path, relative bias is in the 85% to 180% range for the case of N = 20 (155.0% for OLS, 85.8% for PLS, 133.1% for PLSc, 175.6% for DR, and 169.0% for SEM). For the case of N = 100, relative bias remains in the 20% to 50% range (39.0% for OLS, 20.6% for PLS, 35.3% for PLSc, 50.8% for DR, and 53.2% for SEM), which would still be considered unacceptable for publication.

Finally, the weakest of the paths considered here (for the A → C relationship, population value .05) shows some interesting results (see the leftmost panel of Figure 3). In this case, the population value of the path is close to zero. Coupled with our focus on small samples, there is sufficient variability in the sampled data for it to be possible to obtain negative estimates for this parameter. Our results show that some of the statistical approaches considered here were only able to flag as significant estimates that were very negatively biased, whereas other approaches only did so when the estimates were very positively biased. In either case, for the sample size range considered here, none of the examined techniques produced even remotely accurate estimates of this parameter. More precisely, for the case of N = 20 average relative bias was -895.0% for OLS, 502.9% for PLS, 163.5% for PLSc, -910.4% for DR, and -746.6% for SEM. To put these results into context, consider that a path with a population value of .05, as is the case here, would appear to be significant and with an estimate of -.45 when the relative bias is of the order discussed here (e.g., -900%). This is clearly an unacceptable outcome.

- - - Insert Figure 4 about here - - -

Figure 4 shows the same results over the number of indicators used to measure each of the four latent variables in the research model. The results are similar to those previously discussed. The degree of average relative bias worsens with increasingly weaker paths in the research model. For the weakest of the paths examined here, we also see a similar effect of either markedly positive or markedly negative bias estimates. In this case, the magnitudes appear smaller (though still unacceptably large) as the extreme results observed in the preceding discussion are collapsed over levels of indicators. These results also show that the number of indicators used to measure the latent variables does not have a noticeable effect on whether extreme (i.e., biased) and significant estimates are obtained.

- - - Insert Figure 5 about here - - -

Finally, Figure 5 shows the results over different loading patterns. These are either homogenous (all loadings at .7, .8, or .9) or mixed (as noted above, one third at .7, one third at .8, and one third at .9). Whereas the base levels of relative bias for each regression path are as previously discussed, Figure 5 also shows that measurement quality, in terms of the strengths of the loadings relating each indicator to its latent variable, have a noticeable effect on estimation accuracy when only significant paths are considered. In all cases, relative bias is most pronounced when the loadings are weakest, in the condition of homogenous loadings at .7. Estimation accuracy improves as loading strength increases; the condition with mixed loadings is in between homogenous loadings of .8 and .9 in terms of estimate accuracy. Whereas the number of indicators used to measure each latent variable had a limited impact on estimation accuracy, measurement quality in the form of more reliable indicators does affect the accuracy of the resulting estimates. Even so, marked bias remains throughout our simulation results even when loadings are very strong, at .9 in the population.

Discussion and Conclusion

The conduct of quantitative empirical research requires several important planning and design decisions to ensure the quality and validity of the findings. These include the choice of sampling frame and methodology, the items to be used to measure the included constructs, the research design to be used (e.g., survey, experiment, etc.), statistical analysis approach, and the necessary sample size, to name a few. The latter – sample size – is a key component of statistical power analyses, and the only one which is, at least to some extent, under the control of the researchers conducting a study (the other two components of a power analysis being the choice of threshold for statistical significance and the effect size of the relationship of interest).

The seminal work with regards to statistical power in the discipline was conducted by Baroudi and Orlikowski (1989). In this research, the authors first reviewed the fundamentals of statistical power, its calculation, and recommendations to be followed by researchers, and then examined the status quo of power in the discipline at the time. They concluded that, at the time, not only were discussions of power and power analysis not common in the discipline, but also that a large proportion of the research conducted in the field at the time was severely underpowered. The authors saw this as a negative in terms of allocating research resources to studies unlikely to yield significant findings, which may also lead researchers to prematurely abandon what could be potentially interesting areas of research. Baroudi and Orlikowkski (1989) concluded their research with a discussion of different ways in which power analyses could be improved including, but not limited to, increasing sample size. However, what was not recognized at the time, is that significant results resulting from severely underpowered studies – which are the main issue of concern in our research, and which Baroudi and Orlikowski (1989) noted were common in the discipline – could lead to markedly biased estimates for the relationships of interest, as we show with our work.

Subsequent work by Goodhue et al (2012) focused specifically on the issue of whether PLS, as has been frequently been argued, had any particular advantages over other techniques in small sample scenarios (or, less relevant to our research, when data are sampled from non-normal distributions). Their interest was in comparing the performance of PLS and a maximum likelihood estimator for latent variable models for the estimation of structural equation models with latent variables, and the extent to which there was evidence supporting the purported advantages of PLS when applied to small samples. Their results indicated no particular advantage of PLS over other techniques when sample size was small, with the caveat (observed in our research as well) that covariance-based estimators may have a harder time converging with very small samples. Although the authors specifically discussed the implications of their results for published research which found significant results, using PLS, from small samples, their concern was solely on whether PLS or other techniques would be more or less likely to reach the same conclusion – finding significance – under the same conditions (that is, whether PLS or other techniques were more or less powerful to detect those) and on whether false positives, above the 5% threshold, would be more or less common for the different techniques they examined under small sample scenarios. Goodhue et al (2012), however, did not consider the accuracy of the estimates in their research, as their main focus was on statistical power levels. Our work therefore expands on this research by examining not only a larger number of statistical techniques under a variety of conditions, but more importantly considering the accuracy of significant estimates obtained from small samples. As we argue throughout this research, and support with evidence from our simulations, this should be a major issue of concern for quantitative researchers, reviewers, and editors.

Lin et al (2013) cautioned researchers about the interpretation of significant effects arising from the use of very large samples, which have recently been made possible with the new availability of extensive datasets, such as those involving online transactional data. Their concern was that, when very large samples are employed, even minor or trivial effects are likely to be significant, and thus interpretation should shift from merely stating that an effect is significantly different from zero to considering the magnitude or practical relevance of such an effect, i.e., the effect size. In this research, our concern is with the other end of the continuum of sample sizes employed in IS research. Whereas the focus of past research in this area has been which of the many available techniques is more or less appropriate for the study of small samples (e.g, Chin & Newsted, 1999; Goodhue et al., 2012), limited consideration has been placed on the outcome of those studies. Given strong institutional and normative pressures favoring the publication of research containing exclusively or mostly significant findings, it is therefore important to examine the subset of outcomes obtained from small samples which are likely to successfully navigate the review process, as those are representative of what can be found in our journals. As a result, we focus here on the accuracy of estimates obtained from small samples when those estimates conform to patterns of significance that make them more likely to lead to publication.

In this research, we defined a population model with a variety of effect sizes, in terms of both patterns and magnitude. We then generated data for those for several different measurement conditions, varying in terms of the number of indicators with which each latent variable was measured, their strength, and sample size (which is the focus of this

research). For each of these conditions we generated many replications and subjected those to analysis with five different statistical techniques. Of those, we only focused on the replications that conformed to a pattern of results that would be consistent with the true nature of the underlying relationships – that is, significant results for the non-zero paths in the population, and non-significant results for the zero path in the population. These are the replications that would produce results in line with theoretical expectations, and thus more likely to eventually appear in published form.

As our results show, these are a relatively small subset of all possible replications, particularly at the smaller end of the sample size range. This is in line with expectations given the known workings of statistical power such that, everything else being equal, more replications will be significant with larger samples than with smaller samples. However, when we only consider that subset of publishable replications, the results are quite discouraging. When significant estimates are obtained from small samples, even if those conform to the true pattern of relationships in the population, the estimates of those relationships are likely to be markedly biased when compared to their true population values. Whereas the magnitude of the bias is dependent on the statistical technique employed and the true pattern and magnitude of the relationships in the population, such that stronger effects are less affected than weaker effects, it is quite evident from our results that none of the commonly used statistical techniques are immune from the bias discussed here.

The results presented here are subject to the limitations shared by any research that employs simulations; in particular, the choice of population models and research conditions examined in the study. Whereas the structure of the population model employed here is relatively simple, it is not unlike many research models in the IS discipline, which include only a few incoming paths (only one in many cases) on any given portion of a research model (Goodhue et al., 2012). The strength of the loadings relating indicators to their latent variables as well as the number of indicators used to measure each latent variable are well within what would be considered acceptable or better in light of commonly used validation guidelines in effect in IS research. Taken together, while our results could certainly be replicated and extended by focusing on different model, different magnitudes for the relationships involved, different measurement conditions, etc. (and we encourage researchers to undertake these expansions to further refine our findings), it is clear from our results that there is reason to be concerned about the interpretation of research findings obtained from underpowered studies.

Does power analysis address these issues?

Given that the underlying cause of the bias that occurs when only studies with statistically significant results are published is that studies using small samples are underpowered, it would be ideal to state that power analyses could be used to address the problem presented here. Unfortunately, this is not the case. While power analyses are useful when planning the required sample size for the study, they are less informative after a study has been conducted (Hoenig & Heisey, 2001). There are two problems with post-hoc power analysis in this case. First, if a researcher first estimates a model, and then specifies a power analysis based on the estimated effect size, the estimated (observed) power provides no additional information over that provided by the *p* value. Second, the problem that we address here is that non-significant findings from small samples are not publishable in IS research. There is no reason to believe that adding a power analysis to a study demonstrating that a study is underpowered would make a non-significant study more publishable; in fact, it may very well have the opposite effect (as non-significant findings from underpowered studies may be deemed not informative).

Nevertheless, in order to avoid conducting underpowered studies in the first place, we recommend that, if data are difficult to obtain, or there is some other reason making a small sample study the only feasible alternative, a power analysis should be used to determine an appropriate sample size for the study when assuming effect sizes for the true population parameters that are consistent with those of theory or past research. Under these conditions, significant results obtained from severely underpowered studies should be met with a strong dose of skepticism. Doing the same for the particular research model and relationships of interest, and under measurement conditions similar to those in the focal research study, would provide evidence of the degree to which the issues observed here should be a concern. There are approaches that are appropriate for each of the techniques discussed here (for example, Aguirre-Urreta & Rönkkö, 2015 for PLS and PLSc; or Muthén & Muthén, 2002 for SEM) and therefore the conduct of such analyses should not prove to be overly complex, and would anyways be recommended in the design and conduct of any quantitative study.

We would argue that conducting a power analysis as part of research design and planning should be considered a standard practice when conducting quantitative research, and we are of course not the first, in the IS discipline, to note this should be the case (Baroudi & Orlikowski, 1989; Carte & Russell, 2003; Goodhue et al., 2007; Straub, 1989). Earlier in the development of the discipline, the concern regarding statistical power had to do with designing

studies with sufficient power so to both maximize the chances of detecting an effect of interest and also aiding in its interpretation. For example, Straub (1989, p. 152) notes "non-significant results from tests with low power, i.e., a probability of less than 80 percent that the null hypothesis has been correctly rejected […] are inconclusive and do not indicate that the effect is truly not present", and we discuss this issue (interpretation and reporting of non-significant results) in the next section.

In this research, however, we are concerned about the scenario where a study is underpowered, yet significant results are obtained, and with the accuracy (or bias) of those results. While, as noted above, there is a history in the discipline of advising power analyses should be conducted and reported, it is nonetheless the case that studies continue to be published with small samples, well within the range of sample sizes examined here, and which we show will result in possible markedly biased estimates of the relationships of interest. For example, Goodhue et al (2012) noted that 13% of the studies which employed some form of path analysis, published in the premier journals of the discipline in the 2006-2010 period, had sample sizes smaller than 80 which they argued, as do we here as well, are insufficient[15]; yet these studies successfully navigated a review process where the conduct of power analyses, which would have flagged those as being markedly underpowered, has been strongly recommended for decades. One possible explanation is that the negative consequences, in terms of accuracy and bias, of significant results obtained from underpowered studies are not well understood. With our research, we hope to contribute to this discourse to show there is reason to be concerned about these findings.

What can be done about the issue?

Based on our research and findings, we recommend that editors, reviewers, and readers be more skeptical of significant results obtained from small samples than has been the case in the past. But what should be done in the future? Generally, outside of the IS discipline, there is an increasing understanding that *p* values are not a good metric for assessing the quality of a research study or the importance of its findings (Wasserstein et al., 2019). To this end, many journals outside IS have started to emphasize criteria other than *p* values. For example, *Strategic Management Journal*, a leading journal in management, has an explicit policy of also publishing non-significant results (Bettis et al., 2016). Nevertheless, there may still be a long road ahead to educate reviewers to not only pay attention to *p* values when evaluating studies. As another example, there is an initiative where some journals are sending studies to review with the results blinded so that the reviewers must judge the importance of the research question, quality of the theory, and strength of the research design instead of looking at a particular set of results that could have been occurred only by chance in underpowered studies (e.g., Antonakis, 2017a). We hope that our article can raise awareness of the bias due to small samples and of the possible remedies applied in other adjacent fields to address the issue at the journal level.

While many of the most impactful decisions to address the small-sample problem are done at the level of the journal, which decides whether a study is publishable, we argue that there are also actions that individual researchers can take to address this issue. Designing studies for sufficient statistical power and spending more resources on data collection to ensure sufficient power is an obvious solution. However, we recognize that this is not always possible, particularly when studying organizations or teams within organizations, where the target population may be small. Thus, non-significant results from underpowered studies are sometimes inevitable. However, even in these scenarios, a focus on accurate parameter estimation and the presentation of descriptive information, absent significance testing, may still be warranted (Valentine et al., 2015).

To make small-sample studies publishable – and consequently decrease the bias caused by publishing only those underpowered studies which are statistically significant – we suggest that these studies de-emphasize *p* values or even leave them out. For example, instead of reporting $b = 0.2$ ($p = .11$) and declaring a hypothesis not supported, a researcher could report $b = 0.2$ (95% CI:[-0.4, 0.44]) explaining that the best estimate of the effect is 0.2 and that there is great uncertainty in the result, given the width of the confidence interval surrounding the estimate. This approach would thus involve shifting the focus from a yes/no decision based on *p* values to the assessment of the effect size and its uncertainty (Cumming, 2011; Wasserstein et al., 2019). Confidence intervals can also be graphically represented with error bars around an estimate, which visually conveys the degree of uncertainty in that estimate in a way that *p* values cannot replicate, thus lending additional interpretation power to the use of confidence intervals over and above a binary significant/non-significant decision (Fidler & Loftus, 2009). While such tactics may not convince all reviewers, it is consistent with calls to focus more on assessing the magnitude of the effects in IS (Aguirre-Urreta & Rönkkö, 2018) and in other disciplines as well (Bettis et al., 2016; Cumming, 2014; Kelley & Preacher, 2012). For example, as noted above, *Strategic Management Journal* recently revised its editorial policies to 1) publish and welcome submissions of replication studies, 2) publish and welcome submissions which include non-significant results, 3) no longer accept for publication submissions which refer to cutoff values for statistical significance, now requiring standard errors or exact *p* values (or both) as well as their interpretation, and 4) require

that accepted submissions explicitly discuss and interpret effect sizes for the estimated coefficients (Bettis et al., 2016). Other journals, such as *Basic and Applied Social Psychology* (Trafimow & Marks, 2015), *Epidemiology* (Fidler et al., 2004), or *Political Analysis* (Gill, 2018), have outright banned the use of *p* values in their publications.

More generally, while our discussion here centers on significant results obtained from small samples (which, as shown in this research, are likely to be very biased), our results also show that in most instances, small samples will lead to non-significant results, as a result of limited statistical power to detect effects when those are indeed present. The vast majority of these results, in turn, will not make it into a publication (Franco et al., 2014). As a result, it is worth considering what options are available for researchers faced with non-significant findings (see Mehler et al., 2019 for a more detailed discussion of these).

Most notably, researchers should refrain from interpreting a non-significant result as evidence of the absence of an effect (e.g., the null hypothesis of no effect being true), particularly in small sample scenarios, which are severely underpowered. In these scenarios, a non-significant result is compatible with a non-zero effect which was overlooked due to lack of statistical power. To distinguish between a meaningful but overlooked effect, and one that is either negligible or absent we should first determine, based on prior research, theory, and the goals of the study, what is the minimum effect size that would be considered meaningful for the research. Then, *equivalence tests* (Lakens et al., 2018) can be used to determine whether there is enough evidence to determine whether the observed effect could be considered negligible, and thus treated, for all practical purposes, as being absent. Ideally, the expectation of possibly conducting an equivalence test, and thus determining minimum meaningful effect sizes, etc. should be part of the research design and planning (and sample size determination) process. Alternative approaches, based on Bayesian statistics, are also possible, such as the use of regions of practical equivalence (ROPE, Kruschke, 2011) or hypothesis testing based on Bayes factors (Lakens et al., 2020).

Final words

It would be possible to argue that there are particular research settings or populations of interest where obtaining large enough samples is not possible practical, or economically feasible, and we would concede that to be the case in some very specific research scenarios. At the same time, however, researchers should then be mindful of seeking to apply statistical techniques developed on the assumption of large samples to obtain stable and accurate results to populations that do not conform to those requirements. In the past it has been argued that PLS can be a technique that would be particularly applicable to those research scenarios (e.g, Chin & Newsted, 1999; Majchrzak et al., 2005); there is, however, mounting evidence (e.g., Evermann & Rönkkö, 2021; Goodhue et al., 2012, p. 998; Rönkkö et al., 2015, p. 81; Rönkkö & Evermann, 2013, Myth 5, pp. 440-442), including our own research findings, showing that is not the case. For example, Goodhue et al (2012, p. 998) noted the belief among MIS researchers that PLS has special powers at small sample size or with non-normal distributions is strongly and widely held in the MIS research community. Our study, however, found no advantage of PLS over the other techniques for non-normal data or for small sample size". Similarly, Rönkkö and Evermann (2013) listed a number of studies expressing the belief that PLS does not require a large sample size, and categorized those as one of the myths about PLS highlighted in their research ("Myth 5: PLS Has Minimal Requirements on Sample Size"). Going one step forward, however, our research highlights that none of the commonly used techniques in IS research are appropriate for application in these scenarios, and thus calls for a rethinking of our research designs and analytical approaches when faced with small samples, and where larger samples cannot be feasibly obtained.

Notes

[1] In particular, Goodhue et al (2012) note that out of 188 studies in the premier journals of the IS discipline in the 2006-2010 period that employed some form of path analysis, 49% of those chose PLS, 35% of those referred to advantages of PLS when applied to small samples, and 14% analyzed samples with less than 80 participants. Some extreme examples of these include Kahai and Cooper (2003, $N = 31$), Malhotra, Gosain and El Sawy (2007, $N = 41$) and Majchrzak et al. (2005, $N = 17$)

[2] Some of the authors who otherwise argue that PLS has advantages when it comes to working with small samples have also acknowledged that those presumed advantages have been abused in the past and that is a valid criticism of some extant applications of PLS. At the same time, for scenarios where populations may be small (e.g., under 100), they still argue that PLS is the only structural approach which can yield meaningful results (Hair et al, 2019, p. 771). Our concern in this research is whether these techniques should be employed at all under those circumstances.

[3] We note that there may be scenarios where HARKing can produce valid and important research findings, but the fact that a hypothesis was developed after the results were known must be reported transparently (Hollenbeck & Wright, 2017).

[4] This can be done by solving for the critical value of a two-tailed *t*-test, where the statistic would be $t = r\sqrt{\frac{N-2}{1-r^2}}$. This can be rewritten as follows (where *df* = *N* – *2*) to solve for r: $r = \sqrt{\frac{\frac{t^2}{df}}{\frac{t^2}{df}+1}}$. For a given sample size (and associated degrees of freedom), the critical value of the *t*-distribution can be obtained and then used to calculate the critical value of the correlation coefficient. For the case of N = 10 and alpha = 5%, the critical value for a *t*-distribution with 8 degrees of freedom is 2.3066. Using this critical value, we can solve for the critical value of *r* of 0.632 for N = 10 and alpha 5%. Other critical values can be similarly calculated.

[5] All percentage bias reported in this research are calculated on a relative basis, comparing the average of the sample estimates against the population value of the parameter, divided by the latter; specifically: $relative\ bias = \frac{(average\ of\ estimates - population\ value)}{population\ value}$ for a given condition.

[6] The bias of an estimate is a measure of over- and underestimation in the long run, over many replications. Selective reporting of only significant results leads to a tally of the estimates with the largest estimation errors (as only those would reach significance). When considered over multiple replications, this leads to overall biased estimates (that is, $E[\hat{\beta}_{sig}] \neq \beta$,, where β is the population parameter and $\hat{\beta}_{sig}$ those estimates that have been found significant). In other words, the average of only significant estimates will be biased when compared to the true population value of a given parameter.

[7] There is a growing concern in the psychology discipline about the ability, or lack thereof, to replicate past findings. Since unsuccessful replications – those that exhibit negative findings, where the original relationship of interest is not significant in the replication – face an uphill battle in getting published, this compounds the problem (Yong, 2012).

[8] Throughout this research, the SEM label refers to the statistical approach of using maximum likelihood estimation for latent variable models. In the IS literature, this has also been referred to as CBSEM (Covariance-Based SEM) to distinguish it from PLS-based approaches for the estimation of structural equation models with latent variables.

[9] All data generation and analysis code is available for download at a repository accessible through the Open Science Framework (https://osf.io/93azp).

[10] Disattenuated regression was specified in the *matrixpls* package as follows. First, composites with unit weights were created as aggregates of all items measuring a construct (this is unlike PLS, where weighted aggregates are created). Second, reliabilities for those composites were obtained for each composite. This was done by estimating loadings for each item, one block of items at a time, using the *minres* estimator for the loadings. The estimates of the loadings were then used to calculate the reliabilities for each composite. Third, the correlation matrix of the original composites was disattenuated using these composite reliabilities. Finally, from the disattenuated correlation matrix, estimates of the paths in the research model were calculated.

[11] The degree of freedom for the *t* test in OLS and DR is *n – k – 1*, where *n* is the sample size and *k* is the number of predictors. The PLS literature provides various different recommendations for number of degrees of freedom (Rönkkö, McIntosh, Antonakis, et al., 2016). To keep the results comparable across techniques, we used the same degrees of freedom for all techniques.

[12] In their research on hypothesis testing using factor score regression, Devlieger et al (2016) noted that the common formula for the standard error may not be accurate when bias correction is used. We examined the degree to which bias may be introduced in this case, as well as reran our simulation using a bootstrapping approach, similar to the one employed with PLS and PLSc. Our results regarding the presence of bias on significant estimates obtained from small samples, which is the main issue of interest here, are not qualitatively different. All these results are reported in Appendix D. While we believe the development and validation of an accurate standard error measure for the approach is a worthwhile endeavor, it is beyond the scope of our current research (but would be an integral step in the development of disattenuated regression as a complete alternative for the estimation and testing of structural models). We thank an anonymous reviewer for alerting us to the work of Devlieger et al (2016).

[13] We used two-tailed tests for significance throughout all our research. For the same level of statistical significance (e.g., the same alpha), the only difference between these and one-tailed tests lies in the distribution of the critical area of rejection on one or both ends of the distribution of the test statistic. Though two-tailed tests do not involve a specific direction for an effect to be considered significant, it is often the case that there is an expected direction which was hypothesized a prior (e.g., a path is expected to be positive or negative), which would point to the use of a one-tailed test for establishing significance. However, two-tailed tests are used in most applications, whether there is an expected direction for an effect, and they are the default (and, in many cases, the only option) in most statistical software packages. The use of one-tailed tests is controversial and sometimes considered cheating (Abelson, 1995, p. 55).

[14] Researchers working with small samples should also be mindful of convergence and admissibility issues. All results presented here come from 1,000 samples that converged to admissible solutions. When a particular replication did not successfully converge for a given technique, we discarded it and drew a new one. Convergence rates, however, are informative for researchers. Both OLS and DR had perfect convergence, whereas PLS had a 99.2% effectiveness for a single condition and 100% otherwise. SEM had trouble converging with small samples (e.g., N = 20) with an average convergence of 40.3% in those conditions; convergence rates improved markedly as sample size increased. Finally, PLSc suffered from severe admissibility issues in some conditions. For example, the condition with a sample size of 20, 9 indicators, and mixed loadings required 49,047 replications to obtain 1,000 acceptable ones. Similarly, the condition with a sample size of 20, 3

indicators, and loadings at .9 required 46,603 replications. The lowest number of replications required to obtain a set of 1,000 acceptable ones was 1,514 (sample size of 100, 3 indicators, .7 loadings), for a convergence effectiveness of 66%. Similar results were published by Huang (2013), who reported that 39% of her PLSc replications of her small model were inadmissible at N=200, and by Rönkkö, McIntosh, and Aguirre-Urreta (2016). While it is always possible to draw a new sample in a simulation, researchers have only a single sample of data, and are thus advised to consider how likely it is that a given statistical approach would result in at least acceptable – from a convergence and admissibility standpoint – estimates.

[15] Future research could focus on conducting a formal review of research practices in the IS discipline with regards to statistical power analyses, sample size determination, and related research practices, thus updating the work of Goodhue et al (2012) for the case of structural models, but also considering other quantitative research designs as well. In the conduct of our research, we came across two additional exemplars from recently published work. Jenkin et al (2019) conducted a model test, using regression analysis, with a sample size of 13, which the authors acknowledged resulted in low statistical power, yet results were significant nonetheless. In another study, Serrano and Karahanna (2016) noted that low statistical power was a likely cause of lack of significance in their testing of moderation hypotheses, which goes against the recommendations that power analyses be conducted and reported as standard practice in the discipline.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Lawrence Erlbaum Associates.

Adams, R. J., Smart, P., & Huff, A. S. (2017). Shades of grey: Guidelines for working with the grey literature in systematic reviews for management and organizational studies. *International Journal of Management Reviews*, *19*(4), 432–454.

Aguirre-Urreta, M. I., & Rönkkö, M. (2015). Sample size determination and statistical power analysis in PLS using R: An annotated tutorial. *Communications of the Association for Information Systems*, *36*(1). http://aisel.aisnet.org/cais/vol36/iss1/3

Aguirre-Urreta, M. I., & Rönkkö, M. (2018). Statistical inference with PLSc using bootstrap confidence intervals. *MIS Quarterly*, *42*(3), 1001-1020.

Antonakis, J. (2017a). Editorial: The future of The Leadership Quarterly. *Leadership Quarterly*, *28*(1), 1–4.

Antonakis, J. (2017b). On doing better science: From thrill of discovery to policy implications. *Leadership Quarterly, 28*(1), 5-21.

Banks, G., O'Boyle, E., Pollack, J., White, C., Batchelor, J., Whelpley, C., Abston, K., Bennett, A., & Adkins, C. (2016). Questions about questionable research practices in the field of management. *Journal of Management*, *42*(1), 5–20.

Baroudi, J., & Orlikowski, W. (1989). The problem of statistical power in mis research. *MIS Quarterly*, *13*(1), 87–106.

Bedeian, A., Taylor, S., & Miller, A. (2010). Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning and Education*, *9*(4), 715–725.

Bellefontaine, S., & Lee, C. (2014). Between Black and White: Examining Grey Literature in Meta-analyses of Psychological Research. *Journal of Child and Family Studies*, *23*(8), 1378–1388.

Benitez, J., Henseler, J., Castillo, A., & Schuberth, F. (2020). How to perform and report an impactful analysis using partial least squares: Guidelines for confirmatory and explanatory IS research. *Information & Management*, *57*(2), 103–168.

Bettis, R. A., Ethiraj, S., Gambardella, A., Helfat, C., & Mitchell, W. (2016). Creating repeatable cumulative knowledge in strategic management. *Strategic Management Journal*, *37*(2), 257–261.

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology*, *69*(3), 709–750.

Brand, A., & Bradley, M. T. (2016). The precision of effect size estimation from published psychological research: Surveying confidence intervals. *Psychological Reports*, *118*(1), 154–170.

Carte, T., & Russell, C. (2003). In pursuit of moderation: Nine common errors and their solutions. *MIS Quarterly*, *27*(3), 479–501.

Chan, A., Hrobjartsson, A., Haahr, M., Gøtzsche, P., & Altman, D. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association*, *291*, 2457–2465.

Chin, W., & Newsted, P. (1999). Structural equation modeling analysis with small samples using partial least squares. In R. Hoyle (Ed.), *Statistical Strategies for Small Sample Research* (pp. 307–341). Sage Publications.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Copas, J., & Shi, J. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research*, *10*(4), 251–265.

Cumming, G. (2011). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29.

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis Testing Using Factor Score Regression: A Comparison of Four Methods. *Educational and Psychological Measurement*, *76*(5), 741–770.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *Journal of the American Medical Association*, *263*, 1385–1389.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. Rothstein, A. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-analysis: Prevention, Assessment, and Adjustments* (pp. 11–34). Wiley.

Dijkstra, T. K., & Henseler, J. (2015). Consistent partial least squares path modeling. *MIS Quarterly*, *39*(2), 297–316.

Dwan, K., Altman, D., Arnaiz, J., Bloom, J., Chan, A., Cronin, E., Decullier, E., Easterbrook, P., Von Elm, E., Gamble, C., Ghersi, D., Ioannidis, J., Simes, J., & Williamson, P. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One*, *3*(8).

Egger, M., & Smith, G. (1998). Bias in location and selection of studies. *British Medical Journal*, *316*, 61–66.

Evermann, J., & Rönkkö, M. (2021). Recent developments in PLS. *Communications of the Association for Information Systems*, 44.

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from us states data. *PLoS One*, *5*(4).

Ferguson, C., & Brannick, M. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*, 120–128.

Ferguson, C., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561.

Fidler, F., & Loftus, G. (2009). Why figures with error bars should replace *p* values. *Journal of Psychology*, *217*(1), 27–37.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*, 119–126.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Gefen, D., Rigdon, E. E., & Straub, D. W. (2011). An update and extension to SEM guidelines for administrative and social science research. *MIS Quarterly*, *35*(2), iii–xiv.

Gerber, A. S., Green, D. P., & Nickerson, D. (2001). Testing for publication bias in political science. *Political Analysis*, *9*(4), 385–392.

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research*, *37*(1), 3–30.

Gill, J. (2018). Comments from the new editor. *Political Analysis*, *26*, 1–2.

Goodhue, D., Lewis, W., & Thompson, R. (2012). Does PLS have advantages for small sample size or non-normal data? *MIS Quarterly*, *36*(3), 981–1001.

Goodhue, D., Thompson, R., & Lewis, W. (2007). Statistical power in analyzing interaction effects: Questioning the advantage of PLS with product indicators. *Information Systems Research*, *18*(2), 211–227.

Greenwald, A. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Hair, J., Black, W., Babin, B., & Anderson, R. (2019). *Multivariate Data Analysis*. Cengage Learning.

Hair, J., Hult, G., Ringle, C., Sarstedt, M., Danks, N., & Ray, S. (2021). *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R*. Springer.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology*, *13*(3), e1002106.

Henseler, J., Hubona, G., & Ray, P. A. (2016). Using PLS path modeling in new technology research: Updated guidelines. *Industrial Management & Data Systems*, *116*(1), 2–20.

Hoenig, 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.

Hollenbeck, J. R., & Wright, P. M. (2017). Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. *Journal of Management*, *43*(1), 5–18.

Huang, W. (2013). PLSe: Efficient estimators and tests for partial least squares [Doctoral dissertation, University of California]. http://escholarship.org/uc/item/2cs2g2b0

Hubbard, R., & Armstrong, J. (1997). Publication bias against null results. *Psychological Reports*, *80*, 337–338.

Hutton, J., & Williamson, P. (2000). Bias in meta–analysis due to outcome variable selection within studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *49*(3), 359–370.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), 696–701.

Jenkin, T., Chan, Y., & Sabherwal, R. (2019). Mutual understanding in information systems development: Changes within and across projects. *MIS Quarterly*, *43*(2), 649–671.

Joseph, D., Ng, K.-Y., Koh, C., & Ang, S. (2007). Turnover of information technology professionals: A narrative review, meta-analytic structural equation modeling, and model development. *MIS Quarterly*, *31*(3), 547–577.

Kahai, S., & Cooper, R. (2003). Exploring the core concepts of media richness theory: The impact of cue multiplicity and feedback immediacy on decision quality. *Journal of Management Information Systems*, *20*(1), 263–299.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152.

Kepes, S., Banks, G., McDaniel, M., & Whetzel, D. (2012). Publication bias in the organizational sciences. *Organizational Research Methods*, *15*(4), 624-662..

Kepes, S., & McDaniel, M. (2013). How trustworthy is the scientific literature in industrial and organizational psychology? *Industrial and Organizational Psychology*, *6*, 252–268.

Kerr, N. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217.

Kruschke, J. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.

Kühberger, A., Fritz, A., & Schemdl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS One*, *9*(9).

Lakens, D., McLatchie, N., Isager, P., Scheel, A., & Dienes, Z. (2020). Improving inferences about null effects with bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, *75*(1), 45–57.

Lakens, D., Scheel, A., & Isager, P. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 25–269.

Leamer, E. (1983). Let's take the con out of econometrics. *American Economic Review*, *73*(1), 31–43.

Lin, M., Lucas, H., & Shmueli, G. (2013). Too big to fail: Large samples and the p-value problem. *Information Systems Research*, *24*(4), 906–917.

Majchrzak, A., Beath, C., Lim, R., & Chin, W. (2005). Managing client dialogues during information systems design to facilitate client learning. *MIS Quarterly*, *29*(4), 653–672.

Malhotra, A., Gosain, S., & El Sawy, O. (2007). Leveraging standard electronic business interfaces to enable adaptive supply chain partnerships. *Information Systems Research*, *18*(3), 260–279.

Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*(2), 331–348.

McDaniel, M., Rothstein, H., & Whetzel, D. (2006). Publication bias: A case study of four test vendors. *Personnel Psychology*, *59*, 927–953.

Mehler, D., Edelsbrunner, P., & Matić, K. (2019). Appreciating the significance of non-significant findings in

psychology. *Journal of European Psychology Students*, *10*(4), 1–7.

Muthén, L., & Muthén, B. (2002). How to use a monte carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599–620.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301.

Nosek, B., Spies, J., & Motyl, M. (2012). Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, *43*(2), 376–399.

Peng, D. X., & Lai, F. (2012). Using partial least squares in operations management research: A practical guideline and summary of past research. *Journal of Operations Management*, *30*(6), 467–480.

Pornprasertmanit, S., Miller, P., & Schoemann, A. (2020). simsem: SIMulated Structural Equation Modeling (0.5-15). http://CRAN.R-project.org/package=simsem

Qureshi, I., & Compeau, D. (2009). Assessing between-group differences in information systems research: A comparison of covariance- and component-based SEM. *MIS Quarterly*, *33*(1), 197–214.

R Core Team. (2021). R: A Language and Environment for Statistical Computing (4.1.0). R Foundation for Statistical Computing.

Reinartz, W. J., Haenlein, M., & Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of Research in Marketing*, *26*(4), 332–344.

Rigdon, E. E. (2016). Choosing PLS path modeling as analytical method in European management research: A realist perspective. *European Management Journal*, *34*(6), 598–605.

Ringle, C., Sarstedt, M., & Straub, D. (2012). A critical look at the use of PLS-SEM in MIS Quarterly. *MIS Quarterly*, *36*(1), iii–xiv.

Rönkkö, M. (2021). matrixpls: Matrix-based partial least squares estimation (1.0.12). https://github.com/mronkko/matrixpls

Rönkkö, M., & Evermann, J. (2013). A critical examination of common beliefs about partial least squares path modeling. *Organizational Research Methods*, *16*(3), 425–448.

Rönkkö, M., McIntosh, C. N., & Aguirre-Urreta, M. I. (2016). Improvements to PLSc: Remaining problems and simple solutions. Unpublished Working Paper. http://urn.fi/URN:NBN:fi:aalto-201603051463

Rönkkö, M., McIntosh, C. N., & Antonakis, J. (2015). On the adoption of partial least squares in psychological research: Caveat emptor. *Personality and Individual Differences*, *87*, 76–84.

Rönkkö, M., McIntosh, C. N., Antonakis, J., & Edwards, J. R. (2016). Partial least squares path modeling: Time for some serious second thoughts. *Journal of Operations Management*, 47–48, 9–27.

Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Rothstein, H., Sutton, A., & Borenstein, M. (2005). Publication bias in meta-analyses. In H. Rothstein, A. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-analysis: Prevention, Assessment, and Adjustments* (pp. 1–7). Wiley.

Selvin, H., & Stuart, A. (1966). Data-dredging procedures in survey analysis. *American Statistician*, *20*(3), 20–23.

Serrano, C., & Karahanna, E. (2016). The compensatory interaction between user capabilities and technology capabilities in influencing task performance: An empirical assessment in telemedicine consultations. *MIS Quarterly*, *40*(3), 597–622.

Sharma, R., & Yetton, P. (2007). The contingent effects of training, technical complexity, and task interdependence on successful information systems implementation. *MIS Quarterly*, *31*(2), 219–238.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547.

Song, J. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment*, *14*, 1–22.

Starbuck, W. (2016). 60th anniversary essay how journals could improve research practices in social science. *Administrative Science Quarterly*, *61*(2), 165–183.

Sterling, T., Rosenbaum, W., & Weinkam, J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*(1), 108–112.

Straub, D. (1989). Validating instruments in MIS research. *MIS Quarterly*, *13*(2), 147–169.

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its cause and consequences. *Journal of Clinical Epidemiology*, *53*(2), 207–216.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1–2.

Turner, E., Matthews, A., Linardatos, E., & Tell, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, *358*(3), 252–260.

Valentine, J., Aloe, A., & Lau, T. (2015). Life after NHST: how to describe your data without "p-ing" everywhere. *Basic and Applied Social Psychology*, *37*, 260–273.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05." *The American Statistician*, *73*(sup1), 1–19.

Williamson, P., Gamble, C., Altmand, D., & Hutton, J. (2005). Outcome selection bias in meta-analysis. *Statistical Methods in Medical Research*, *14*(5), 515–524.

Wold, H. (1982). Soft Modeling—The basic design and some extensions. In K. G. Jöreskog & S. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (pp. 1–54). North-Holland.

Wu, J., & Lederer, A. (2009). A meta-analysis of the role of environment-based voluntariness in information technology acceptance. *MIS Quarterly*, *33*(2), 419-432.

Yong, E. (2012). Replication studies: Bad copy. *Nature*, *485*(7398), 298–300.

About the Authors

Miguel I. Aguirre-Urreta is an associate professor in the Department of Information Systems and Business Analytics, College of Business, at Florida International University, and the Director of Doctoral Programs for the College. Before joining FIU, he was on the faculty at Texas Tech University and DePaul University. Miguel received his Ph.D. in Information Systems from the University of Kansas, and his MBA in Information Systems and Finance from the Kelley School of Business at Indiana University. His undergraduate degree is in public accounting from the Universidad de Buenos Aires, in Argentina. Miguel is interested in quantitative research methods, computer self-efficacy, human–computer interaction, technology acceptance and diffusion, and formal modeling and theory development. He is also a past chair of the AIS Special Interest Group on Human-Computer Interaction. His research has appeared or is forthcoming at *MIS Quarterly*, *Information Systems Research*, *Psychological Methods*, *Communications of the Association for Information Systems*, *Research Synthesis Methods*, *IEEE Software*, *Measurement*, *Organizational Research Methods*, and *The DATA BASE for Advances in Information Systems*, among others.

Mikko Rönkkö is associate professor of entrepreneurship at Jyväskylä University School of Business and Economics (JSBE) and a docent at Aalto University School of Science in the field of statistical methods in management research. His current research interests are growth entrepreneurship and quantitative research methods in management research. His work on methods has been published in e.g., *Organizational Research Methods*, *Psychological Methods*, *Structural Equation Modeling*, and *MIS Quarterly*. He serves on the editorial boards of *Organizational Research Methods* and *Entrepreneurship Theory and Practice*. In the past Mikko has also been an entrepreneur.

Cameron N. McIntosh is the manager of the Data Development Unit within the Canadian Centre for Justice and Community Safety Statistics, which is part of Statistics Canada. His current research interests are focused on crime and quantitative methods in the social sciences. His work on statistical methods has appeared in, e.g., *Psychological Methods* and *Organizational Research Methods*. He serves on the editorial boards of the *Journal of Sport and Exercise Psychology* and *Quality of Life Research*.

Appendix A. Full Results with All Effects Significant

The following tables present full results for all conditions and statistical approaches where all four paths were found to be significant. Empty rows indicate not a single replication (out of the 1,000 generated for each condition) in that condition had all four effects flagged as significant. Rows with NA are those where analyses would either not converge at all (for the case of PLSc) or were not run due to model identification issues (for the case of SEM).

Table A1. OLS Results

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
20	3	Mixed	-1155%	69%	141%	-8%	2
20	3	.7	-1272%	65%	159%	18%	3
20	3	.8	-1310%	63%	155%	23%	4
20	3	.9	-1072%	83%	134%	-10%	1
20	6	Mixed	-1234%	20%	220%	19%	1
20	6	.7					
20	6	.8					
20	6	.9	-1036%	5%	212%	24%	1
20	9	Mixed					
20	9	.7	-332%	47%	145%	9%	2
20	9	.8					
20	9	.9	52%	47%	133%	-1%	3
40	3	Mixed	-135%	21%	82%	-18%	4
40	3	.7	-840%	7%	180%	-26%	3
40	3	.8	-490%	10%	93%	-17%	7
40	3	.9	-179%	35%	26%	-10%	6
40	6	Mixed	77%	0%	31%	-18%	5
40	6	.7	-878%	5%	127%	-15%	5
40	6	.8	-669%	4%	123%	-4%	7
40	6	.9	-583%	4%	120%	-3%	10
40	9	Mixed	-729%	3%	126%	-2%	9
40	9	.7	-438%	11%	140%	-4%	7
40	9	.8	-875%	6%	128%	-1%	8
40	9	.9	-242%	12%	131%	3%	7
60	3	Mixed	-156%	-10%	76%	-10%	7
60	3	.7	-228%	2%	52%	-19%	8
60	3	.8	-243%	6%	57%	-17%	13
60	3	.9	-37%	-14%	75%	0%	11
60	6	Mixed	-297%	-3%	72%	-7%	12
60	6	.7	-153%	-6%	80%	-13%	7
60	6	.8	-326%	-8%	82%	-5%	10
60	6	.9	-259%	-6%	84%	1%	11
60	9	Mixed	-255%	-7%	83%	-1%	11
60	9	.7	-134%	-11%	84%	-8%	9
60	9	.8	-328%	-5%	85%	-2%	10
60	9	.9	-186%	-9%	84%	1%	12
80	3	Mixed	319%	-21%	44%	-18%	12
80	3	.7	-131%	-14%	38%	-24%	16
80	3	.8	-25%	-8%	44%	-15%	20
80	3	.9	263%	-14%	42%	-11%	14
80	6	Mixed	141%	2%	19%	-9%	19
80	6	.7	89%	-10%	39%	-10%	13
80	6	.8	128%	-5%	40%	-5%	15
80	6	.9	174%	1%	39%	-3%	18
80	9	Mixed	175%	-4%	37%	-4%	18
80	9	.7	281%	-4%	51%	-13%	9

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
80	9	.8	129%	-7%	43%	-2%	15
80	9	.9	162%	-2%	62%	-4%	12
100	3	Mixed	290%	-8%	29%	-19%	19
100	3	.7	213%	-18%	33%	-22%	16
100	3	.8	284%	-17%	34%	-11%	21
100	3	.9	230%	0%	32%	-11%	25
100	6	Mixed	115%	-3%	39%	-10%	25
100	6	.7	82%	-11%	43%	-14%	22
100	6	.8	130%	-9%	44%	-10%	23
100	6	.9	130%	-5%	45%	-6%	27
100	9	Mixed	103%	-5%	45%	-8%	24
100	9	.7	168%	-4%	36%	-9%	24
100	9	.8	116%	-6%	43%	-8%	25
100	9	.9	137%	3%	41%	-3%	29

Table A2. DR Results

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
20	3	Mixed	-869%	44%	158%	25%	12
20	3	.7	-871%	62%	165%	28%	24
20	3	.8	-1269%	77%	250%	23%	23
20	3	.9	-1215%	44%	149%	35%	5
20	6	Mixed	-895%	41%	180%	14%	8
20	6	.7	-821%	34%	181%	11%	23
20	6	.8	-1108%	64%	184%	14%	8
20	6	.9	-1171%	55%	194%	-2%	3
20	9	Mixed	-1112%	60%	168%	4%	4
20	9	.7	-351%	43%	85%	17%	14
20	9	.8	-1080%	57%	179%	6%	8
20	9	.9	42%	51%	139%	1%	3
40	3	Mixed	-336%	24%	107%	5%	32
40	3	.7	-404%	30%	119%	2%	68
40	3	.8	-337%	19%	121%	3%	37
40	3	.9	-62%	19%	82%	2%	12
40	6	Mixed	-321%	8%	74%	1%	17
40	6	.7	-381%	24%	120%	-1%	29
40	6	.8	-479%	10%	126%	1%	21
40	6	.9	-513%	-1%	126%	3%	19
40	9	Mixed	-544%	3%	136%	3%	20
40	9	.7	-406%	22%	139%	4%	24
40	9	.8	-546%	3%	137%	3%	20
40	9	.9	-315%	16%	127%	0%	13
60	3	Mixed	-154%	1%	87%	3%	38
60	3	.7	-379%	19%	97%	4%	67
60	3	.8	-327%	13%	95%	1%	44
60	3	.9	-101%	-1%	73%	0%	26
60	6	Mixed	-288%	6%	73%	0%	24
60	6	.7	-240%	7%	64%	5%	33
60	6	.8	-227%	5%	68%	2%	21
60	6	.9	-329%	-4%	88%	6%	13
60	9	Mixed	-286%	0%	88%	5%	17

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
60	9	.7	-295%	8%	71%	8%	30
60	9	.8	-286%	0%	87%	5%	17
60	9	.9	-281%	1%	88%	6%	15
80	3	Mixed	132%	-3%	60%	-3%	40
80	3	.7	-111%	7%	77%	1%	74
80	3	.8	-50%	3%	53%	-1%	46
80	3	.9	225%	-7%	56%	-3%	29
80	6	Mixed	-29%	8%	46%	0%	37
80	6	.7	6%	6%	66%	-5%	46
80	6	.8	86%	2%	50%	-3%	34
80	6	.9	151%	-1%	43%	-1%	23
80	9	Mixed	119%	-1%	48%	-2%	28
80	9	.7	93%	2%	65%	-3%	26
80	9	.8	130%	1%	49%	-2%	25
80	9	.9	109%	0%	69%	0%	13
100	3	Mixed	56%	10%	51%	-2%	55
100	3	.7	-110%	7%	66%	8%	83
100	3	.8	23%	-2%	54%	6%	55
100	3	.9	210%	3%	40%	-4%	36
100	6	Mixed	57%	1%	38%	-1%	47
100	6	.7	80%	1%	49%	-1%	52
100	6	.8	85%	-2%	51%	-1%	43
100	6	.9	142%	-3%	50%	-1%	34
100	9	Mixed	95%	0%	49%	-2%	38
100	9	.7	22%	9%	48%	0%	46
100	9	.8	100%	-1%	51%	-2%	39
100	9	.9	144%	4%	44%	0%	30

Table A3. PLS Results

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
20	3	Mixed	601%	38%	177%	-32%	1
20	3	.7					
20	3	.8	928%	41%	78%	-33%	1
20	3	.9	-902%	32%	252%	11%	2
20	6	Mixed	505%	28%	17%	-7%	7
20	6	.7	760%	41%	-151%	-25%	2
20	6	.8	662%	13%	74%	-12%	7
20	6	.9					
20	9	Mixed	598%	57%	151%	-30%	3
20	9	.7	583%	37%	167%	-7%	3
20	9	.8	544%	23%	146%	5%	3
20	9	.9					
40	3	Mixed	501%	0%	69%	-11%	6
40	3	.7	636%	7%	51%	-15%	2
40	3	.8	365%	7%	62%	-17%	8
40	3	.9	164%	11%	57%	-11%	13
40	6	Mixed	496%	9%	89%	-18%	15
40	6	.7	506%	-4%	66%	-5%	11
40	6	.8	260%	5%	98%	-11%	14
40	6	.9					

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
40	9	Mixed	559%	9%	69%	-13%	18
40	9	.7	535%	-7%	72%	-8%	18
40	9	.8	433%	10%	81%	-16%	15
40	9	.9					
60	3	Mixed	447%	-11%	35%	-17%	17
60	3	.7	366%	3%	55%	-31%	17
60	3	.8	379%	-3%	60%	-20%	18
60	3	.9	397%	-10%	33%	-14%	18
60	6	Mixed	375%	-3%	31%	-8%	31
60	6	.7	388%	-7%	63%	-16%	27
60	6	.8	439%	-11%	37%	-10%	21
60	6	.9					
60	9	Mixed	392%	-11%	50%	-6%	21
60	9	.7	441%	-4%	56%	-12%	31
60	9	.8	405%	5%	22%	-10%	28
60	9	.9					
80	3	Mixed	417%	-8%	23%	-18%	26
80	3	.7	322%	-3%	25%	-28%	16
80	3	.8	314%	-12%	33%	-17%	40
80	3	.9	168%	-6%	48%	-11%	18
80	6	Mixed	290%	-4%	39%	-8%	36
80	6	.7	339%	-12%	33%	-14%	43
80	6	.8	375%	-14%	26%	-5%	42
80	6	.9					
80	9	Mixed	242%	-9%	35%	-4%	40
80	9	.7	369%	-13%	34%	-10%	46
80	9	.8	320%	-8%	32%	-5%	48
80	9	.9					
100	3	Mixed	293%	-12%	21%	-15%	40
100	3	.7	287%	-21%	20%	-25%	29
100	3	.8	147%	-12%	22%	-15%	21
100	3	.9	170%	-7%	24%	-5%	40
100	6	Mixed	311%	-8%	17%	-9%	65
100	6	.7	346%	-11%	19%	-16%	45
100	6	.8	328%	-10%	12%	-10%	42
100	6	.9	65%	7%	24%	-7%	9
100	9	Mixed	239%	-5%	21%	-6%	53
100	9	.7	278%	-13%	22%	-11%	55
100	9	.8	318%	-9%	29%	-7%	43
100	9	.9					

Table A4. SEM Results

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
20	3	Mixed	-713%	52%	192%	21%	20
20	3	.7	-761%	82%	139%	15%	19
20	3	.8	-874%	65%	166%	11%	15
20	3	.9	-658%	38%	178%	18%	17
20	6	Mixed	NA	NA	NA	NA	NA
20	6	.7	NA	NA	NA	NA	NA
20	6	.8	NA	NA	NA	NA	NA

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
20	6	.9	NA	NA	NA	NA	NA
20	9	Mixed	NA	NA	NA	NA	NA
20	9	.7	NA	NA	NA	NA	NA
20	9	.8	NA	NA	NA	NA	NA
20	9	.9	NA	NA	NA	NA	NA
40	3	Mixed	-504%	31%	97%	8%	22
40	3	.7	-607%	55%	166%	0%	18
40	3	.8	-570%	34%	148%	2%	19
40	3	.9	-306%	24%	52%	2%	15
40	6	Mixed	-224%	18%	127%	2%	15
40	6	.7	-699%	26%	161%	6%	19
40	6	.8	-651%	17%	124%	5%	21
40	6	.9	-404%	5%	106%	0%	23
40	9	Mixed	-355%	7%	126%	0%	22
40	9	.7	-361%	19%	130%	5%	22
40	9	.8	-480%	16%	110%	4%	21
40	9	.9	-252%	11%	112%	3%	19
60	3	Mixed	-109%	5%	75%	1%	26
60	3	.7	-672%	38%	129%	10%	24
60	3	.8	-537%	26%	109%	5%	20
60	3	.9	-173%	8%	73%	-2%	20
60	6	Mixed	-314%	4%	79%	0%	16
60	6	.7	-590%	4%	106%	10%	13
60	6	.8	-447%	4%	92%	9%	14
60	6	.9	-260%	0%	86%	4%	16
60	9	Mixed	-202%	2%	60%	5%	16
60	9	.7	-490%	18%	83%	9%	16
60	9	.8	-333%	6%	90%	6%	16
60	9	.9	-196%	-1%	89%	4%	17
80	3	Mixed	32%	2%	62%	1%	21
80	3	.7	-371%	17%	98%	2%	24
80	3	.8	-231%	9%	76%	0%	24
80	3	.9	229%	-8%	59%	-3%	22
80	6	Mixed	-95%	5%	46%	0%	25
80	6	.7	-105%	7%	69%	-1%	16
80	6	.8	57%	0%	54%	-1%	23
80	6	.9	106%	3%	45%	-2%	27
80	9	Mixed	62%	6%	47%	-2%	26
80	9	.7	-5%	1%	79%	1%	17
80	9	.8	91%	4%	45%	-2%	21
80	9	.9	108%	0%	60%	-2%	17
100	3	Mixed	91%	8%	50%	-1%	27
100	3	.7	-178%	15%	89%	7%	19
100	3	.8	90%	5%	66%	2%	22
100	3	.9	218%	7%	42%	-5%	26
100	6	Mixed	95%	0%	44%	0%	42
100	6	.7	-124%	2%	75%	1%	25
100	6	.8	47%	-2%	56%	-1%	28
100	6	.9	107%	-2%	44%	-1%	36
100	9	Mixed	159%	-6%	48%	-1%	30
100	9	.7	-45%	10%	55%	3%	32
100	9	.8	64%	-4%	56%	-1%	33
100	9	.9	139%	5%	40%	-2%	34

Table A5. PLSc Results

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
20	3	Mixed					
20	3	.7					
20	3	.8					
20	3	.9	-147%	13%	204%	22%	2
20	6	Mixed	3%	52%	221%	-1%	3
20	6	.7					
20	6	.8					
20	6	.9	NA	NA	NA	NA	NA
20	9	Mixed	439%	46%	31%	3%	4
20	9	.7					
20	9	.8					
20	9	.9	NA	NA	NA	NA	NA
40	3	Mixed	465%	10%	107%	7%	5
40	3	.7	814%	25%	77%	-15%	1
40	3	.8	238%	55%	107%	-7%	8
40	3	.9	276%	13%	121%	3%	6
40	6	Mixed	702%	-2%	97%	3%	4
40	6	.7	504%	16%	132%	-2%	2
40	6	.8	395%	13%	117%	-3%	14
40	6	.9	NA	NA	NA	NA	NA
40	9	Mixed	398%	5%	99%	-2%	11
40	9	.7	611%	-5%	100%	-4%	6
40	9	.8	276%	15%	62%	-2%	13
40	9	.9	NA	NA	NA	NA	NA
60	3	Mixed	191%	18%	81%	-5%	4
60	3	.7	449%	19%	91%	-23%	1
60	3	.8	508%	19%	60%	-10%	7
60	3	.9	115%	6%	90%	-3%	20
60	6	Mixed	333%	7%	78%	-5%	20
60	6	.7	525%	13%	28%	-6%	7
60	6	.8	448%	5%	59%	-5%	16
60	6	.9	NA	NA	NA	NA	NA
60	9	Mixed	296%	4%	62%	-12%	16
60	9	.7	512%	-4%	81%	-7%	12
60	9	.8	432%	-3%	56%	-7%	30
60	9	.9	NA	NA	NA	NA	NA
80	3	Mixed	60%	8%	82%	-4%	17
80	3	.7	-18%	35%	107%	-8%	7
80	3	.8	101%	-2%	82%	3%	12
80	3	.9	360%	-12%	43%	4%	23
80	6	Mixed	408%	-5%	44%	-1%	27
80	6	.7	404%	5%	71%	-4%	21
80	6	.8	301%	-4%	60%	-1%	34
80	6	.9	NA	NA	NA	NA	NA
80	9	Mixed	324%	-2%	49%	-4%	37
80	9	.7	376%	-2%	63%	-5%	30
80	9	.8	428%	-8%	29%	-3%	41
80	9	.9	NA	NA	NA	NA	NA
100	3	Mixed	274%	-1%	45%	-4%	21
100	3	.7	488%	-2%	43%	-7%	7

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
100	3	.8	338%	-1%	44%	-4%	17
100	3	.9	206%	-2%	34%	-2%	31
100	6	Mixed	254%	-2%	33%	-1%	38
100	6	.7	396%	-1%	42%	-5%	23
100	6	.8	393%	-7%	23%	-2%	31
100	6	.9	NA	NA	NA	NA	NA
100	9	Mixed	354%	-6%	26%	0%	43
100	9	.7	305%	0%	48%	-2%	45
100	9	.8	330%	-3%	29%	-2%	42
100	9	.9	NA	NA	NA	NA	NA

Appendix B. Full Results with 3 or More Effects Significant

The following tables present full results for all conditions and statistical approaches where three of more paths were found to be significant. Empty rows indicate not a single replication (out of the 1,000 generated for each condition) in that condition had all four effects flagged as significant. Rows with NA are those where analyses would either not converge at all (for the case of PLSc) or were not run due to model identification issues (for the case of SEM).

Table B.1. OLS Results

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
20	3	Mixed	-129%	19%	125%	-5%	26
20	3	.7	-410%	31%	117%	-8%	18
20	3	.8	-211%	31%	112%	-11%	35
20	3	.9	-75%	15%	116%	-4%	43
20	6	Mixed	-251%	27%	132%	1%	43
20	6	.7	-251%	30%	98%	3%	26
20	6	.8	-359%	26%	119%	8%	42
20	6	.9	-333%	24%	133%	8%	47
20	9	Mixed	-338%	23%	127%	8%	50
20	9	.7	-164%	36%	104%	-6%	46
20	9	.8	-380%	25%	125%	8%	48
20	9	.9	-72%	26%	95%	-3%	70
40	3	Mixed	-38%	6%	70%	-13%	121
40	3	.7	-56%	4%	67%	-23%	93
40	3	.8	-16%	5%	67%	-14%	146
40	3	.9	-44%	3%	75%	-6%	184
40	6	Mixed	6%	8%	70%	-13%	158
40	6	.7	-27%	6%	64%	-17%	156
40	6	.8	-17%	8%	66%	-12%	189
40	6	.9	-19%	8%	70%	-7%	221
40	9	Mixed	-13%	9%	69%	-9%	199
40	9	.7	-7%	2%	69%	-8%	161
40	9	.8	-18%	9%	67%	-9%	207
40	9	.9	-25%	3%	76%	-1%	215
60	3	Mixed	-17%	-10%	50%	-15%	214
60	3	.7	-2%	-8%	35%	-25%	152
60	3	.8	-26%	-6%	46%	-15%	225
60	3	.9	-23%	-7%	55%	-6%	264
60	6	Mixed	-3%	0%	47%	-10%	262
60	6	.7	-16%	-4%	48%	-16%	226
60	6	.8	-21%	-2%	53%	-10%	272
60	6	.9	-31%	-1%	57%	-4%	307
60	9	Mixed	-24%	-2%	58%	-6%	281
60	9	.7	-14%	-5%	49%	-10%	261
60	9	.8	-22%	-1%	57%	-7%	283
60	9	.9	-18%	-1%	54%	-3%	315
80	3	Mixed	-6%	-10%	34%	-15%	356
80	3	.7	2%	-15%	28%	-24%	256
80	3	.8	-1%	-11%	33%	-15%	337
80	3	.9	-4%	-5%	39%	-8%	417
80	6	Mixed	7%	-5%	36%	-9%	397
80	6	.7	-3%	-10%	35%	-15%	326
80	6	.8	-5%	-6%	38%	-9%	377
80	6	.9	-3%	-3%	40%	-5%	409

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
80	9	Mixed	-1%	-4%	39%	-7%	397
80	9	.7	12%	-6%	37%	-10%	347
80	9	.8	-3%	-4%	39%	-7%	396
80	9	.9	11%	0%	40%	-3%	401
100	3	Mixed	-13%	-11%	25%	-15%	437
100	3	.7	-1%	-18%	18%	-24%	348
100	3	.8	-9%	-11%	25%	-14%	423
100	3	.9	-17%	-5%	29%	-6%	507
100	6	Mixed	2%	-6%	26%	-9%	478
100	6	.7	-6%	-13%	25%	-14%	425
100	6	.8	-9%	-8%	29%	-8%	476
100	6	.9	-11%	-4%	31%	-4%	514
100	9	Mixed	-11%	-6%	30%	-6%	498
100	9	.7	-4%	-8%	27%	-10%	442
100	9	.8	-10%	-6%	29%	-6%	499
100	9	.9	-5%	-2%	31%	-2%	504

Table B2. DR Results

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
20	3	Mixed	-173%	19%	126%	9%	120
20	3	.7	-55%	23%	57%	7%	137
20	3	.8	-265%	32%	106%	9%	116
20	3	.9	-143%	24%	122%	3%	88
20	6	Mixed	-181%	25%	123%	5%	104
20	6	.7	-228%	27%	99%	8%	152
20	6	.8	-255%	27%	120%	6%	113
20	6	.9	-313%	27%	133%	7%	86
20	9	Mixed	-297%	28%	128%	7%	95
20	9	.7	-149%	26%	99%	4%	122
20	9	.8	-311%	27%	128%	7%	93
20	9	.9	-83%	29%	98%	0%	83
40	3	Mixed	-52%	11%	78%	3%	286
40	3	.7	-114%	17%	83%	1%	312
40	3	.8	-75%	14%	78%	0%	288
40	3	.9	-33%	9%	78%	0%	248
40	6	Mixed	-22%	11%	73%	-4%	246
40	6	.7	-47%	17%	72%	-2%	293
40	6	.8	-43%	14%	71%	-3%	269
40	6	.9	-25%	10%	70%	-3%	257
40	9	Mixed	-33%	12%	69%	-3%	264
40	9	.7	-41%	7%	72%	3%	268
40	9	.8	-36%	12%	74%	-3%	262
40	9	.9	-38%	5%	78%	2%	234
60	3	Mixed	-30%	-1%	56%	1%	369
60	3	.7	-71%	9%	53%	0%	397
60	3	.8	-40%	5%	53%	-1%	366
60	3	.9	-19%	-2%	52%	0%	352
60	6	Mixed	-17%	4%	54%	-2%	333
60	6	.7	-35%	5%	55%	-1%	364

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
60	6	.8	-31%	4%	57%	-2%	355
60	6	.9	-31%	2%	57%	-1%	342
60	9	Mixed	-37%	2%	57%	-1%	346
60	9	.7	-15%	2%	52%	0%	364
60	9	.8	-36%	2%	58%	-1%	348
60	9	.9	-27%	0%	57%	0%	338
80	3	Mixed	-16%	3%	47%	-1%	499
80	3	.7	-12%	1%	44%	0%	551
80	3	.8	-13%	0%	43%	1%	519
80	3	.9	-20%	1%	45%	-1%	484
80	6	Mixed	1%	2%	43%	-1%	478
80	6	.7	6%	2%	39%	-1%	494
80	6	.8	2%	1%	40%	-1%	473
80	6	.9	-1%	0%	40%	-1%	453
80	9	Mixed	1%	1%	40%	-1%	457
80	9	.7	8%	2%	42%	-1%	453
80	9	.8	2%	0%	40%	-1%	465
80	9	.9	6%	2%	41%	-1%	430
100	3	Mixed	-22%	2%	34%	1%	608
100	3	.7	-18%	4%	38%	1%	588
100	3	.8	-18%	2%	36%	1%	574
100	3	.9	-21%	1%	33%	1%	587
100	6	Mixed	3%	0%	32%	-1%	549
100	6	.7	-2%	0%	33%	0%	569
100	6	.8	-3%	-1%	33%	0%	564
100	6	.9	-7%	-1%	33%	0%	544
100	9	Mixed	-6%	0%	33%	0%	553
100	9	.7	-10%	0%	35%	0%	537
100	9	.8	-8%	0%	33%	0%	552
100	9	.9	-8%	-1%	33%	0%	528

Table B3. PLS Results

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
20	3	Mixed	144%	29%	100%	-9%	26
20	3	.7	265%	44%	84%	-19%	19
20	3	.8	215%	31%	124%	-12%	32
20	3	.9	-71%	30%	108%	-6%	62
20	6	Mixed	215%	31%	74%	-9%	66
20	6	.7	388%	26%	123%	-13%	36
20	6	.8	365%	17%	90%	-2%	63
20	6	.9					
20	9	Mixed	276%	31%	124%	-9%	55
20	9	.7	389%	22%	109%	-6%	55
20	9	.8	204%	26%	130%	-6%	74
20	9	.9					
40	3	Mixed	132%	9%	60%	-16%	180
40	3	.7	179%	9%	51%	-20%	76
40	3	.8	140%	8%	58%	-15%	155

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
40	3	.9	72%	7%	60%	-8%	202
40	6	Mixed	136%	9%	74%	-9%	221
40	6	.7	274%	9%	47%	-15%	172
40	6	.8	152%	5%	63%	-8%	225
40	6	.9					
40	9	Mixed	157%	10%	65%	-8%	237
40	9	.7	198%	7%	70%	-11%	206
40	9	.8	142%	9%	68%	-10%	243
40	9	.9					
60	3	Mixed	61%	-3%	49%	-15%	275
60	3	.7	115%	-2%	48%	-24%	167
60	3	.8	48%	-3%	45%	-14%	285
60	3	.9	49%	-2%	44%	-9%	336
60	6	Mixed	91%	-1%	46%	-8%	333
60	6	.7	130%	-1%	52%	-14%	303
60	6	.8	86%	-2%	50%	-8%	349
60	6	.9					
60	9	Mixed	51%	2%	55%	-6%	378
60	9	.7	115%	-2%	54%	-9%	357
60	9	.8	83%	2%	45%	-8%	389
60	9	.9					
80	3	Mixed	57%	-6%	33%	-14%	373
80	3	.7	93%	-7%	25%	-24%	263
80	3	.8	77%	-8%	30%	-16%	414
80	3	.9	23%	-3%	35%	-8%	437
80	6	Mixed	34%	-3%	41%	-7%	471
80	6	.7	118%	-5%	30%	-15%	423
80	6	.8	76%	-4%	38%	-8%	427
80	6	.9					
80	9	Mixed	47%	-2%	39%	-6%	482
80	9	.7	102%	-5%	38%	-9%	431
80	9	.8	55%	0%	40%	-6%	518
80	9	.9					
100	3	Mixed	32%	-8%	23%	-14%	476
100	3	.7	64%	-16%	23%	-24%	369
100	3	.8	27%	-10%	23%	-15%	438
100	3	.9	-4%	-7%	29%	-6%	547
100	6	Mixed	47%	-4%	29%	-8%	571
100	6	.7	76%	-8%	27%	-14%	466
100	6	.8	31%	-5%	25%	-8%	555
100	6	.9	13%	0%	30%	-4%	101
100	9	Mixed	18%	-4%	32%	-5%	586
100	9	.7	48%	-6%	36%	-9%	527
100	9	.8	29%	-2%	32%	-6%	568
100	9	.9					

Table B4. SEM Results

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
20	3	Mixed	-291%	37%	148%	8%	156
20	3	.7	-240%	46%	129%	6%	114
20	3	.8	-189%	33%	128%	7%	139
20	3	.9	-129%	27%	104%	3%	155
20	6	Mixed	NA	NA	NA	NA	NA
20	6	.7	NA	NA	NA	NA	NA
20	6	.8	NA	NA	NA	NA	NA
20	6	.9	NA	NA	NA	NA	NA
20	9	Mixed	NA	NA	NA	NA	NA
20	9	.7	NA	NA	NA	NA	NA
20	9	.8	NA	NA	NA	NA	NA
20	9	.9	NA	NA	NA	NA	NA
40	3	Mixed	-70%	18%	84%	1%	250
40	3	.7	-150%	33%	102%	0%	181
40	3	.8	-93%	21%	90%	1%	222
40	3	.9	-40%	10%	75%	1%	263
40	6	Mixed	10%	7%	65%	1%	271
40	6	.7	-89%	18%	80%	-2%	246
40	6	.8	-50%	16%	72%	-2%	260
40	6	.9	-11%	11%	64%	-3%	310
40	9	Mixed	-23%	12%	62%	-3%	303
40	9	.7	-82%	13%	74%	2%	245
40	9	.8	-27%	13%	65%	-4%	294
40	9	.9	-4%	4%	64%	1%	291
60	3	Mixed	-65%	4%	59%	1%	317
60	3	.7	-111%	18%	73%	2%	195
60	3	.8	-89%	10%	63%	0%	280
60	3	.9	-18%	0%	54%	0%	343
60	6	Mixed	-10%	6%	55%	-2%	314
60	6	.7	-55%	9%	63%	-2%	273
60	6	.8	-36%	5%	58%	-2%	326
60	6	.9	-28%	2%	53%	-1%	366
60	9	Mixed	-21%	2%	54%	-1%	353
60	9	.7	-41%	4%	57%	0%	324
60	9	.8	-34%	3%	57%	-1%	336
60	9	.9	-18%	0%	50%	0%	380
80	3	Mixed	-23%	1%	48%	0%	414
80	3	.7	-53%	11%	66%	-1%	306
80	3	.8	-32%	3%	56%	-1%	382
80	3	.9	-17%	1%	48%	-1%	447
80	6	Mixed	-8%	1%	43%	-1%	463
80	6	.7	-20%	4%	52%	-1%	366
80	6	.8	-9%	2%	45%	-2%	431
80	6	.9	-5%	1%	39%	-1%	469
80	9	Mixed	-3%	0%	39%	-1%	462
80	9	.7	-1%	3%	48%	-1%	400
80	9	.8	-8%	1%	43%	-1%	448

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
80	9	.9	6%	1%	39%	0%	462
100	3	Mixed	-18%	3%	39%	0%	499
100	3	.7	-55%	9%	56%	1%	354
100	3	.8	-28%	4%	45%	1%	451
100	3	.9	-26%	2%	36%	1%	547
100	6	Mixed	-2%	1%	33%	0%	526
100	6	.7	-25%	0%	43%	0%	459
100	6	.8	-12%	0%	36%	0%	515
100	6	.9	-5%	0%	32%	0%	554
100	9	Mixed	-9%	0%	33%	0%	548
100	9	.7	-13%	1%	39%	0%	482
100	9	.8	-8%	0%	34%	0%	530
100	9	.9	-6%	-1%	32%	0%	541

Table B5. PLSc Results

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
20	3	Mixed	-268%	78%	130%	25%	13
20	3	.7					
20	3	.8	-217%	42%	242%	7%	12
20	3	.9	-170%	39%	177%	3%	42
20	6	Mixed	163%	30%	166%	4%	16
20	6	.7	-389%	85%	260%	-6%	3
20	6	.8	84%	46%	196%	-5%	34
20	6	.9	NA	NA	NA	NA	NA
20	9	Mixed	-44%	55%	141%	-3%	33
20	9	.7	129%	49%	206%	0%	10
20	9	.8	12%	45%	183%	-4%	26
20	9	.9	NA	NA	NA	NA	NA
40	3	Mixed	44%	22%	112%	-4%	90
40	3	.7	-47%	45%	157%	7%	24
40	3	.8	76%	29%	108%	-3%	85
40	3	.9	31%	13%	84%	-3%	184
40	6	Mixed	47%	17%	107%	-4%	148
40	6	.7	50%	29%	143%	-4%	62
40	6	.8	120%	18%	102%	-3%	145
40	6	.9	NA	NA	NA	NA	NA
40	9	Mixed	91%	13%	100%	-4%	166
40	9	.7	39%	28%	132%	-4%	99
40	9	.8	60%	18%	100%	-3%	155
40	9	.9	NA	NA	NA	NA	NA
60	3	Mixed	14%	10%	85%	-1%	187
60	3	.7	24%	24%	118%	-2%	76
60	3	.8	18%	15%	88%	-3%	187
60	3	.9	40%	2%	64%	1%	292
60	6	Mixed	61%	5%	76%	-1%	285
60	6	.7	59%	18%	93%	-3%	169
60	6	.8	46%	8%	73%	-2%	252
60	6	.9	NA	NA	NA	NA	NA

Sample Size	Indics.	Loadings	% Bias A → C	% Bias A → D	% Bias B → C	% Bias C → D	Count
60	9	Mixed	29%	5%	79%	-2%	317
60	9	.7	44%	13%	97%	-2%	209
60	9	.8	48%	7%	77%	-3%	301
60	9	.9	NA	NA	NA	NA	NA
80	3	Mixed	1%	5%	65%	-1%	298
80	3	.7	45%	17%	85%	-2%	142
80	3	.8	36%	5%	62%	1%	295
80	3	.9	29%	1%	44%	1%	412
80	6	Mixed	25%	2%	56%	0%	402
80	6	.7	63%	10%	76%	-2%	250
80	6	.8	45%	3%	59%	-1%	376
80	6	.9	NA	NA	NA	NA	NA
80	9	Mixed	9%	3%	61%	-1%	458
80	9	.7	24%	5%	78%	0%	352
80	9	.8	42%	1%	54%	0%	417
80	9	.9	NA	NA	NA	NA	NA
100	3	Mixed	27%	2%	50%	-1%	380
100	3	.7	4%	13%	77%	-2%	234
100	3	.8	27%	2%	47%	-1%	403
100	3	.9	5%	0%	37%	0%	509
100	6	Mixed	26%	0%	45%	0%	501
100	6	.7	45%	5%	58%	0%	388
100	6	.8	19%	2%	43%	0%	494
100	6	.9	NA	NA	NA	NA	NA
100	9	Mixed	21%	2%	45%	0%	493
100	9	.7	33%	4%	59%	0%	459
100	9	.8	8%	1%	46%	0%	523
100	9	.9	NA	NA	NA	NA	NA

Appendix C. R Code for Correlation Significance Example

The code below produces the correlation significance example (see Table 1 and Figure 1).

```
library(MASS)
library(dplyr)

### Simulation Parameters
SAMPLE_SIZE <- c(10, 20, 30, 40, 50, 60, 70)
POP_CORRELATION <- c(0, .1, .2, .3, .4, .5)
REP <- 10000
ALPHA <- .05

### Accumulate results over replications
out <- matrix(NA, nrow = length(SAMPLE_SIZE) * length(POP_CORRELATION) * REP, ncol = 6)
colnames(out) <- c('replication', 'sample_size', 'pop_correlation', 'estimate', 'pvalue', 'is_sig')

### Loop over the various combinations of conditions
### Generate data and calculate correlations and p values
### Store in the output matrix
ro <- 0
for(ss in 1:length(SAMPLE_SIZE)){

  for(pc in 1:length(POP CORRELATION)){

    for(r in 1:REP){

      dataset <- mvrnorm(n = SAMPLE_SIZE[ss], mu = c(0,0), Sigma = matrix(c(1, POP_CORRELATION[pc],
POP_CORRELATION[pc], 1), nrow = 2))
      res <- cor.test(x = dataset[,1], y = dataset[,2], method = 'pearson')
      ro <- ro + 1
      out[ro,'replication'] <- r
      out[ro,'sample_size'] <- SAMPLE_SIZE[ss]
      out[ro,'pop_correlation'] <- POP_CORRELATION[pc]
      out[ro,'estimate'] <- res$estimate
      out[ro,'pvalue'] <- res$p.value
      out[ro,'is_sig'] <- res$p.value < ALPHA

    }

  }

}

### Summarized output by population correlation and sample size
### Results can be used to build Table 1 in the research
grouped <- group_by(as.data.frame(out), sample_size, pop_correlation)
filtered <- filter(grouped, is_sig == 1)
table_output_all <- summarise(grouped, .groups = 'keep', mean_all = mean(estimate),
                                                         percentage_sig = mean(is_sig))
table_output_sig <- summarise(filtered, .groups = 'keep', mean_only_sig = mean(estimate))

### Histograms shown in Figure 1 in the research
hist(out[which(out[,'is_sig'] == 1 & out[,'sample_size'] == 10 & out[,'pop_correlation'] ==
0.10),'estimate'],
     main = 'Correlation = 0.10, Sample Size = 10',
     xlab = 'Significant Correlations',
     col = 'red', breaks = seq(from = -1, to = 1, by = .05))
hist(out[which(out[,'is_sig'] == 1 & out[,'sample_size'] == 70 & out[,'pop_correlation'] ==
0.10),'estimate'],
     main = 'Correlation = 0.10, Sample Size = 70',
     xlab = 'Significant Correlations',
     col = 'red', breaks = seq(from = -1, to = 1, by = .05))
hist(out[which(out[,'is_sig'] == 1 & out[,'sample_size'] == 10 & out[,'pop_correlation'] ==
0.50),'estimate'],
     main = 'Correlation = 0.50, Sample Size = 10',
     xlab = 'Significant Correlations',
     col = 'red', breaks = seq(from = -1, to = 1, by = .05))
hist(out[which(out[,'is_sig'] == 1 & out[,'sample_size'] == 70 & out[,'pop_correlation'] ==
0.50),'estimate'],
     main = 'Correlation = 0.50, Sample Size = 70',
     xlab = 'Significant Correlations',
     col = 'red', breaks = seq(from = -1, to = 1, by = .05))
```

Appendix D. Standard Errors and Disattenuated Regression

Devlieger et al (2016) examined four different alternatives for conducting hypothesis testing using factor score regression: a regression factor score approach, which uses the regression predictor to compute factor scores, a second approach which uses the Bartlett predictor to do the same, a bias avoiding method which uses the regression predictor for the independent variable and the Bartlett predictor for the dependent variable, and a bias correcting method which uses either of these predictors and then uses the variances and covariances of the factor scores to compute those of the true latent variable scores, which are then used to calculate the regression coefficients. The first three approaches perform a linear regression, for which the standard error formula is well-known; however, this is not the case for the bias correcting method. Specifically, the standard error formula refers to the uncorrected regression coefficient and employing it with the corrected regression coefficient could introduce bias and result in an incorrect t value. Our disattenuated regression approach is similar to the bias correcting method employed by Devlieger et al (2016) and, therefore, it is worth considering the impact of the issue for our research.

We first consider the difference between the standard deviation of the regression coefficients in our research, within a given condition, with the average standard error from those, which was computed using the standard formula (following Devlieger et al, 2016, the standard deviation of the coefficients within a given condition is their empirical standard deviation, while the average standard error of those is their mean standard error; see Table 3 in their work). The results shown in Table D1 indicate the presence of differences between the two when the measurement conditions are the poorest in our research design (smallest sample size, fewest and weakest indicators). However, as those improve, there is no noticeable difference between the two quantities.

Second, we examined an alternative approach to establishing the significance of the regression coefficients for the case of disattenuated regression. Instead of computing the standard error using the standard formula, we bootstrapped each replication 1,000 times. From those bootstrapped replicates, we computed both a standard error as well as 95% confidence intervals, using the percentile approach, the same process recommended for significance testing in PLSc (e.g., Aguirre-Urreta & Rönkkö, 2018). Table D2 compares the standard deviation of the regression coefficients within a given condition, as shown in Table D1, with the average standard error from the bootstrap replicates. From the results it is clear that, under the weakest of measurement conditions, there are very marked differences between the two quantities, with the average standard error from the bootstrap replicates being a multiple of standard deviation of the regression coefficients, which we attribute to the presence of extreme values in small samples due to the disattenuation process. However, for sample sizes of 80 and above, there are no noticeable differences.

In addition, we calculated the average relative bias (as a percentage of the original value) for each relationship in the research model, for the case when all four regression coefficients were significant (as was done in the main body of our research). These results, obtained from a 95% percentile confidence interval using bootstrap replicates, are not qualitatively different from those shown in Table A2, highlighting that the marked bias due to only considering significant results when sample sizes are small is not solely due to the approach used to establish statistical significance, but rather it is a systematic issue resulting from the combination of a preference for significant results with obtaining those from small samples. Average coefficient estimates and relative bias for each one of those, by simulation condition, are reported in Table D3.

Table D1. Standard Error Bias in Disattenuated Regression

Sample Size	Indics.	Loadings	SE Diff. A → C	SE Diff. A → D	SE Diff. B → C	SE Diff. C → D
20	3	0.7	0.15	0.46	0.13	0.44
40	3	0.7	0.09	0.07	0.09	0.05
60	3	0.7	0.07	0.06	0.06	0.05
80	3	0.7	0.05	0.05	0.04	0.04
100	3	0.7	0.05	0.05	0.04	0.03
20	6	0.7	0.08	0.07	0.07	0.04
40	6	0.7	0.04	0.03	0.03	0.03
60	6	0.7	0.03	0.04	0.03	0.02
80	6	0.7	0.03	0.03	0.02	0.01
100	6	0.7	0.03	0.02	0.02	0.01
20	9	0.7	0.04	0.04	0.04	0.02

Sample Size	Indics.	Loadings	SE Diff. A → C	SE Diff. A → D	SE Diff. B → C	SE Diff. C → D
40	9	0.7	0.03	0.03	0.02	0.01
60	9	0.7	0.02	0.02	0.02	0.01
80	9	0.7	0.02	0.02	0.01	0.00
100	9	0.7	0.02	0.01	0.01	0.00
20	3	0.8	0.08	0.07	0.06	0.05
40	3	0.8	0.05	0.04	0.05	0.02
60	3	0.8	0.04	0.03	0.03	0.02
80	3	0.8	0.03	0.02	0.02	0.02
100	3	0.8	0.03	0.03	0.02	0.01
20	6	0.8	0.04	0.04	0.02	0.01
40	6	0.8	0.02	0.02	0.02	0.01
60	6	0.8	0.02	0.02	0.02	0.00
80	6	0.8	0.02	0.01	0.01	(0.00)
100	6	0.8	0.01	0.01	0.01	0.00
20	9	0.8	0.03	0.03	0.01	0.00
40	9	0.8	0.02	0.01	0.01	(0.00)
60	9	0.8	0.01	0.01	0.01	(0.00)
80	9	0.8	0.01	0.01	0.01	(0.00)
100	9	0.8	0.01	0.01	0.01	(0.00)
20	3	0.9	0.02	0.03	0.02	0.01
40	3	0.9	0.01	0.02	0.01	(0.00)
60	3	0.9	0.01	0.01	0.01	0.00
80	3	0.9	0.01	0.01	0.01	(0.00)
100	3	0.9	0.01	0.01	0.01	(0.00)
20	6	0.9	0.02	0.02	0.00	(0.01)
40	6	0.9	0.01	0.00	0.00	(0.01)
60	6	0.9	0.01	0.01	0.01	(0.00)
80	6	0.9	0.01	0.01	0.00	(0.01)
100	6	0.9	0.01	0.00	0.00	(0.01)
20	9	0.9	0.01	0.00	0.00	(0.02)
40	9	0.9	0.01	0.01	0.00	(0.01)
60	9	0.9	0.00	0.00	0.00	(0.01)
80	9	0.9	0.00	0.00	0.00	(0.01)
100	9	0.9	0.00	0.00	0.00	(0.01)
20	3	Mixed	0.07	0.06	0.07	0.04
40	3	Mixed	0.03	0.04	0.03	0.02
60	3	Mixed	0.03	0.03	0.03	0.02
80	3	Mixed	0.02	0.03	0.02	0.01
100	3	Mixed	0.02	0.03	0.02	0.01
20	6	Mixed	0.03	0.03	0.03	0.01
40	6	Mixed	0.02	0.03	0.01	0.01
60	6	Mixed	0.02	0.02	0.01	0.00
80	6	Mixed	0.01	0.01	0.01	0.00
100	6	Mixed	0.01	0.01	0.01	0.00
20	9	Mixed	0.03	0.03	0.01	0.00
40	9	Mixed	0.02	0.01	0.01	(0.00)
60	9	Mixed	0.01	0.01	0.01	(0.00)
80	9	Mixed	0.01	0.01	0.01	(0.00)
100	9	Mixed	0.01	0.01	0.01	(0.00)

Note: Reported values are the difference between the standard deviation of coefficient estimates for a given condition and the average of the standard errors for that coefficient, within the same condition.

Table D2. SE Bias from Bootstrapping

Sample Size	Indics.	Loadings	A → C		A → D		B → C		C → D	
			StDev	Mean	StDev	Mean	StDev	Mean	StDev	Mean
20	3	Mixed	0.294	2.127	0.208	0.623	0.293	2.126	0.186	0.603
20	3	0.7	0.322	3.062	0.229	1.623	0.321	3.060	0.198	1.607
20	3	0.8	0.340	2.197	0.207	5.443	0.332	2.194	0.186	5.427
20	3	0.9	0.260	0.461	0.185	0.231	0.262	0.459	0.161	0.216
20	6	Mixed	0.268	1.171	0.193	0.343	0.267	1.168	0.167	0.328
20	6	0.7	0.308	2.106	0.212	1.343	0.301	2.098	0.186	1.398
20	6	0.8	0.272	0.957	0.191	0.338	0.263	0.949	0.174	0.322
20	6	0.9	0.257	0.515	0.178	0.184	0.243	0.513	0.154	0.166
20	9	Mixed	0.273	0.456	0.182	0.237	0.257	0.453	0.165	0.215
20	9	0.7	0.279	3.436	0.201	0.621	0.271	3.198	0.169	0.643
20	9	0.8	0.264	0.421	0.179	0.255	0.255	0.415	0.163	0.252
20	9	0.9	0.248	0.252	0.169	0.192	0.243	0.249	0.149	0.176
40	3	Mixed	0.197	0.228	0.150	0.154	0.199	0.226	0.130	0.139
40	3	0.7	0.239	1.216	0.180	0.323	0.238	1.215	0.158	0.308
40	3	0.8	0.209	0.220	0.149	0.159	0.213	0.217	0.130	0.143
40	3	0.9	0.177	0.180	0.130	0.127	0.177	0.177	0.110	0.113
40	6	Mixed	0.186	0.180	0.138	0.129	0.177	0.178	0.119	0.114
40	6	0.7	0.204	0.228	0.143	0.154	0.197	0.226	0.130	0.139
40	6	0.8	0.188	0.182	0.126	0.132	0.182	0.179	0.117	0.116
40	6	0.9	0.177	0.169	0.115	0.118	0.169	0.166	0.107	0.104
40	9	Mixed	0.182	0.173	0.120	0.123	0.174	0.170	0.112	0.109
40	9	0.7	0.195	0.186	0.139	0.135	0.190	0.186	0.122	0.120
40	9	0.8	0.182	0.173	0.119	0.123	0.176	0.171	0.112	0.109
40	9	0.9	0.174	0.164	0.118	0.113	0.170	0.164	0.104	0.100
60	3	Mixed	0.166	0.164	0.122	0.123	0.159	0.164	0.112	0.109
60	3	0.7	0.196	0.218	0.146	0.156	0.184	0.217	0.128	0.143
60	3	0.8	0.170	0.165	0.122	0.123	0.161	0.164	0.109	0.111
60	3	0.9	0.149	0.144	0.105	0.103	0.142	0.143	0.095	0.091
60	6	Mixed	0.151	0.146	0.108	0.104	0.148	0.145	0.095	0.093
60	6	0.7	0.164	0.162	0.127	0.121	0.167	0.160	0.108	0.107
60	6	0.8	0.151	0.146	0.113	0.106	0.152	0.145	0.096	0.093
60	6	0.9	0.142	0.137	0.102	0.096	0.143	0.135	0.087	0.084
60	9	Mixed	0.145	0.140	0.106	0.100	0.146	0.139	0.091	0.087
60	9	0.7	0.153	0.151	0.110	0.109	0.153	0.148	0.097	0.096
60	9	0.8	0.146	0.140	0.107	0.100	0.147	0.139	0.091	0.087
60	9	0.9	0.138	0.135	0.095	0.093	0.138	0.133	0.083	0.081
80	3	Mixed	0.138	0.139	0.106	0.104	0.137	0.137	0.092	0.093
80	3	0.7	0.168	0.167	0.125	0.131	0.161	0.164	0.116	0.119
80	3	0.8	0.145	0.140	0.103	0.105	0.136	0.138	0.095	0.093
80	3	0.9	0.123	0.123	0.090	0.088	0.122	0.122	0.078	0.078
80	6	Mixed	0.130	0.125	0.093	0.090	0.127	0.124	0.080	0.079
80	6	0.7	0.143	0.137	0.105	0.103	0.138	0.136	0.089	0.091
80	6	0.8	0.132	0.125	0.093	0.091	0.127	0.124	0.078	0.080
80	6	0.9	0.124	0.117	0.084	0.082	0.120	0.116	0.071	0.072
80	9	Mixed	0.127	0.120	0.088	0.086	0.123	0.119	0.074	0.075
80	9	0.7	0.132	0.129	0.094	0.093	0.130	0.126	0.083	0.083
80	9	0.8	0.128	0.120	0.088	0.086	0.123	0.119	0.074	0.075
80	9	0.9	0.120	0.116	0.080	0.080	0.117	0.114	0.072	0.071
100	3	Mixed	0.127	0.123	0.096	0.093	0.124	0.122	0.082	0.081
100	3	0.7	0.150	0.147	0.118	0.116	0.147	0.145	0.101	0.105

Sample Size	Indics.	Loadings	A → C		A → D		B → C		C → D	
			StDev	Mean	StDev	Mean	StDev	Mean	StDev	Mean
100	3	0.8	0.129	0.124	0.095	0.093	0.125	0.122	0.080	0.083
100	3	0.9	0.114	0.110	0.081	0.079	0.111	0.109	0.069	0.068
100	6	Mixed	0.115	0.112	0.080	0.080	0.112	0.111	0.071	0.070
100	6	0.7	0.128	0.123	0.090	0.091	0.125	0.120	0.080	0.080
100	6	0.8	0.118	0.112	0.079	0.080	0.115	0.110	0.070	0.070
100	6	0.9	0.111	0.106	0.072	0.073	0.108	0.104	0.064	0.064
100	9	Mixed	0.114	0.108	0.075	0.076	0.110	0.106	0.066	0.066
100	9	0.7	0.118	0.115	0.083	0.083	0.114	0.113	0.072	0.073
100	9	0.8	0.114	0.108	0.075	0.076	0.111	0.106	0.067	0.067
100	9	0.9	0.107	0.104	0.071	0.072	0.103	0.102	0.061	0.062

Note: Reported values are the standard deviation of coefficient estimates for a given condition and the average of the standard errors for that coefficient, within the same condition. The standard errors were calculated from bootstrap resamples.

Table D3. Significant Estimation Bias (Bootstrapping)

Sample Size	Indics.	Loadings	A → C		A → D		B → C		C → D	
			Mean	Bias	Mean	Bias	Mean	Bias	Mean	Bias
20	3	Mixed								
20	3	0.7								
20	3	0.8								
20	3	0.9	(0.456)	-1012%	0.484	38%	0.479	139%	0.779	30%
20	6	Mixed	(0.511)	-1121%	0.605	73%	0.411	106%	0.810	35%
20	6	0.7								
20	6	0.8	0.296	493%	0.309	-12%	(0.093)	-147%	0.414	-31%
20	6	0.9	(0.453)	-1006%	0.508	45%	0.622	211%	0.646	8%
20	9	Mixed								
20	9	0.7	0.167	235%	0.350	0%	0.280	40%	0.598	0%
20	9	0.8								
20	9	0.9	(0.218)	-535%	0.544	55%	0.520	160%	0.722	20%
40	3	Mixed	(0.488)	-1076%	0.488	39%	0.593	196%	0.627	4%
40	3	0.7	0.230	360%	0.368	5%	0.133	-33%	0.607	1%
40	3	0.8	(0.321)	-742%	0.511	46%	0.609	204%	0.635	6%
40	3	0.9	(0.183)	-466%	0.480	37%	0.352	76%	0.619	3%
40	6	Mixed	(0.071)	-243%	0.413	18%	0.313	57%	0.546	-9%
40	6	0.7	(0.447)	-995%	0.490	40%	0.616	208%	0.633	5%
40	6	0.8	(0.254)	-608%	0.417	19%	0.480	140%	0.598	0%
40	6	0.9	(0.174)	-448%	0.404	15%	0.451	126%	0.582	-3%
40	9	Mixed	(0.159)	-418%	0.404	15%	0.448	124%	0.596	-1%
40	9	0.7	(0.246)	-591%	0.447	28%	0.500	150%	0.647	8%
40	9	0.8	(0.180)	-460%	0.408	17%	0.455	128%	0.592	-1%
40	9	0.9	(0.164)	-429%	0.422	21%	0.440	120%	0.608	1%
60	3	Mixed	(0.211)	-521%	0.412	18%	0.411	105%	0.650	8%
60	3	0.7	0.066	33%	0.358	2%	0.132	-34%	0.582	-3%
60	3	0.8	(0.174)	-448%	0.435	24%	0.437	119%	0.622	4%
60	3	0.9	(0.103)	-306%	0.384	10%	0.362	81%	0.632	5%
60	6	Mixed	(0.126)	-353%	0.376	7%	0.392	96%	0.609	1%
60	6	0.7	(0.127)	-354%	0.443	27%	0.343	72%	0.627	4%
60	6	0.8	(0.099)	-298%	0.389	11%	0.341	70%	0.631	5%
60	6	0.9	(0.112)	-323%	0.360	3%	0.366	83%	0.636	6%

Sample Size	Indics.	Loadings	A → C		A → D		B → C		C → D	
			Mean	Bias	Mean	Bias	Mean	Bias	Mean	Bias
60	9	Mixed	(0.117)	-333%	0.363	4%	0.373	87%	0.637	6%
60	9	0.7	(0.208)	-515%	0.404	16%	0.422	111%	0.656	9%
60	9	0.8	(0.144)	-388%	0.368	5%	0.375	87%	0.638	6%
60	9	0.9	(0.030)	-159%	0.343	-2%	0.391	95%	0.620	3%
80	3	Mixed	0.038	-23%	0.359	2%	0.367	84%	0.603	0%
80	3	0.7	(0.293)	-686%	0.443	27%	0.453	127%	0.643	7%
80	3	0.8	(0.072)	-244%	0.404	15%	0.358	79%	0.582	-3%
80	3	0.9	0.118	136%	0.318	-9%	0.340	70%	0.605	1%
80	6	Mixed	(0.002)	-105%	0.396	13%	0.276	38%	0.600	0%
80	6	0.7	(0.073)	-246%	0.358	2%	0.347	73%	0.613	2%
80	6	0.8	(0.027)	-155%	0.337	-4%	0.312	56%	0.606	1%
80	6	0.9	(0.001)	-103%	0.328	-6%	0.291	46%	0.592	-1%
80	9	Mixed	(0.003)	-106%	0.341	-3%	0.292	46%	0.598	0%
80	9	0.7	(0.045)	-190%	0.378	8%	0.369	84%	0.585	-2%
80	9	0.8	0.011	-78%	0.352	1%	0.312	56%	0.599	0%
80	9	0.9	0.052	4%	0.356	2%	0.327	64%	0.582	-3%
100	3	Mixed	0.060	20%	0.380	9%	0.316	58%	0.589	-2%
100	3	0.7	(0.144)	-388%	0.431	23%	0.380	90%	0.696	16%
100	3	0.8	0.006	-88%	0.398	14%	0.336	68%	0.622	4%
100	3	0.9	0.120	140%	0.366	4%	0.291	45%	0.580	-3%
100	6	Mixed	0.091	83%	0.358	2%	0.298	49%	0.594	-1%
100	6	0.7	0.072	45%	0.355	2%	0.307	54%	0.610	2%
100	6	0.8	0.066	32%	0.337	-4%	0.316	58%	0.618	3%
100	6	0.9	0.077	55%	0.343	-2%	0.297	49%	0.602	0%
100	9	Mixed	0.070	41%	0.340	-3%	0.305	52%	0.605	1%
100	9	0.7	0.079	57%	0.368	5%	0.301	50%	0.604	1%
100	9	0.8	0.095	89%	0.341	-3%	0.303	51%	0.598	0%
100	9	0.9	0.131	162%	0.355	2%	0.264	32%	0.594	-1%

Note: Reported values, for each relationship in the research model, are the average estimate and the relative bias over all replications within a given condition where all coefficients were significant. An estimate was deemed significant when zero was not included in a 95% percentile confidence interval.

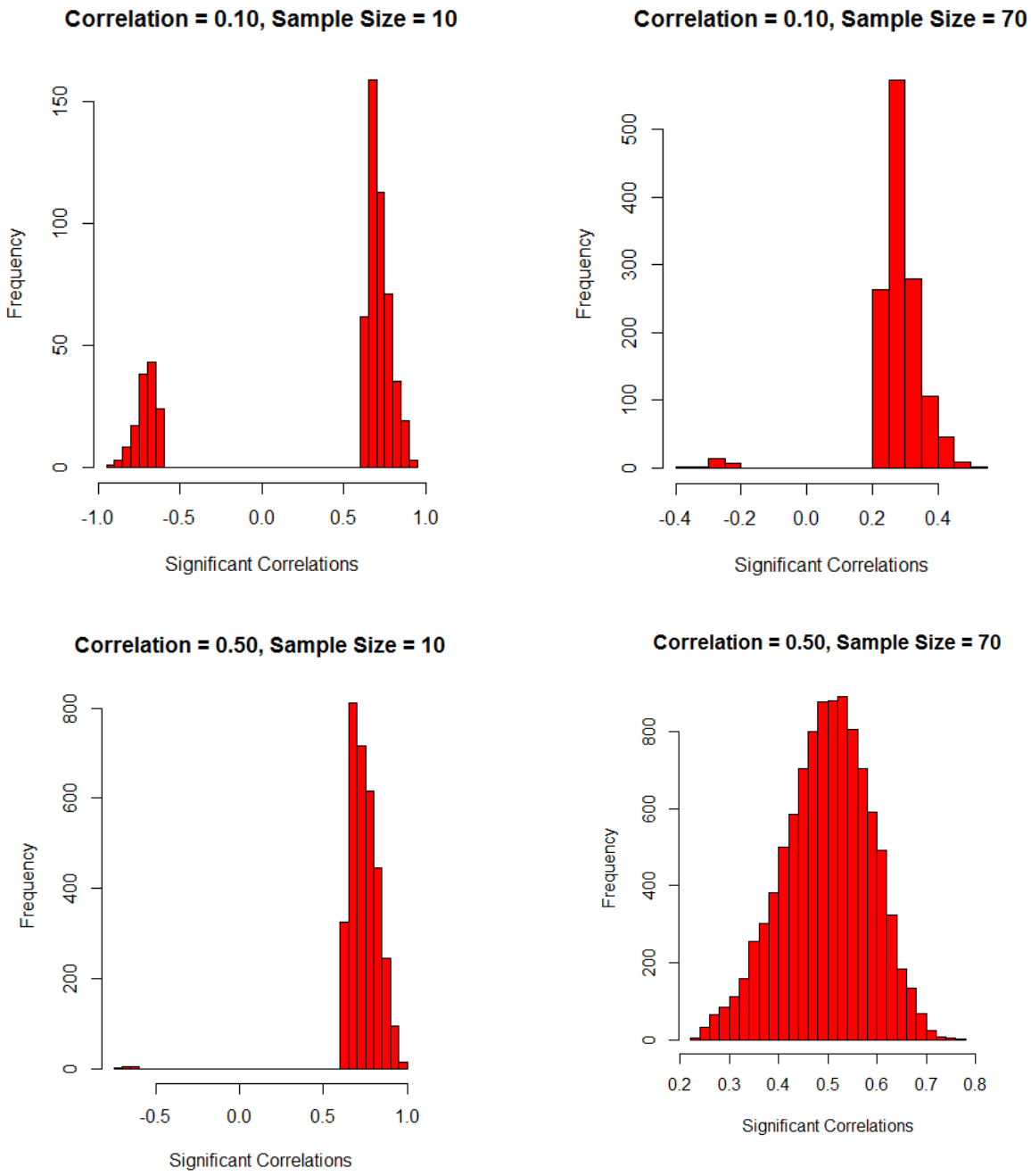

Figure 1. Significant Correlations Distribution (Sample Cases)

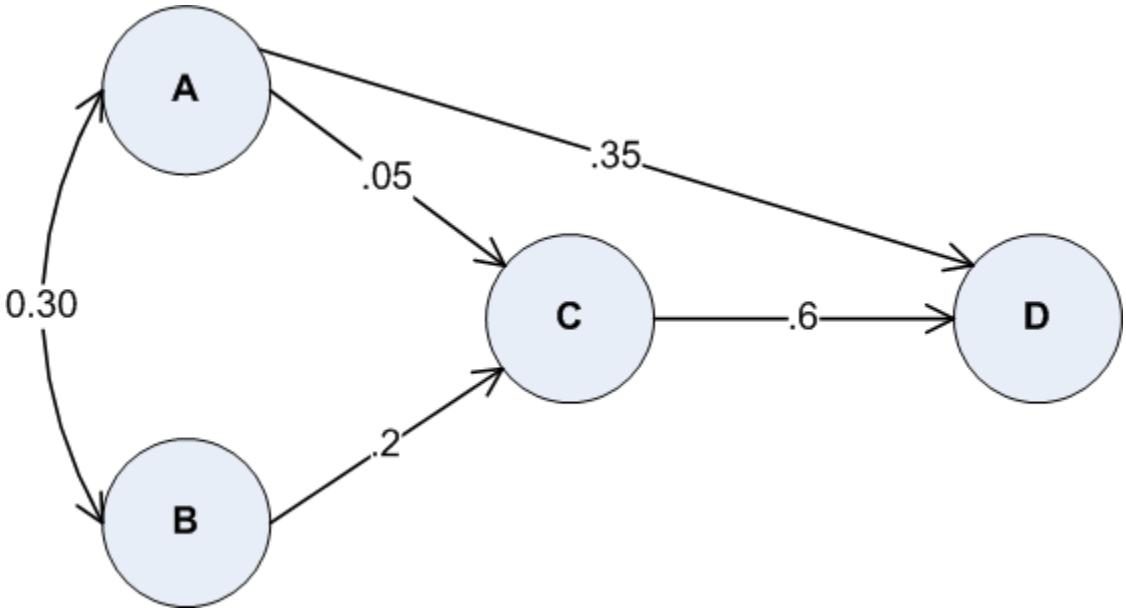

Figure 2. Population Model (Structural Portion)

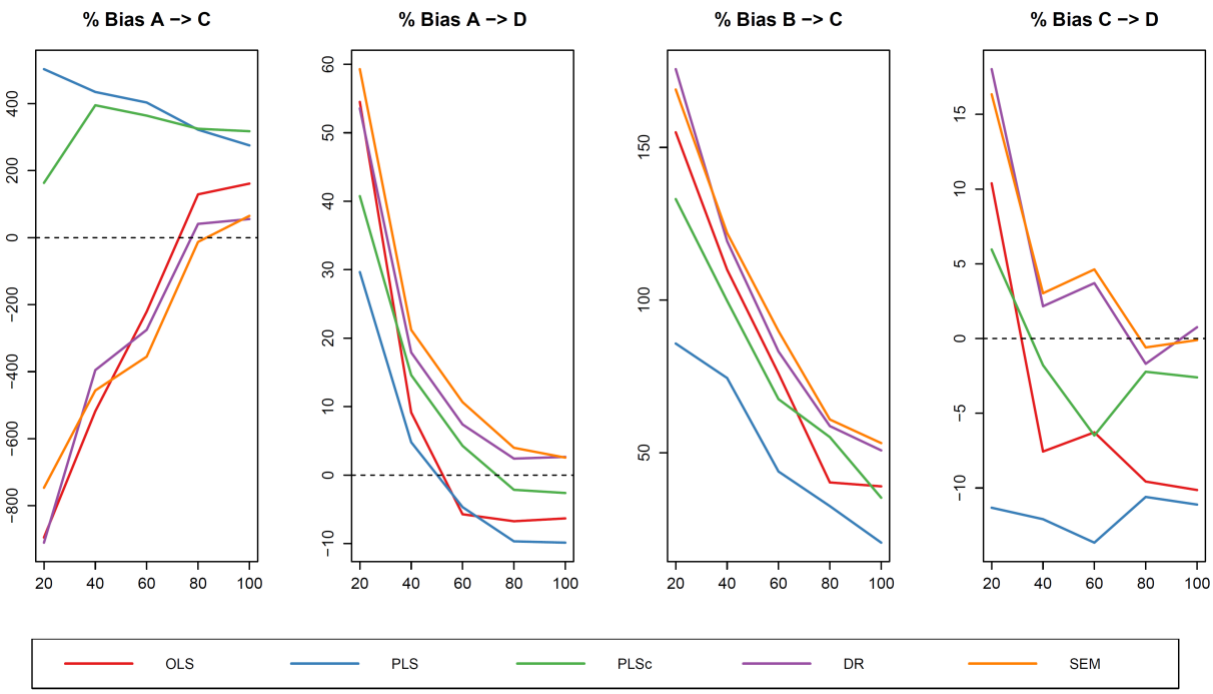

Figure 3. Estimation Bias by Path and Statistical Approach over Sample Size

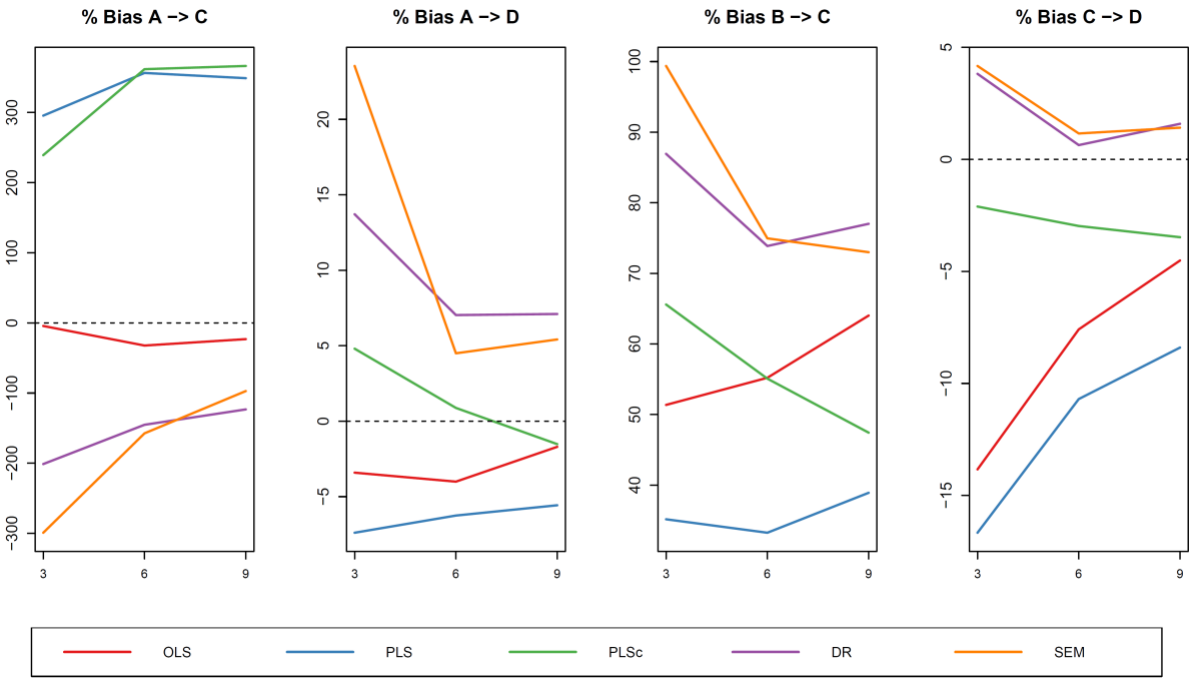

Figure 4. Estimation Bias by Path and Statistical Approach over Number of Indicators

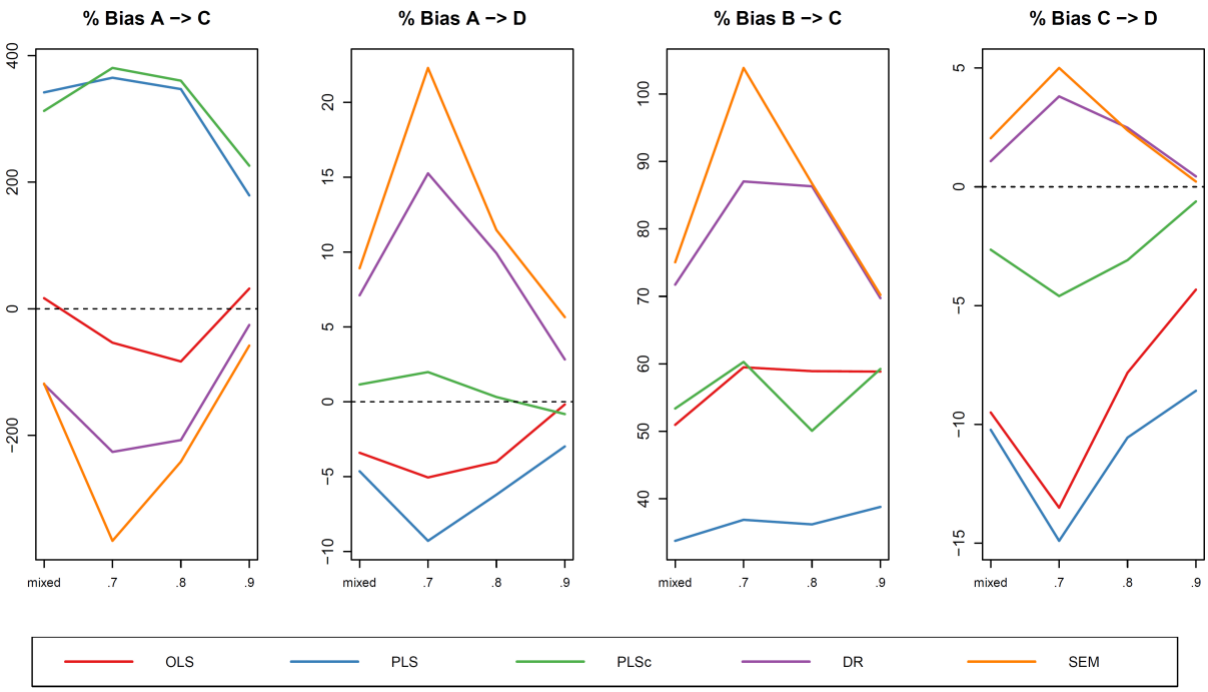

Figure 5. Estimation Bias by Path and Statistical Approach over Loading Pattern

Table 1. Average Sample Correlation, Relative Frequency of Significant Correlations, and Average of Significant Sample Correlation

Sample size	Population Correlation								
	0			.1			.2		
	Mean (All)	% Sig.	Mean (Sig.)	Mean (All)	% Sig.	Mean (Sig.)	Mean (All)	% Sig.	Mean (Sig.)
10	0.007	5.0 %	0.044	0.102	6.0 %	0.396	0.191	8.3 %	0.619
20	0.002	5.2 %	0.015	0.095	6.9 %	0.387	0.196	13.7 %	0.517
30	0.000	4.8 %	-0.030	0.097	7.4 %	0.356	0.197	19.0 %	0.440
40	-0.001	5.0 %	-0.012	0.100	9.6 %	0.347	0.197	24.4 %	0.393
50	0.002	5.3 %	0.006	0.102	10.7 %	0.311	0.196	28.2 %	0.358
60	0.000	5.2 %	-0.005	0.099	11.9 %	0.291	0.201	34.7 %	0.333
70	0.002	5.1 %	0.012	0.100	13.0 %	0.282	0.199	38.6 %	0.313
Sample size	**Population Correlation**								
	.3			.4			.5		
	Mean (All)	% Sig.	Mean (Sig.)	Mean (All)	% Sig.	Mean (Sig.)	Mean (All)	% Sig.	Mean (Sig.)
10	0.282	12.8 %	0.696	0.385	21.0 %	0.734	0.482	32.8 %	0.741
20	0.290	24.1 %	0.546	0.390	42.9 %	0.565	0.491	64.7 %	0.598
30	0.295	36.8 %	0.465	0.391	59.9 %	0.496	0.493	83.0 %	0.540
40	0.297	47.8 %	0.417	0.395	74.4 %	0.456	0.496	92.7 %	0.516
50	0.298	57.7 %	0.388	0.398	83.5 %	0.436	0.496	96.8 %	0.505
60	0.297	65.4 %	0.366	0.400	90.2 %	0.422	0.497	98.7 %	0.501
70	0.299	72.7 %	0.351	0.396	93.7 %	0.410	0.497	99.6 %	0.498

Note: Correlation over 10,000 simulated samples. Mean of all correlations and only significant (Sig.) correlations. % Sig refers to either Type I Error (for those conditions where the true correlation is 0) and statistical power otherwise.

Table 2. Statistical Power by Simulation Condition

Condition	OLS	DR	PLS	PLSc	SEM
(a) Sample Size					
20	0.14%	1.13%	0.24%	0.09%	0.59%
40	0.65%	2.60%	1.00%	0.70%	1.97%
60	1.01%	2.88%	1.91%	1.33%	1.78%
80	1.51%	3.51%	2.96%	2.49%	2.19%
100	2.33%	4.65%	3.68%	3.05%	2.95%
(b) Number of Indicators					
3	1.06%	4.03%	1.67%	0.95%	2.10%
6	1.12%	2.68%	2.09%	1.65%	2.24%
9	1.17%	2.15%	2.13%	2.20%	2.24%
(c) Loadings					
.7	0.96%	4.26%	2.30%	1.08%	2.03%
.8	1.19%	2.94%	2.34%	1.77%	2.13%
.9	1.25%	1.83%	0.67%	1.78%	2.22%
Mixed	1.12%	2.78%	2.53%	1.67%	2.37%

Note: Percentages shown are the number of replications where all four estimates were significant divided by the total number of replications in each level for a given simulation condition.