The Issue of Statistical Power for Overall Model Fit in Evaluating Structural Equation Models: Examples from Industrial-Organizational Psychology Research

Richard Hermida, Joseph N. Luchman, Vias Nicolaides, & Cristina Wilcox

George Mason University

Abstract

Statistical power is an important concept for psychological research. However, examining the power of a structural equation model (SEM) is rare in practice. This article provides an accessible review of the concept of statistical power for the Root Mean Square Error of Approximation (RMSEA) index of overall model fit in structural equation modeling. By way of example, we examine the current state of power in the literature by reviewing studies in top Industrial-Organizational (I/O) Psychology journals using SEMs. Results indicate that in many studies, power is very low, which implies acceptance of invalid models. Additionally, we examined methodological situations which may have an influence on statistical power of SEMs. Results showed that power varies significantly as a function of model type and whether or not the model is the main model for the study. Finally, results indicated that power is significantly related to model fit statistics used in evaluating SEMs. The results from this quantitative review imply that researchers should be more vigilant with respect to power in structural equation modeling. We therefore conclude by offering methodological best practices to increase confidence in the interpretation of structural equation modeling results with respect to statistical power issues.

Keywords: statistical power, structural equation modeling, measurement, RMSEA

The Issue of Statistical Power for Overall Model Fit in Evaluating Structural Equation Models: Examples from Industrial-Organizational Psychology Research

Structural equation modeling (SEM) is an increasingly popular analysis framework in many areas of scientific inquiry, including psychology, management, and sociology (MacCallum & Austin, 2000; Swanson & Holton, 2005). However popular, SEM is highly complex and its statistical mechanics are often not well understood by users. As a result, SEM has the potential to be misapplied, affecting the interpretation of scientific findings. For example, researchers often allow data to dictate which measurement errors should correlate, as opposed to appealing to *a priori theory*, which can make poorly fitting models appear "passable" (Hermida, Conjar, Najab, Kaplan, & Cortina, 2010; Landis, Edwards, & Cortina, 2009), or give inappropriate statements regarding causality without the backing of theoretical assumptions within SEM (Pearl, 2009; 2012).

While there are a number of methodological subtleties to SEM (Bagozzi & Yi, 2012; Bentler & Bonnett, 1980; MacCallum & Austin, 2000), one issue that has largely escaped the attention of SEM users is consideration of statistical power in SEM with respect to overall model fit. Understanding power in model fit is important because power reflects the probability that a model will differentiate between good and bad theory-implied constraints or specifications (Cohen, 1988; 1992). Since overall model fit is one of the main standards by which empirical evidence provided by structural equation models (SEMs) are judged, understanding issues related to power of SEM fit indices can provide the understanding necessary to effectively conduct SEM-based data analysis, improve inferences regarding SEMs, and increase the rate of scientific progress enjoyed by users of SEM.

The primary purposes of the present study are fivefold. First, we wish to inform researchers how sampling variability and power can potentially harm our inferences regarding our judgments of SEMs. Second, we wish to give sufficient background as to how power is calculated in a structural equation modeling context for overall model fit. Third, we wish to benefit researchers by explaining the main influencers of power so as to aid researchers in study design efforts. Fourth, we wish to examine certain methodological situations that could signal a need for the researcher to pay particular attention to statistical power. Fifth, we wish to conduct a quantitative review regarding power in order to a) gain understanding of the distribution of power as it exists in published journal articles, and b) test the degree to which power is associated with certain methodological situations.

The results of our review speak to the level of uncertainty related to the *decisions* scholars make about model fit. Thus our survey provides best practices to researchers in terms of "powering" their study to detect non-trivial problems related to model misfit. Our research is primarily based on MacCallum, Browne, and Sugawara (1996), who conducted the pioneering work in developing the concept of statistical power in overall model fit indexes. It is our hope that after reading this review, researchers will understand the statistics of power as it relates to overall model fit, and moreover be able to identify how some methodological issues might provide signals to the researcher regarding power of their tested models.

Overview of Power

The primary strength of SEM, and the root of its popularity, is in integrating the measurement model focus of factor analysis with a structural or theoretical model that has been the focus of path analytic or regression modeling. An important issue to both measurement and structural models is examining how well the model implied by theory fits to the data collected

(Kaplan, 1995; Specht, 1975). To the extent that a theoretical model fits empirical data, the theoretical model is confirmed, as it is a plausible explanation for the covariance structure amongst the variables (Mulaik et al., 1989). The issue of how to evaluate model fit is complicated, and opinions have yet to converge on the most appropriate method. As a result of different opinions related to how fit should be assessed (Barrett, 2007; Hayduk et al., 2007), a number of model fit indices have been developed for SEMs (Bagozzi & Yi, 1998; Bentler, 1990), each with different properties across a number of dimensions, such as absolute vs. relative fit (Mulaik, 2009).

Of the many structural equation model fit indices available in the literature, the RMSEA is a popular index of absolute fit (i.e., it is not *relative* to the null model as are indices such as the confirmatory fit index or CFI) and is noted for its insensitivity to estimator by comparison to relative fit indices (Sugawara & MacCallum,1993). The RMSEA is, most fundamentally, a function of the model chi-square value, but also includes the model degrees of freedom, and sample size as seen in Equation 1.

$$\text{RMSEA} = \frac{\sqrt{(\chi^2 - \text{df})}}{\sqrt{(\text{df})(\text{N} - 1)}} \qquad (1)$$

Important to note is that model degrees of freedom, for traditional maximum likelihood SEM with continuous factor indicators, are computed by obtaining the total number of elements in the variance-covariance matrix that can be analyzed minus the number of estimated parameters. Readers who wish to review these concepts should consult a more in-depth explanation of degrees of freedom by Rigdon (1994).

The functional form of the RMSEA in Equation 1 can be explained by noting first that nested within the RMSEA index is the assumption that the model being estimated is misspecified

to some extent and, consequently, the model chi-square statistic follows what is known as a *non-central chi-square distribution*. The non-central chi-square distribution can be thought of as the chi-square distribution when chi-square possesses any non-zero value (Patnaik, 1949). Relevant to this review, the non-central chi-square *distribution* comes into play when one wants to know the chance that chi-square exceeds a particular chi-square value when the true population value of chi-square is non-zero (Cox & Reid, 1987). Of chief concern is the noncentrality parameter, which is simply the parameter that occurs in a distribution that is a transformation of the normal distribution (like the non-central chi-square distribution), and how this parameter relates to power.

Let us walk through a brief statistical sample to illustrate these interactions. At this point, the reader is encouraged to walk through these steps in order to become more intimate with the procedure. It would also be helpful to have the seminal work (MacCallum et al., 1996) on hand for easy access to referenced graphs and figures, as well as tools to easily calculate power on hand (Preacher & Coffman, 2006).

Suppose a researcher tested a model with 20 degrees of freedom, a sample size of 200. Next suppose the researcher wanted to obtain the probability of rejecting the null hypothesis that the obtained chi-square would be equal to or less than RMSEA = .05 (i.e. – $\chi^2 = 29.95$ in this context), if the true value was RMSEA = .08 (i.e. – $\chi^2 = 45.47$ in this case), with alpha equal to the traditional .05 level.

Calculation of the noncentrality parameters associated with the null and alternative hypothesis is easily done via the following formula:

$$\lambda = (N-1)(df)(RMSEA^2) \qquad (2)$$

Therefore, in this example the noncentrality parameter associated with the null hypothesis (ncp0) is 9.95, and the noncentrality parameter associated with the alterative hypothesis (ncpa) is 25.47. It is important to note at this point that the extent to which the model is correctly specified, the model is better approximated by the "central" chi-square (which has a mean or expected value equal to the model degrees of freedom) and λ approaches 0. The more the model is misspecified is the degree to which the noncentrality parameter and central chi-square diverge (see figure 1, pg. 136; MacCallum et al., 1996). We will now turn to specific hypothesis tests taken from MacCallum et al. (1996) that will be relevant for our quantitative review.

Exact, Close and Not Close Fit. MacCallum et al. (1996) use the non-central chi-square distribution to propose three hypothesis tests which evaluate different aspects of model fit by assessing the degree of overlap between a pair of non-central chi-square distributions (i.e., the null and alternative distributions). The first test proposed by MacCallum et al. is a test of *exact fit*. Exact fit is analogous to the central chi-square test of model fit in that it evaluates whether a model's fit to the data is sufficiently good to be "exactly" as the specified model dictates (see footnote 1). The null value used for the exact fit test is not, however, 0 (i.e., no model discrepancies from the data), but rather some very small RMSEA value, such as .01. In the instance that the estimated RMSEA is sufficiently large—that is, large enough to be significantly different from a small value such as .01 (MacCallum et al., 1996)—then we can infer that the fit of the model is not likely to be exact.

A departure from the exact fit idea is proposed through the second test of *close fit*. Close fit differs from exact fit in that it evaluates whether the confidence interval about RMSEA centers around, but does not exceed, 0.05. Hence, the purpose of the close fit test is to evaluate a model in which the null hypothesis is .05, with an alternative hypothesis that is *larger* than .05—

suggesting that the RMSEA in the population is likely to be ≤ .05. Models that are not significantly larger than .05 are inferred to have a close fit, and although close fit is not exact, MacCallum et al. argue that the data approximates the model "closely" or well enough to be of use scientifically.

The final test proposed is *not close fit*. Not close fit supplements conceptual deficiencies in the previous tests by evaluating whether the estimated RMSEA ≥.05 therefore indicating that the model *is* likely to be a poor fit to the data. Similar to the test of close fit, the null distribution centers on RMSEA of .05, however the alternative distribution for not close fit is fixed at a value less than .05. Thus, as MacCallum et al. show (p. 136-8), the not-close fit test adds to the information provided by the close fit test by distinguishing between situations where the RMSEA's confidence interval falls relatively close to a value of .05. Specifically, based on the pattern of tests accepted and rejected the researcher can triangulate on the likely "true" RMSEA of the model. For example, when the test of not close fit is rejected and the test of close fit is accepted, a researcher can infer that the true RMSEA falls somewhere below .05. Alternatively, when close fit is rejected but not close fit is accepted, a researcher can infer that the true RMSEA falls above .05. A final possibility is that both tests are accepted, which suggests that RMSEA's confidence interval centers around .05. In combination, all three tests permit a researcher to evaluate the degree of model fit more flexibly than using only "rule of thumb" cut off values for fit indices or a single chi-square test, as the "three test" procedure admits to the idea that degrees of freedom and sample size play an important role in the precision of the estimates obtained using SEM and the level of uncertainty we have about their true values—which is thereby reflected in the confidence interval associated with the estimated model's fit (McQuitty, 2004).

Brief Review of Hypothesis Testing for RMSEA. The hypothesis testing procedure for tests of overall model fit in SEM differ slightly from the way hypothesis tests are structured for traditional statistics such as ANOVA F-tests or t-tests. To be precise, accepting the null is a sought after result in the case of close and exact fit. Moreover, accepting the null of a not close fit test does not imply *unacceptably poor* fit, but only *not good* fit. Whereas the interpretation of the tests differs, the logic of the hypothesis testing procedure does not differ from the usual procedure as outlined in introductory statistics texts.

Hypothesis testing for RMSEA proceeds by evaluating the distributions of two values of RMSEA—which for the sake of consistency with MacCallum, et al., we will, for the remainder of the present section, refer to as ϵ—the null value: ϵ_o and the alternative value: ϵ_a. Using ϵ_o and ϵ_a we then can compare the non-central chi-square distribution associated with ϵ_o to the non-central chi-square distribution associated with ϵ_a. The overlap between ϵ_o and ϵ_a is overall covariance model power—based on ϵ. The null hypothesis value for each of the tests (exact, close, or not close) we describe above (e.g., .01 for exact fit) and similar to differences between means in a t-test, the differences between ϵ_o and ϵ_a values can be conceptualized as the "effect size" component that factors into the power calculations for t-tests. When $\epsilon_o > \epsilon_a$, power is estimated as:

$$\pi = P(\chi_d^2, \lambda_a < \chi_c^2) \qquad (3)$$

Whereas when $\epsilon_o < \epsilon_a$, power is estimated as:

$$\pi = P(\chi_d^2, \lambda_a > \chi_c^2) \qquad (4)$$

In both cases, χ_c^2, χ_d^2,, and λ_a represent the non-central chi-square distributions associated with ϵ_o and ϵ_a and π represents statistical power . When $\epsilon_o > \epsilon_a$, the power is measured as the portion of χ_d^2, or λ_a, that lies to the left of alpha (α) in the left tail of χ_c^2. When $\epsilon_o < \epsilon_a$, power is measured as

the portion of χ_d^2, or λ_a, that lies to the right of the critical value α in the right tail ofχ_c^2. When ϵ = 0, the non-centrality parameter λ is also 0 (see Equation 3). In this case, ϵ is distributed as a regular, central chi-square and perfect fit is implied. We will now move to a discussion of statistical factors that influence power.

Factors Affecting Power

Closeness of null and alternative RMSEA values. One major area that influences power (all else equal), is the closeness of the RMSEA values associated with the null and alternative hypotheses. Closeness of null and alternative RMSEA values is *negatively* associated with power. That is, the closer the RMSEA values, the more power decreases.

This pattern occurs because to the degree that null and alternative RMSEA values are similar is the degree to which the non-central chi-square distributions overlap. To the degree the distributions overlap, is the degree to which there is a lack of ability to find area under the alternative non-central chi-square distribution that is beyond the critical value associated with the null hypothesis *and* not overlapping with the non-central chi-square distribution associated with the null. To use our previous examples, the ncp0 and ncpa associated with RMSEA = .05 and RMSEA = .08 are 9.95 and 25.47 (difference of 15.52) and generate a power coefficient of .45. If the null was moved to RMSEA = .07, the ncp0 would shift to 19.50 (difference of 5.97). This would cause the null and alternative distributions to move closer together, creating more overlap and less power (in this case .13). However, if the null was moved to RMSEA = .01, the ncp0 would shift to 0.40 (difference of 25.07). This would cause the null and alternative distributions to move further apart, creating less overlap and more power (in this case .88). The bottom line here is that the more differentiation there is between the null RMSEA and the alternative RMSEA, the more power will increase, all else being equal.

Sample Size. Sample size has a positive association with power. That is, as sample size increases, power increases. This is because as sample size increases, the ability for the null and alternative noncentrality parameters to separate themselves from one another increases as well. For example, in our running example with a sample size of 200, the difference between the ncp0 and ncpa was 15.52 (derived from 25.47-9.95), which equates to a power coefficient of 0.45. If the sample size is increased from 200 to 500, this difference increases to 38.92 (derived from 63.87-24.95), and drives power to 0.86. If the sample size decreased to 100, the difference between the parameters would drop to 7.72 (derived from 12.67-4.95), causing power to 0.24. This is because the individual noncentrality parameter is calculated via a multiplicative term involving degrees of freedom, sample size, and the square of the RMSEA null or alternative hypothesis in question (see equation 2). Ultimately, as sample size increases, power will increase as well, all else being equal.

Degree of Model Misfit. A more subtle influence on power is the degree of model misspecification. All else being equal, it is easier to obtain power as model misfit increases. For example, with df = 20, N = 200, RMSEA null = .00, and RMSEA alt = .05, power is only 0.40. However, if the null and alt RMSEA were .05 and .10, power would increase to 0.84. While these examples are contrived to illustrate the general principle, the general theme here is that the *more precise your models, and the more precise of a comparison you wish to make, the more difficult it is to obtain power*. Just as before, this occurs because the noncentrality parameters are connected to degrees of freedom, sample sizes, and RMSEA. For high RMSEA values, there is more *potential* for the ncp0 and ncpa to differ than for low values of RMSEA, holding all else equal. The bottom line here is that as the degree of model misfit increases, power will increase as well, all else being equal.

Degrees of Freedom. Finally, the relationship between degrees of freedom and power is positive. That is, as degrees of freedom increase, power increases as well. This relationship takes place on two different fronts. First, degrees of freedom impact obtained noncentrality parameters. As the degrees of freedom increase, so too does the noncentrality parameter. Applied to the current context, as degrees of freedom increase, the noncentrality parameters associated with the null and alternative hypotheses increase, but the *degree of difference* between the noncentrality parameters also increases. This is because the individual noncentrality parameter is calculated via a multiplicative term involving degrees of freedom, sample size, and the square of the RMSEA null or alternative hypothesis in question. For example, with df = 5, N = 200, RMSEA null = .05, and RMSEA alt = .08, power is only 20, with the ncpa and ncp0 possessing a difference of 3.88 (6.37-2.49). However, if the degrees of freedom were increased to 50, power would increase to 0.73, with the ncpa and ncp0 possessing a difference of 34.92 (57.31-22.39).

A more obscure way that degrees of freedom impact power is through manipulation of the shape of noncentral chi-square distributions. This is because the shape (variance) associated with the noncentral chi-square distribution is dependent on both degrees of freedom and the noncentrality parameter, as represented in the formula below:

$$2(df + 2\lambda) \qquad (3)$$

This can be seen further by examining the formulas associated with the skewness and kurtosis of the noncentral chi-square distribution, as shown in formulas 4 and 5, respectively:

$$\frac{2^{1.5} * (df + 3\lambda)}{(df + 2\lambda)^{1.5}} \qquad (4)$$

$$\frac{12(df + 4\lambda)}{(df + 2\lambda)^{1.5}} \qquad (5)$$

To the degree the noncentral chi-square distributions change, power will change as well, all else being equal. The bottom line here is that as the number of degrees of freedom increases, power will increase as well, all else being equal.

To summarize the described effects on power to detect model misfit, power increases as sample size, degrees of freedom, difference between null and alternative RMSEA values, and degree of model misfit increase. It is important to keep in mind that all of the aforementioned effects were described in the context of all other influences being held constant. It is critical to note that in reality, all of these elements interact with one another to produce statistical power, and consequently, it is possible to have several of the elements oriented towards low power, but to have a single element that is so strong as to compensate for the weakness of the other elements in producing power, or vice versa.

To give an extreme example in order to illustrate the principle, suppose a researcher sought to find power of a model that had extremely few degrees of freedom (5) and null and alternative RMSEA values that were both small *and* close together (.00 and .01). While these elements would generate low power in most situations, a researcher that obtained a sample size of 50,000 would obtain a power coefficient of .98. Conversely, suppose a researcher sought to find power about a model that had 70 degrees of freedom, a sample size of 500, and an alternative RMSEA value that had a fairly higher degree of misfit (.08). While these elements would generate high power in most situations, if the null RMSEA was set at .07, power would only be 0.53.

Ultimately, it is our desire for researchers to understand the main influencers of power, so that researchers can more easily ascertain power in different scenarios, appreciate what might

need to be done to obtain more power in a particular research setting, and ultimately engage in meaningful power analysis at the planning stages of the research process.

Rationale for Quantitative Review

Hypothesis testing and power estimation for overall model fit is conceptually identical to hypothesis testing for less technically-complicated statistical analyses such as bivariate correlation or analysis of variance (ANOVA). Unfortunately, until relatively recently, no computational algorithm or computer program has been available from the literature to allow practicing researchers to *easily* compute a priori power values for the RMSEA index (see Preacher & Coffman, 2006). Moreover, the technical documentation of the RMSEA hypothesis testing procedure was presented in a very technical way by MacCallum et al. (1996). We believe that the combined influence of both of these factors have contributed to the relative neglect of power vis-à-vis overall model fit in structural equation modeling.

As opposed to the use of hypothesis testing, SEMs in the literature are usually evaluated on the basis of commonly accepted point estimate "cut-offs" such as .05 for "excellent" or .08 for "adequate" fitting models (Chen, Curran, Bollen, Kirby, & Paxton, 2008). The use of cut-off values are necessary, however, using only cut-off values and omitting hypothesis tests altogether does not allow a researcher to account for sampling variability, as we note above. For example, our confidence in a model with a RMSEA of .06 and a standard error of .01 is very different from the same RMSEA value with a standard error of .05. In particular, a RMSEA with a smaller standard error will, in the long run, be more similar to the result just obtained than an *identical* RMSEA value with a larger standard error. Hence, our certainty about the true value of the RMSEA, and thus the true fit of the model to the data, is necessarily different.

An implication of the relative neglect of power in SEM is the potential for models to have acceptable RMSEA values, yet high levels of sampling variability—suggesting the possibility that the value a study's fit index obtained is merely due to chance. The issue of chance values of fit is an important one, as obtaining a RMSEA value that is deemed "adequate" in magnitude but not "adequate at beyond chance levels" in terms of its confidence intervals is much like a large correlation that is not sufficiently larger than 0 to be statistically significant. In our view, the current lack of attention to overall model fit sampling variability casts doubt on the fidelity of the results obtained in our literature for SEMs. Stated differently, owing to the neglect of power-related issues in SEM, it is possible that published research using SEMs does not have acceptable levels of power to differentiate failing from adequate models and, thus, do not have adequate power to make an accurate decision about model fit based on the data. Although it is possible that our SEMs in the organizational sciences do not have acceptable levels of power, and thus are not particularly informative about model fit, the extent of the problem is an empirical question. Therefore, in order to evaluate the possibility that SEMs in the organizational sciences do not offer adequate information about model fit, we conduct an extensive survey of the organizational science literature to ascertain the state of the field regarding power of SEMs in influential research from top-tier journals. Specifically, the current survey will consider many important aspects to the topic of power in SEM, such as the distribution of power across all studies included, differences in power between models that are ultimately deemed to be the best model in a study vs. those models that are not deemed to be the best model, differences in power across different types of model configurations (i.e., measurement vs. structural models), and we also incorporate information regarding sample size issues—as sample size is directly linked to statistical power. Finally, our study seeks to contribute to the discourse regarding good practice

in SEM, and therefore will conclude by make recommendations related to "best practices" for power in SEM.

Variables to be Reviewed

We have reason to believe that there are several critical issues to examine vis-à-vis power in structural equation modeling. We list these factors, with emphasis on why we think these factors are important to examine in the quantitative review. These are the exploratory variables we will review, as we do not have a specific hypothesis associated with these variables.

Distribution of Power. The first and most important attribute that we wish to review is the distribution of power across models in Industrial-Organizational Psychology journals. Of particular interest in the percentage of studies that exhibit low levels of power, which in this case would correspond to inadequately falsifiable models. This is important because models that lack falsifiability have the potential to lead researchers down incorrect paths with respect model fit and the understanding of psychological relationships.

Sample Size Issues. A second and related issue to be analyzed is the sample size associated with models to be tested in structural equation modeling. This is important because sample size is a somewhat controllable issue in model testing, and directly related to a model's ability to be falsified with respect to overall model fit. It is our suspicion that in some cases, models possess a sample size that is grossly inadequate to meet a reasonable level of falsifiability and power. If this is true, it could signal the needs for researchers to increase the sample size involved in testing their models, in order to achieve power that signifies a reasonable level of falsifiability.

Moderators

We have reason to believe that several methodological artifacts will have an influence on the overall power of tested models in this quantitative review. We will list these variables and associated hypotheses now, with emphasis on why we think these artifacts will influence power.

Model Type. We have reason to believe that the type of model researchers conducting structural equation modeling on will have an impact on the power of the tested model. Specifically, we believe that measurement models will possess higher levels of power as compared to models that are purely structural in nature. This is because measurement models tend to have greater degrees of freedom than structural models, owning primarily to the use of multiple indicators in modeling a single latent variable. It seems unlikely that on average, structural models will have enough variables to counteract the degrees of freedom obtained through the use of multiple indicators in measurement of latent variables. All else being equal, lower degrees of freedom for tested models equates to lower overall power. Therefore, our first hypothesis is:

Hypothesis 1: Power will be associated with model type such that measurement models will have significantly more power than structural models.

Main Models vs. Competing Models. We also have reason to believe that the model selection process in and of itself could lead to differing levels of power. Often, researchers will choose the best fitted model as their model of choice after adding pathways suggested by modification indices (MacCallum & Austin, 2000). Additionally, many researchers will judge models on the basis of overall model fit, and chose the best fitted model as their final model. If this is true, degrees of freedom in competing models will be lower than in the final model, which would in turn cause lower power, all else being equal. Additionally, if researchers are choosing models that have the best fit as their best models, it is possible that because power to detect

misfit is negatively associated with fit (all else being equal), that main models will have best fit, but only because of lower power. Our hypothesis is therefore:

Hypothesis 2: Power will be associated with main models such that main models will have significantly less power than competing models.

Teams/Groups Models vs. Other Models. The previous two hypotheses dealt with methodological artifacts that could reduce power via reduction in the number of degrees of freedom of the tested model. We believe that power is also susceptible to being lower in models that deal with team and group topics than other types of models through sample size reduction. It is common in teams/groups studies to reduce the tested sample size in terms of N, because team/group variables usually required the original sample size to be divided by some factor in order to produce teams or groups to study. For example, a sample size of 600 *individuals* might be reduced to 200 *teams* composed of three individuals each. This process means that it is much harder for these studies to obtain sample sizes on which the final tested model will have an appreciably high N, than non-group/teams studies where the division of sample size does not take place, all else being equal. Because sample size is a factor that influences power, we hypothesize that:

Hypothesis 3: Power will be associated with teams/groups models such that teams/groups models will have significantly less power than non-teams/groups models.

Fit Index Values. A third issue to be analyzed is the relationship between fit indices and power in models published in I/O Psychology journals. This is important because model fit index values are one of the main standards on which the value of a model is judged. To the degree that model fit is associated with power is the degree to which there is potential for fit indices to be artificially high and not reflective of true population values. Logically, because fit indices are all

based on the degree of model misfit in some way, it follows that the degree to which the model is powered to detect misfit is the degree to which the model fit index will be worsened. It could be the case that for some models, fit is only seen as acceptable because the power to detect misfit is low. This review seeks to quantify these issues. We therefore hypothesize:

Hypothesis 4: Power will be associated with fit index quality such that as power increases, fit index quality decreases.

Journal Quality. Finally, we believe that journal quality may be related to overall power in that higher quality journals may be associated with more highly powered models. If higher powered models are more scientifically sound, and higher quality journals publish more scientifically sound researcher than lower quality journals, then there seems to be reason to believe that all else being equal, higher quality journals would present findings related to more falsifiable models. Our specific hypothesis is:

Hypothesis 5: Power will be associated with journal quality such that as journal quality increases, model power increases.

Method

Sample of Studies

As the goal of the present work was to review trends of power in Organizational Psychology we conducted a comprehensive literature search of studies using SEM in journals frequently referenced in organizational psychology. Literature searches were conducted using the PSYCINFO, ProQuest, ERIC, AB-INFORM databases. Journals included were *Journal of Applied Psychology*, *Personnel Psychology*, *Academy of Management Journal*, *Human Performance*, *International Journal of Selection and Assessment*, *Journal of Management*, *Journal of Organizational Behavior*, *Organizational Behavior and Human Decision Processes*,

and *Journal of Vocational Behavior.* As such, we only included studies using SEM from each of these 9 journals. We limited our search from 1996 to 2012, as 1996 is the year in which MacCallum et al.'s (1996) article was disseminated to the research community. In all databases, we used the keywords "covariance models," "confirmatory factor analysis," "structural equation modeling," and "SEM" to identify articles that used SEM.

Selection Criteria

In order to be included in the present quantitative review, each study was required to report information needed to ascertain power of at least one model. Specifically, degrees of freedom and sample size for a structural equation model estimated was necessary. We also collected information on overall fit indexes such as the chi-square, RMSEA, CFI, and NFI. We identified a total of 365 studies across the 9 journals that met initial inclusion criteria. However, after reviewing each of the articles we excluded 25 for a grand total of 340 usable studies. In general, the articles that were excluded were measurement equivalence studies. We elected to exclude measurement equivalence studies due to the fact that these studies did not include information necessary for model comparison in that they rarely had a "main model" for coding. Additionally, measurement equivalence studies are generally less focused on absolute fit index values, and more focused on parameter equivalence between groups. Since the focus of the present study was on statistical power of overall model fit, we elected to discard measurement equivalence studies. Within the valid articles, 1,692 individual SEMs were included in the present study.

Article Coding

Once we had identified a set of usable studies, we coded each article for relevant variables. First, we coded features of the articles such as the year, authorship, and journal where the article appeared.

Second, each SEM within each article was coded for its reported degrees of freedom (df) and sample size (n). Using information on df and n, we calculated the power coefficient for the tests of exact, close, and not-close fit, using software available from Preacher and Coffman (2006). In addition, we calculated the sample size that would be required to obtain a power coefficient of .80 and recorded the difference between this sample size and the actual, reported sample size.

Third, we recorded aspects of each SEM. The first coding task was to evaluate whether the model was a measurement only, structural only (i.e., with no estimated measurement-related parameters), or combination of measurement and structural model. The second coding task focused on whether the model in question was a model that evaluated phenomena about team and/or group functioning by dispersing the original sample size across teams or groups. This variable was included because we hypothesized that the necessary reduction in sample size to study variables at a team-level as opposed to individual level would decrease n and thus decrease power, all else being equal.

Fourth, we evaluated each reported SEM within an article to arrive at the "main" model(s) of the article. We considered the main models of the article to be models that were either most justified by theory presented earlier in the article and the focus of the study's hypotheses or the model that the study's authors deemed the "final" model either explicitly or indirectly in the language included in their discussion and results sections. The final model was always deemed to be the model that the authors declared their "final" model, even when the

deemed “final” model differed from the model hypothesized in the introduction section of the article in question.

Finally, we included information about each of the overall fit indices of the SEMs. We coded for all possible fit indices, including a category of “other”, for fit indices that are not traditionally reported in I/O Psychology, such as the Akaike Information Criterion (AIC).

Each of 1,692 structural equation models was double coded for accuracy. Descriptive statistics are presented in Table 1. The most problematic area with respect to coding was identification of main models, with an interrater agreement statistic of .76. Full interrater agreement statistics are displayed in Table 2. While no individual category possessed unacceptable, or even mediocre degrees of agreement, the difficulties in reliability centered almost exclusively on identification of the main model of the article. Specifically, there were instances in studies where the language used by authors made it unclear what the final model was meant to be. Often, these difficulties arose in studies where it appeared the authors engaged in post-hoc modeling while not explicitly stating they were doing so. In these cases, we carefully examined the introduction sections of the studies in order to ascertain the likelihood of authors truly supporting a particular post-hoc model vs. simply arriving at a post-hoc model through model modification.

A second area of importance with respect to judgment calls in coding came from identification of whether the model in question was a measurement model, structural model, or combination measurement/structural model. While this may seem surprising, we encountered studies that depicted misleading model figures (i.e., graphics) along with misleading text in indicating what model actually went into a particular statistical program. Often, this was indicated by degree of freedom figures that were dramatically misaligned with information

presented in pictures and text. The most common feature of this idea was when authors only presented information in graphics relating to structural models, but in fact simultaneously tested a measurement/structural model, while neglecting to include this information in a footnote or text. In the instances where the degrees of freedom were dramatically misaligned with what would be a structural or measurement model in isolation and otherwise had no strong textual evidence to indicate what type of modeling was actually conducted, we elected to code the model as a combined measurement/structural model.

Power Calculations

This quantitative review followed suggestions from MacCallum et al. (1996) by using values of 0.00 and 0.05 for testing exact fit, 0.05 and 0.08 for testing close fit, and 0.05 and 0.01 for testing not close fit; each value corresponding to ϵ_o and ϵ_a, respectively. However, because the general trend of power associated with the types of hypothesis tests were strongly correlated (r = .99), we elected to report the results of the close fit test, following recommendations from MacCallum et al. (1996).

Exploratory Review

Distribution of Estimated Power. A primary goal of this quantitative review was to examine the *distribution* of estimated power coefficients in I/O Psychology. The distribution of estimated power coefficients can be seen in Table 3. Across all studies and models, approximately 22 % of models had a power coefficient less than .50. Therefore, 22% of all SEMs tested in Organizational Psychology have less than a 50 % chance of correctly rejecting an invalid model again where the null RMSEA of .05 and the alternative RMSEA of .08. Also, the majority of the power coefficients fell in the range above .90. Therefore, although most models

have high levels of power, a non-trivial percentage of models have unacceptably liberal levels of power.

Sample Size Issues. A goal of this quantitative review was to evaluate the degree to which models were judged to have too small a sample size, relative to the number of degrees of freedom of the model, in order to obtain a certain level of power. For this study, we used a power coefficient of .80 as the definition of a properly "powered" study. The distribution of sample size differences can be seen in Figure 2. Interestingly, approximately 27% of models needed at least 100 more participants to reach a power coefficient of .80, while approximately 11% of models needed at least 500 more participants to reach a power coefficient of .80. In general, our results suggest that a nontrivial amount of models have sample sizes that are grossly inadequate to test their theoretical models. The results of the sample size analysis can be seen in Figure 1.

Fit Index Values. Because statistical power is likely to predict the fit of SEMs, we used obtained power coefficients as a predictor of model fit for the RMSEA, CFI, NNFI, and Chi-Square. Because we were concerned with SEMs in published articles, we limited our analysis to values of fit indices that were in the range of values likely to be published, and within the 95 percent confidence interval for models included in this review. Therefore, we examined the relationship between power and fit for values of RMSEA that were between 0.00 and 0.08, values of CFI between 0.90 and 1.00, and values of NFI between 0.90 and 1.00 (additionally, 95 percent of observations fell between these values for each fit index). We also examined Chi-Square via the probability value associated with the Chi-Square test statistic.

Results indicated that for all fit indices, statistical power was associated with worsened fit, as judged by the fit index in question. This finding held for RMSEA $r(784) = .25$, $p < .05$, CFI $r(889) = -.27$, $p < .05$, and NFI $r(278) = -.20$, $p < .05$. The correlation coefficient associated

with the Chi-Square probability value was significantly associated with power, $r(1547) = .29$, $p < .05$, which indicates that as power increases, the likelihood of finding *nonsignificant misfit* decreases. These correlations were all statistically significant at the .05 level. In terms of interpretation, this means that within published studies and fit index values commonly seen in the literature, power is negatively related to fit index quality—liberally powered SEMs are more likely to obtain better fitted models as compared to conservatively powered SEMs, as judged by common fit indices.

Moderators

Differences in Model Power across Model Type. A goal of this quantitative review was to examine power across SEM types, specifically measurement vs. structural models. For measurement and structural models, the average estimated power coefficients for the test of close fit were .84 for measurement and .65 for structural models. The distributions of estimated power are indicated in Tables 4 and 5 for both model types. Across all studies, approximately 17% of measurement and 39% of structural models had a power coefficient less than .50 under the test of close fit. Stated differently, 16% of measurement and 39% of structural models tested have less than a 50% chance of correctly rejecting close fit. Further analysis indicated that the power of structural models was indeed significantly less than the power of measurement models, $t(1028) = 9.36$, $p < .05$, $d = .64$. The difference between measurement and structural models was most driven by the difference in degrees of freedom between measurement models and structural models, as the median value for degrees of freedom was 33 for structural models and 104 for measurement models. Thus, on average, structural models had a 19% *less* chance of correctly rejecting invalid models, compared to measurement models—owing to fewer degrees of freedom observed in studies focusing on evaluation of structural models. As such, structural models close

to RMSEA of .08 in reality may appear—simply owing to chance—to be sufficiently close to a true RMSEA of .05 to accept the model as "good fitting."

Differences in Model Power across Main Models. A goal of this quantitative review was to determine the degree of difference in power between models that were deemed to be the final accepted model by the researcher and models that were deemed not to be the final accepted model of the researcher. This analysis was conducted on independent samples across studies. The average difference in estimated power coefficients between final (M = 0.73, SD = 0.31) and competing models (M = 0.81, SD = 0.28) was statistically significant, $t(1430) = 4.36$, $p < .05$, d = .27. The main driving force between the power differences between main and non-main models was degrees of freedom, with main models being more associated with less complex models and lower degrees of freedom (median = 50) than competing models (median = 96). This means that within the population of SEM, models that were interpreted as the 'correct" model in a given study had an 8% *less* chance of being correctly rejected as invalid, compared to models that were interpreted as incorrect models. Because the power between "main" and "non-main" models differ, it is entirely possible that the reduction in statistical power between the models is a contributing factor in the reason that the "main" model was the accepted model in the final published article.

Difference in Models Teams/Groups vs. Others. An additional goal of this quantitative review was to determine the degree of difference in power between models that aggregated participants to team or group levels using composition models, and models that did not. It was our expectation that aggregation would lower the final sample size, thus lowering power compared to models that did not aggregate. The average difference in estimated power coefficients between team/groups models (M = 0.65, SD = 0.33) and other models (M = 0.80, SD

= 0.29) was statistically significant, $t(1672) = 5.63$, $p < .05$, d = .48. The difference between aggregated and non-aggregated models was driven by the difference in sample size between team/groups models (median = 155) and other models (median = 288). Thus, on average, team/group models that aggregated responses at the team or group level had a 12% *less* chance of correctly rejecting invalid models, compared to models that did not aggregate—owing to smaller sample sizes observed in studies focusing on the team-level of aggregation. Therefore, these types of models tend to be *less falsifiable* than other types of models in organizational psychology, and may warrant special attention in both model construction and experimentation before the statistical testing of the model, as well as evaluation of the model after testing.

Journal Quality. The final hypothesized variable to impact power is journal quality. We specifically hypothesized that journal quality would be positively associated with power such that as quality increased, power increased. Journal quality was indeed statistically significantly related to power, although the effect size was extremely modest $r(1420) = .06$, $p < .05$.

Discussion

The primary purposes of the present study were fourfold. First, we attempted to inform researchers how sampling variability and power can potentially harm our inferences regarding our judgments of SEMs. Second, we attempt to provide guidance as to how power is calculated in a structural equation modeling context. Third, we illuminated the main influencers of power so as to aid researchers in study design efforts. Fourth, we examined certain methodological situations that could signal a need for the researcher to pay special attention to power. We conducted a quantitative review of the literature to help address these issues. The summary of our findings confirms the need to explicitly consider power in structural equation modeling, as

recommended by several methodologists (e.g., Kim, 2005; Kaplan, 1995; MacCallum, Browne, & Cai, 2006; MacCallum et al., 1996, MacCallum & Hong, 1997).

Summary of Findings

We find our results disconcerting in that nearly one-quarter of SEMs have less than a 50% chance of correctly invalidating a bad model. As such, we can conclude that it is likely that the results obtained from at least some of the models included that fall into the less than 50% power category have model fit that is suspect in nature, especially in cases where the model fit in the particular study was adequate (i.e., near . 08) and not excellent (i.e., <.05). Related to this point, we also discovered that a significant number of studies possessed sample sizes that were far removed from what the sample size ought to have been to have a more appropriately powered test of the SEM in question. These findings are particularly disillusioning as these models all appeared in influential journals for Industrial-Organizational Psychology. Therefore, we speculate that invalid models have likely been accepted into top journals, and consequently accepted by researchers in the scientific community. As was previously discussed, neglect of power can slow the advance of scientific progress by leading researchers toward theory that departs from reality as low model fit power reduces the likelihood of rejecting incorrect models. Owing to our findings, these errors are possible as essentially none of the sampled studies conducted a power analysis on their SEM–in spite of the availability of web-based, power analysis tools from Preacher and Coffman (2006).

Finally, the results from this quantitative review pinpoint particular situations where overly liberal statistical power is more likely to occur in research. Specifically, research that involves the evaluation of structural models or models that evaluate team or group level phenomena are more likely to be susceptible with respect to overly liberal statistical power. For

structural models, this occurs because of reduced amounts of degrees of freedom, as compared to measurement models. For team/groups models, this occurs because of a reduced sample size, as compared to most individual-level models. Additionally, we found that models that were seen as the correct model for the study in question had significantly more liberal levels of power than competing models. It is our intention that bringing light to these situations, and their implications for science, will alert researchers and practioners when they need to pay particular attention to statistical power.

Deriving from our discussion of SEM power as well as our quantitative review of current practice in organizational psychology, in the coming sections we outline what we believe are important recommendations for researchers as a whole, and introduce issues and recommendations related to power that will most likely require the combined attention and consideration from researchers, consumers, and editorial gatekeepers to better advance scientific advancement in organizational psychology via improved research methodology.

Recommendations for the User

Before Data Collection. A desired end in the present study is to prompt researchers to conduct a-priori statistical power analysis for SEMs. Given the availability and user-friendly nature of statistical power tools (see Preacher & Coffman, 2006); we believe very little stands in the way of researchers conducting SEM overall fit index power analysis. As we suggest throughout the present work, understanding the level of power of a SEM is instrumental in the interpretation of overall model fit. To that end, we attempt to present an ordered list of recommendations for the common user of SEMs in research.

First, the researcher should construct a theoretical model of interest and study design. Ideally, the model should maximize the tradeoffs amongst explanatory power, parsimony,

potential for true model fit, and analysis of model misfit. These types of tradeoffs can most likely be estimated by reviewing similar styled models in the particular research domain of interest.

Second, the researcher should determine the sample size they are likely to acquire in testing the theoretical model of interest under their current study design. As with all study designs, it is important to make allowances for methodological artifacts that will decrease sample size over the course of a study.

Third, the user should determine the other elements of the hypothesis test. For this quantitative review, we focused our review around what MacCallum and his associates (1996) dubbed the "test of close fit", whereby the null hypothesis was a RMSEA value of .05 and the alternative RMSEA value was a value of .08. However, it is important to note that any values can be used for the null and alternative RMSEA, even values outside the three tests discussed in the MacCallum et al. (1996) paper. Similarly, any value of alpha can be used, theoretically. It might not always be the case that the researcher is interested in testing null and alternative RMSEA values in line with the tests described by MacCallum et al. (1996). Readers interested in this line of thought can consult work on *isopower* by MacCallum, Lee, & Browne (2010).

With all of these elements, the researcher should conduct a power analysis *before* starting the study. Provided the researcher knows the model degrees of freedom, sample size, alpha, and null/alternative hypothesis tests, power can be calculated for any of the aforementioned power tests using tools provided by Preacher and Coffman (2006) or direct R syntax (available from the first author upon request). These tools require no programming knowledge and can be done via graphical interface (i.e. – "point and click" or "copy and paste"). The user will then have an obtained power coefficient. At this point, some recommendations are warranted, depending on the level of the power coefficient.

After Power Analysis. After a-priori power analysis has been conducted, the first recommendation is that the user should immediately revise their research plan if the power coefficient is extremely low. In such situations researchers are particularly susceptible to *accepting invalid models* as a result of the lack of falsifiability of the model and lack of ability to establish the *verisimilitude* (i.e. – likelihood of truth) of the theoretical model in a meaningful way. This is particularly important in a field that is wed to the use of approximate fit indices and rule of thumb interpretations of such fit indices that contain a relative lack of nuance and appreciation for how dependent approximate fit indices are to statistical artifacts that are usually not even explored, let alone reported (Marsh, Hau, & Wen, 2004; Nye & Drasgow, 2010; Williams & O'Boyle, 2010). This recommendation might beg the question of what power coefficients are considered unacceptably low. Since we do not wish to establish a mechanistic "rule of thumb" regarding this topic, we would simply encourage researchers to "think continuously" as phrased by Cortina & Landis (2011) instead of "thinking discretely" with respect to power coefficients. In the case of extremely low powered models, researchers can remedy low power by obtaining a larger sample size or finding some way to gain degrees of freedom (removing added paths or adding variables of interest to the model are two examples), or both. We suspect in most cases it would be more methodologically sound to increase sample size than degrees of freedom for reasons we will explain later. If a researcher is interested in how large a sample size they need to reach a certain level of power, they can refer to the aforementioned tool from Preacher and Coffman (2006) for guidance.

In instances where models have sample sizes large enough to generate high power (i.e. – near 1) in two or more models, cross-validation recommendations are warranted. For example, a model with 30 degrees of freedom and a sample size of 2,000 would generate a power coefficient

of 1 for a single model, but could also generate power coefficients of 1 for an original model and cross-validation model (or even near 1 if two cross-validated models were tested), if the sample size was split into two equal groups. While, a larger sample size is always better from a statistical standpoint all else being equal, there are issues of diminishing returns with respect to sample size and power to detect misfit. Consequently, there are instances where the benefits derived from cross-validation vastly exceed the very minor benefits made to statistical power. In these instances we recommend cross-validation.

However, not all situations will allow for cross-validation by splitting the original sample into multiple groups. One situation that does not allow for cross-validation in such a way is when splitting sample size into two groups would compromise power, reducing power from one high powered model to two modestly powered models. For example, a model with 17 degrees of freedom and a sample size of 500 would have a power coefficient of .80. If that sample size was split into two equal groups, the power coefficient for those groups would reduce to .50, which means that likelihood of rejecting an invalid model would decrease by approximately 38 percent from the original model, or framed another way, going from incorrect acceptance of bad models (by chance) one out of every five times to one out of every two times, assuming the models are bad in the population.

The value of cross-validation depends in part on the falsifiability of the original and cross-validated model. There will be cases where the benefits normally associated with cross-validation do not keep pace with the problems associated with decreasing levels of falsifiability of the tested models. When this is the case, we recommend testing a single model with the collected data and not engaging in any cross-validation efforts with the *original* data (collecting new data to sufficiently cross-validate the model is always welcome however). As is the case

with defining "extremely low power", we do not wish to establish rote, mechanistic rules of thumb for when to cross-validate with the original sample. We would simply encourage researchers to conduct a-priori power analysis on what power would be both before and after cross-validation. To the degree power stays *the same* and is *high*; cross-validation should *occur*. To the degree power *drops* and is *low*; cross-validation should be *avoided*. If this situation occurs and new data cannot be obtained to cross-validate the model, the researcher should report the expected cross-validation index (ECVI; Browne & Cudeck, 1989), which is computed as an index of how well a solution obtained in one sample is likely to fit independent samples.

Another time when cross-validation should be avoided is when doing so would compromise the stability of parameter estimates. Theoretically, there are cases when splitting a sample size into two or more groups could drive down the sample size enough to cause parameter estimates to go from sufficiently stable to unacceptably unstable. In order to obtain stable parameters estimates, the researcher should aim for a ratio of five units for every free parameter (Bentler & Chou, 1987). Our recommendation is similar to the previous, in that we would encourage researchers to conduct a-priori analysis on the parameter stability of the model parameters before and after cross-validation. To the degree stability stays the *same* and is *high*; cross-validation should *occur* and to the degree stability *drops* and is *low*; cross-validation should be *avoided*.

Model Testing. A final over-arching recommendation is to avoid testing models through "two-step" processes (Anderson & Gerbing, 1988; Anderson & Gerbing, 1992) and having *final* evaluations regarding the utility of models determined through separately analyzing measurement and structural components. The main reason is that in at least some cases, separating one single model evaluation into two smaller models (measurement and structural),

has the potential to severely compromise the power to detect model misfit in what would otherwise be a more strongly powered model via reduction in the degrees of freedom of the tested model in much the same way that cross-validation has the potential to compromise power via sample size reduction.

When it comes to ultimate declarations about the utility of a model, we recommend testing the entire model (measurement + structural) in a single step in order to determine overall model fit, as well as individual parameter values. However, one valid criticism of this one-step method proposed by some methodologists (Fornell & Yi, 1992a; Fornell & Yi, 1992b; Hayduk 1996) is that it does not easily lend itself to investigation of model problems and misfit (Bollen, 2000). To that end, we advise researchers to consider the *jigsaw piecewise technique* advocated by Bollen (2000) in combination with the one-step method. Under this technique, researchers fit pieces of the overall model together and then as a whole, ideally evaluating all possible subcomponents of the overall model to assess where and when model misfit experiences radical upward shifts.

In models that combined measurement and structural elements, it is particularly important to test all aspects of the model, as a model's overall fit if judged in a single step, could be disproportionately influenced by the measurement model (Mulaik et al., 1989). In these cases, poor structural fit could be masked by excellent measurement model fit, resulting in an overall model fit statistic that is deemed as acceptable, although the lack of structural fit would still render the model unsound to the researcher. These problems could be addressed from the aforementioned jigsaw piecewise technique approach specified earlier, examination of parameter values, examination of the residual matrix for the overall model, and calculation and examination of fit indices which focus more on path model relationships (such as the RMSEA-P; Williams &

O'Boyle, 2011). A more ambitious approach to this issue could involve the *combinatorics* approach taken by Meehl and Waller (2002), whereby within a path analysis framework, path analytic model parameters are estimated using only a subset of the elements of the sample correlation matrix, and the resulting parameter estimates are then tested by determining how well they account for the other unused, elements of the correlation matrix. This procedure is conducted for the original model as well as for a set of similar alternative models, and the original model is then compared with the alternatives with respect to results of the risky tests. Support for *verisimilitude* of the original path analytic model is enhanced to the degree that it outperforms the alternative models. If a researcher finds their original structural model is outperformed by another model, this can provide clues about the validity of the proposed structural model.

Ultimately, it is important to recognize that there is no single procedure, and certainly no single mechanical ritualistic procedure that will address all possible methodological issues within structural equation modeling at once, and that a variety of procedures are needed to test the utility of an SEM in a rigorous manner.

Limitations

The current study has a number of limitations. First, the approach taken by MacCallum et al. (1996) in conceptualizing type I and type II errors can arguably be seen as backwards, given that these conceptualizations run contrary to traditional understandings of type I and type II errors in Psychology for more traditional statistics such as ANOVA. However, we feel that despite the potential confusion over "reversing" the usual terms, there is still a great deal of utility in the overall approach. In discussing these issues with colleagues, we have found it useful to discuss the issues presented in this paper in terms of "liberal" and "conservative" levels of

power, rather than in the language of type I and type II errors. Moreover, the general issues that we have discussed and illuminated in this study can also be studied analyzed with type I and type II errors that are aligned in more traditional ways (Hancock 2006; Hancock & Freeman, 2001) if desired.

Another potential limitation of this study is that it necessarily relies on the RMSEA fit index, as well as specific cutoffs for what constitute close fitting models. There are two issues to consider here. The first issue revolves around the use of specific cutoffs for close fitting models under RMSEA. The second issue revolves around the use of approximate model fit indices in SEM in evaluating model fit. We will consider both of these points in turn.

Strict cutoff values for model fit in SEM are oversimplifications of complex statistical situations. We do not see much value in the rules of thumb often used to evaluate model fit, as they tend to be dependent on statistical issues that are often ignored in practice (Marsh, Hau, & Wen, 2004; Nye & Drasgow, 2010; Williams & O'Boyle, 2010). Like Marsh et al. (2004), we feel that despite the cautions offered by Hu and Bentler (1999), problems with the rules of thumb surrounding model fit have been frequently ignored in the practice of SEM. In fact, a current citation count of Hu & Bentler's study reports approximately 617 overall citations *per yea*r. For the sake of comparison, Marsh et al. (2004), which signifies the very serious problems associated with strict adherence to rules of thumb, averages 59 citations per year, with a large number of these citations occurring in methodological journals (Harzing, 2010).

The reason we chose to use the rules of thumb outlined in the introduction is because we assumed that these would be the values most salient to the greatest number of researchers, a conjecture which has been vetted by the above evidence. Consequently, we thought that our findings would resonate with researchers more when using rules of thumb to which most

researchers most likely adhered in order to see the interplay between power and focusing solely on passing the .08 RMSEA "goodness" threshold.

Similar reasoning was used with respect to the second issue: the use of approximate model fit indices to evaluate model fit in the first place. A lively discussion on the internet listserv SEMNET has developed regarding the usefulness of evaluating model fit using goodness of fit (GOF) indices. Some researchers argue that the Chi-Square test is the only acceptable fit index to use in testing SEMs (Barrett, 2007; Hayduk et al., 2007). The argument for the sole use of the Chi-Square in evaluating models is centered on the following points: 1) there are no single thresholds for GOF indices that can be applied to any fit index under all possible measurement and data conditions, 2) GOF indices often allow for researchers to avoid careful model specification and examination, 3) GOF indices can allow rather weak models to make it through the peer-review process, 4) the potential for the degree of casual misspecification of models to be uncorrelated with GOF indices values, and 5) Chi-Square does a better job at detecting model misspecification than does any other fit index.

Readers who are interested in examining the issue of approximate versus exact fit indexes at length can examine the above issues and counterarguments in the special 2007 issue of the *Personality and Individual Differences* journal (42nd volume 5th edition) that summarizes this debate. While we recognize the importance of the fit index debate, the RMSEA fit index is still commonly used and evaluated by reviewers as a basis for the adequacy of model-to-data fit. Indeed, virtually no model in our quantitative review was evaluated strictly on the use of Chi-Square test. It is also worth noting, that even in the event that a researcher only wants to use the Chi-Square to evaluate models, the power analysis of MacCallum et al. (1996) can be extended

in such a way, by aligning the null hypothesis value of RMSEA to zero, which represents perfect fit and is analogous to the Chi-Square test (McIntosh, 2007).

Future Research

There are multiple areas for future research as it relates to this study. First, it could be useful for power analysis of SEM to be extended to other areas of psychology. It is possible that the problems presented in this quantitative review are more severe in other areas of psychology that frequently employ SEM. Second, this line of research could potentially extend to comparisons between nested SEMs. Recent work on statistical power as it relates to nested SEM has been proposed by MacCallum and Browne (2006) and Li & Bentler (2011). Third, this line of research could potentially extend to analysis of power about specific parameter estimates. Specifically, using the statistical program Mplus, and work introduced by Muthen & Muthen (2002), one can calculate the power to detect that a specific parameter will be different than zero.

Conclusions

The present study provides strong evidence for the importance of statistical power in SEM in organizational psychology research. Statistical power has been emphasized in experimental and correlational research. We believe the emphasis on statistical power should extend to SEM vis-à-vis overall model fit, with stricter control of model quality in journals via the editors, simple tools that can be used by researchers to actually calculate power coefficients for SEMs, and better quality education with respect to the teaching of power in modeling to peers and students.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103, 411-423. doi:10.1037/0033-2909.103.3.411

Anderson, J. C., & Gerbing, D. W. (1992). Assumptions and comparative strengths of the two-step approach: Comment on Fornell and Yi. *Sociological Methods and Research*, 20, 321-333. doi:10.1177/0049124192020003002

Bagozzi, R.P., & Yi, Y. (1998). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, 16, 74-94. doi:10.1007/BF02723327

Bagozzi, R.P., & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. *Journal of the Academy of Marketing Science*, 40, 8-34. doi:10.1007/s11747-011-0278-x

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences*, 42, 815-824. doi:10.1016/j.paid.2006.09.018

Bentler, P.M. & Bonnett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 588-606. doi:10.1037//0033-2909.88.3.588

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Methods*, 107, 238-246. doi:10.1037//0033-2909.107.2.238

Bentler, P. M., & Chou, C. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, 16, 78-117. doi:10.1177/0049124187016001004

Bollen, K. A. (1989). *Structural equations with latent variables*. Oxford, England: John Wiley and Sons.

Bollen, K. A. (2000) Modeling strategies: In search of the holy grail. *Structural Equation Modeling*, 7, 74-81. doi:10.1207/S15328007SEM0701_03

Browne, M.W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, 24, 445-55. doi:10.1207/s15327906mbr2404_4

Browne, M.W., & Cudeck, R. (1992). *Alternative ways of assessing model fit. Sociological Methods and Research*, 21, 230-258. doi:10.1177/0049124192021002005

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen, K. A., & Long, J. S. (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Chen, F., Curran, P.J., Bollen, K.A., Kirby, J., and Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods and Research*, 36, 462-494. doi:10.1177/0049124108314720

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

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, 112, 155-159. doi:10.1037//0033-2909.112.1.155

Cohen, J. (1994). The earth is round (p > .05). *American Psychologist*, 49, 997-1003. doi:10.1037//0003-066X.49.12.997

Cortina, J.M. & Landis, R.S. (2011). The earth is NOT round (p = .00). *Organizational Research Methods*, 14, 332-349. doi:10.1177/1094428110391542

Cox, D.R., & Reid, N. (1987). Approximations to noncentral distributions. *Canadian Journal of Statistics*, 15, 105-114. doi:10.2307/3315199

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 56-83. doi:10.1080/10705519909540119

Fornell, C., & Yi, Y. (1992a). Assumptions of the two-step approach to latent variable modeling. *Sociological Methods and Research*, 20, 291–320. doi:10.1177/0049124192020003001

Fornell, C., & Yi, Y. (1992b). Assumptions of the two-step approach to latent variable modeling: Reply to Anderson and Gerbing. *Sociological Methods and Research*, 20, 334–339. doi:10.1177/0049124192020003003

Hancock, G. R. (2006). Power analysis in covariance structure models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course*. Greenwood, CT: Information Age Publishing, Inc.

Hancock, G.R., & Freeman, M.J. (2001). Power and sample size for the RMSEA test of not close fit in structural equation modeling. *Educational and Psychological Measurement*, 61, 741-758.

Hayduk, L. (1996). *LISREL: Issues, debates, and strategies*. Baltimore: Johns Hopkins University Press.

Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! testing! One, two, three - Testing the theory in structural equation models! *Personality and Individual Differences*, 42, 841–850. doi:10.1016/j.paid.2006.10.001

Harzing, A.W. (2010). Publish or Perish, version 3.1, available at www.harzing.com/pop.htm.

Hermida, R., Conjar, E.A., Najab, J.A., Kaplan, S.A., & Cortina, J.M. (2010) On the Practice of Allowing Correlated Residuals in Structural Equation Models. Unpublished Manuscript, Department of Psychology, George Mason University, Fairfax, Virginia, United States.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55. doi:10.1080/10705519909540118

Kaplan, D. (1995). Statistical power in structural equation modeling. In R.H. Hoyle (ed.), *Structural Equation Modeling: Concepts, Issues, and Applications* (pp. 100-117). Newbury Park, CA: Sage Publications, Inc.

Kim, K. H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling*, 12, 368-390. doi:10.1207/s15328007sem1203_2

Landis, R., Edwards, B. D., & Cortina, J. (2009). Correlated residuals among items in the estimation of measurement models. In C. E. Lance & R. J. Vandenberg (Eds.). *Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences* (pp. 195-214). New York: Routledge.

Le, H., Schmidt, F., Harter, J.K., & Lauver, K.J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Research Methods*, 112, 112-125. doi:10.1016/j.obhdp.2010.02.003

Li, L., & Bentler, P.M. (2011). Quantified choice of root-mean-square errors of approximation for evaluation and power analysis of small differences between structural equation models. *Psychological Methods*, 16, 116-126. doi:10.1037/a0022657

MacCallum, R.C., & Austin, J.T. (2000). Applications of Structural Equation Modeling in Psychological Research. *Annual Review of Psychology*, 51, 201-226. doi:10.1146/annurev.psych.51.1.201

MacCallum, R.C., Browne, M.W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, 11, 19-35. doi:10.1037/1082-989X.11.1.19

MacCallum, R.C., Browne, M.W., & Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130-149. doi:10.1037//1082-989X.1.2.130

MacCallum, R.C., & Hong, S. (1997). Power analysis in covariance structure modeling using GFI and AGFI. *Multivariate Behavioral Research*, 3, 193-210. doi:10.1207/s15327906mbr3202_5

MacCallum, R., Lee, T., & Browne, M.W. (2010). The issue of isopower in power analysis for tests of structural equation models. *Structural Equation Modeling*, 17, 23-41. doi: 10.1080/10705510903438906

Marsh, H. W., Hau, K. T., & Wen, Z. L. (2004) In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu & Bentler (1999) findings. *Structural Equation Modeling*, 11, 320-341. doi:10.1207/s15328007sem1103_2

McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences*, 42, 859–867. doi:10.1016/j.paid.2006.09.020

McQuitty, S. (2004). Statistical power and structural equation models in business research. *Journal of Business Research*, 57, 175-183. doi:10.1016/S0148-2963(01)00301-0

Meehl, P.E., & Waller, N.G. (2002). The path analysis controversy: A new statistical approach to Strong Appraisal of Verisimilitude. *Psychological Methods*, 7, 283-300. doi:10.1037//1082-989X.7.3.283

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, 105, 430-445.doi:10.1037/0033-2909.105.3.430.

Muthen, L.K., & Muthen, B.O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 9, 599-620. doi:10.1207/S15328007SEM0904_8

Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods*, 14, 548-570. doi:10.1177/1094428110368562

Patnaik, P.B. (1949). The noncentral χ^2 and F-distributions and their applications. *Biometrika*, 36, 202-232. doi:10.2307/2332542

Preacher, K. J., & Coffman, D. L. (2006). Computing power and minimum sample size for RMSEA [Computer software]. Available from http://quantpsy.org/.

Rigdon, E.E. (1994). Calculating degrees of freedom for a structural equation model. *Structural Equation Modeling*, 1, 274-278. doi:10.1080/10705519409539979

Specht, D. A. (1975). On the evaluation of causal models. *Social Science Research*, 4, 113-133. doi:10.1016/0049-089X(75)90007-1

Steiger, J.H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25, 173-180. doi:10.1207/s15327906mbr2502_4

Sugawara, H. M. & MacCallum, R. C. (1993). Effect of estimation method on incremental fit indexes for covariance structure models. *Applied Psychological Measurement 17,* 365-77. doi: 10.1177/014662169301700405

Steiger J. H., & Lind J. M. (1980). Statistically based tests for the number of factors. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Swanson, R.A., & Holton, E.F. III. (2005). *Research in Organizations: Foundations and Methods of Inquiry*. Berrett-Koehler: San Francisco, CA.

Williams, L. J. & Holahan, P. J. (1994). Parsimony-based fit indices for multiple-indicator models: do they work? *Structural Equation Modeling*, 1, 161–189. doi:10.1080/10705519409539970

Williams, L.J., & O'Boyle Jr., E.H. (2011). The myth of global fit indices and alternatives for assessing latent relations. *Organizational Research Methods*, 14, 350-369 doi:10.1177/1094428110391472

FOOTNOTES

1. The non-central χ^2 distribution was chosen by MacCallum et al. (1996) as it does not require that the model being estimated fit the data *exactly* (i.e., the model-to-data discrepancy or fit function is exactly 0; e.g., Brown & Cudeck, 1993; Chen, Curran, Bollen, Kirby, & Paxton, 2008; Steiger & Lind, 1980) in the population, as does the central χ^2 distribution. As such, the non-central χ^2 allows for misspecified, yet practically or scientifically useful models to still fit the data adequately and is therefore a more tenable distribution under most data analytic situations (e.g., MacCallum et al., 1996; Saris & Satorra, 1993).

Table 1.

Descriptive Statistics for Coded Variables

Variable	Mean	Standard Deviation
Eigenfactor Score	0.0125	0.0092
Degrees of Freedom	71.51	111.28
Sample Size	421.64	448.52
Model Power	0.79	0.29
Chi-Square	841.72	1927.14
RMSEA	.08	.06
CFI	.90	.10
NFI	.90	.10
Sample Size Required	428.77	124.57
Sample Size Difference	-12.95	167.86
Team/Groups Model vs. Other Model	1.92	2.78
Measurement vs. Structural Model	1.27	0.44
Main Model vs. Competing Model	1.66	0.94

Note. Codes were as follows: Teams/Groups models = 1, non-team/groups models = 2, structural model = 1, measurement model = 2, main model = 1, competing model = 2.

Table 2.

Interrater Agreement Statistics

Quantitative Review Category	Type of Agreement	Agreement Value
Sample Size	ICC	0.98
Degrees of Freedom	ICC	0.98
Statistical Power Coefficient	ICC	0.98
Team Aggregation (yes/no)	Cohen's Kappa	0.98
Model Type (measurement, structural, combined)	Cohen's Kappa	0.85
Main Model (yes/no)	Cohen's Kappa	0.76

Note. ICC = Intraclass Correlation Coefficent

Table 3.

Distribution of Power, Test of Close Fit

Power	N	Proportion	Cumulative Proportion
0.00-0.09	21	0.01	0.01
0.10 – 0.19	88	0.05	0.06
0.20 – 0.29	67	0.04	0.10
0.30 – 0.39	124	0.07	0.17
0.40 – 0.49	75	0.05	0.22
0.50 – 0.59	85	0.05	0.27
0.60 – 0.69	47	0.03	0.30
0.70 – 0.79	66	0.04	0.34
0.80 – 0.89	104	0.06	0.40
0.90 – 1.00	1015	0.60	1.00

Note. $N = 1692$

Table 4.

Distribution of Power, Measurement Models, Close Fit

Power	N	Proportion	Cumulative Proportion
0.00 – 0.09	00	0.00	0.00
0.10 – 0.19	33	0.04	0.04
0.20 – 0.29	30	0.04	0.08
0.30 – 0.39	33	0.04	0.12
0.40 – 0.49	30	0.04	0.16
0.50 – 0.59	31	0.04	0.20
0.60 – 0.69	16	0.02	0.22
0.70 – 0.79	30	0.04	0.26
0.80 – 0.89	42	0.06	0.32
0.90 – 0.99	511	0.68	1.00

Note. $N = 756$; models that were combined measurement and structural models are not included but are available upon request from the first author.

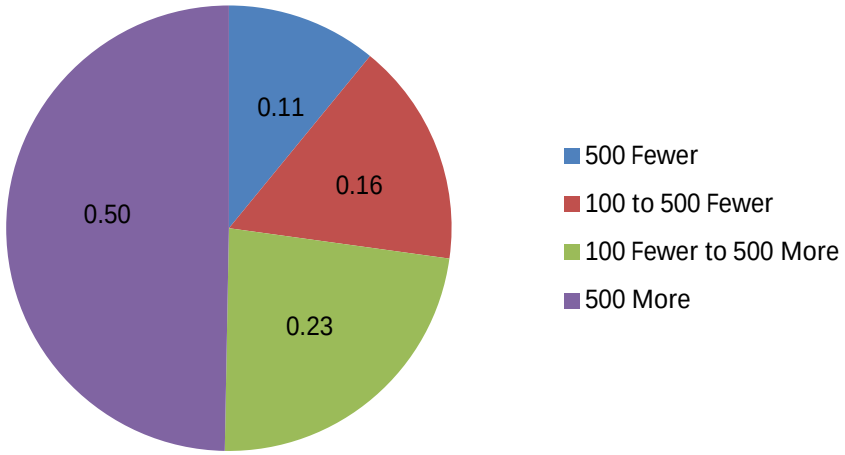
Table 5.

Distribution of Power, Structural Models, Close Fit

Power	N	Proportion	Cumulative Proportion
0.00 – 0.09	05	0.02	0.02
0.10 – 0.19	32	0.12	0.14
0.20 – 0.29	18	0.06	0.20
0.30 – 0.39	29	0.11	0.31
0.40 – 0.49	24	0.09	0.40
0.50 – 0.59	16	0.06	0.46
0.60 – 0.69	17	0.06	0.52
0.70 – 0.79	09	0.03	0.55
0.80 – 0.89	09	0.03	0.58
0.90 – 0.99	115	0.42	1.00

Note. $N = 274$; Models that were combined measurement and structural models are not included but are available upon request from the first author.

Figure 1



Note. $N = 1692$. Labels refer to the sample size required to obtain power coefficient of .80 subtracted from the sample size used to test the model.