Statistical and Methodological Myths and Urban Legends

Doctrine, Verity and Fable in the Organizational and Social Sciences

Edited by

Charles E. Lance and Robert J. Vandenberg

Routledge
Taylor & Francis Group
New York London

Routledge
Taylor & Francis Group
270 Madison Avenue
New York, NY 10016

Routledge
Taylor & Francis Group
27 Church Road
Hove, East Sussex BN3 2FA

Routledge is an imprint of Taylor & Francis Group, an Informa business

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-0-8058-6238-6 (Softcover) 978-0-8058-6237-9 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Statistical and methodological myths and urban legends : doctrine, verity and fable in the organizational and social sciences / [edited by] Charles E. Lance & Robert J. Vandenberg.
p. cm.
Includes bibliographical references.
ISBN 978-0-8058-6237-9 (hardcover) -- ISBN 978-0-8058-6238-6 (pbk.)
1. Organization--Research--Methodology. 2. Organization--Research--Statistical methods. 3. Social sciences--Statistical methods. 4. Social sciences--Research--Statistical methods. I. Lance, Charles E., 1954- II. Vandenberg, Robert J.

HD30.4.S727 2009
300.72--dc22 2008019657

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the Routledge Web site at
http://www.routledge.com

11

Sample Size Rules of Thumb

Evaluating Three Common Practices

Herman Aguinis and
Erika E. Harden

This chapter provides a description and critical analysis of three rules of thumb related to sample size that are commonly used by researchers in the organizational and social sciences. Thus, similar to the chapter by Vandenberg and Grelle (2008), our chapter does not address faulty assumptions or improper citations that can be traced back to an original source and have risen to the category of "statistical and methodological myths and urban legends." Instead, we provide a critical analysis of these rules of thumb that we hope will provide information that will be useful to researchers in their own work as well as journal reviewers who evaluate the work of others. We also hope that by discussing these rule of thumbs critically we will prevent them from possibly becoming statistical and methodological myths and urban legends in the future.

Our chapter is about inferences regarding estimated relationships between variables and latent constructs or between observed indicators and latent constructs. Thus, our chapter addresses rules of thumb about sample size related to internal, construct, and statistical conclusion validity but does not address issues of external validity (i.e., what sample size is needed to be able to generalize results across populations).

We wanted to minimize the impact of our subjective opinion on the process of identifying any existing rules of thumb. So, rather than discussing what *we think* are some of the existing rules of thumb that researchers use, we adopted an inductive approach for identifying any existing rules. Specifically, we conducted an in-depth review of

the Method, Results, and Discussion sections for each of approximately 1,260 articles published between 2000 and 2006 in the following journals:

- *Academy of Management Journal*
- *Administrative Science Quarterly*
- *Journal of Applied Psychology*
- *Personnel Psychology*
- *Strategic Management Journal*

We selected the above journals because they arguably publish some of the most methodologically sophisticated and rigorous empirical research in the field of management. If rules of thumb that may not be appropriate, or are used inappropriately, are invoked frequently by researchers publishing in these journals, it is likely that these rules are used by researchers publishing in many other journals as well.

Our inductive study consisted of searching for statements and justifications that authors used that involved sample size. We found 102 articles (i.e., about 8.2% of all articles included in our literature review) that included a statement in which authors explained how they chose the sample size they had, described consequences of their particular sample size, or explained or justified a result in relationship to their sample size. We identified the following commonly invoked rules of thumb related to sample size:

1. Determine whether sample size is appropriate by conducting a power analysis using Cohen's definitions of small, medium, and large effect size.
2. Increase the a priori Type I error rate to .10 because of a small sample size.
3. Sample size should include at least 5 observations per estimated parameter in covariance structure analyses.

Next, we critically analyzed each of these three rules of thumb by answering the following questions: Where did these rules come from? What did the attributed sources really say about them? How much merit do these rules really have? Should we continue using these rules of thumb or should we abandon them altogether?

Determine Whether Sample Size Is Appropriate by Conducting a Power Analysis Using Cohen's Definitions of Small, Medium, and Large Effect Size

A crucial step in designing a study is determining sample size because N is one of the key determinants of statistical power. Statistical power is the probability of detecting an effect that exists in the population. The greater the sample size, the greater the statistical power. Statistical power is $1 - \beta$, where β is the Type II error rate (i.e., the probability of not detecting an existing effect). In addition to sample size, power is affected by the size of the effect in the population (i.e., the greater the effect, the greater the power), and by the Type I error rate (i.e., α), which is the probability of falsely concluding that an effect exists. Note that Type I and Type II error have an inverse relationship. In order to conduct a power analysis to determine what sample size is sufficient to detect an effect, or whether the sample size in hand is sufficient to detect an effect, there is a need to choose a targeted effect size (Aguinis, Boik, & Pierce, 2001).

Our review uncovered that a common rule of thumb in conducting a power analysis is to use Cohen's (1988) definitions of small, medium, and large effect size. For instance, Raver and Gelfand (2005) conducted a power analysis using Cohen's values and concluded that "[a] power analysis indicated that the power to detect a medium effect with an alpha level of .05 was 46 percent, and the power to detect a large effect was 86 percent (Cohen, 1988)" (p. 394). Similarly, Morgeson and Campion (2002) also used Cohen's definitions and noted that "[s]tatistical power to detect a significant R^2 in the regression analysis was 35% for a small effect ($R^2 = .0196$, $p < .05$) and 99% for a medium effect ($R^2 = .13$, $p < .05$; Cohen, 1988)" (p. 598). A perusal of articles published recently in some of the major journals in the organizational and social sciences reveals many additional examples. Consider Kim, Hoskisson, and Wan (2004), who noted that "the precise estimates of effect sizes are generally difficult to obtain, which is a major obstacle to implementing power analysis. Following Lane and colleagues (1998), we rely on general approximations of small, medium, and large effect size as suggested by Cohen (1992)" (p. 625). Likewise, Brews and Tucci (2004) argued that "[o]ur large sample size alleviates concerns about statistical power (Schwenk & Dalton, 1991; Ferguson & Ketchen, 1999). We have adequate power to detect small, medium, and large effects" (p. 437). Finally, Brown (2001) also used

Cohen's definitions in his power analysis and stated that "this study is limited by a relatively small sample size and modest reliabilities of some measures. Although the power to detect moderate effects (r = .30) at the .05 alpha level with this sample is .78, the power to detect small effects (r = .10) is only .14 (Cohen & Cohen, 1983)" (p. 292).

What did Cohen really recommend about the procedures to select a targeted effect size in conducting a power analysis to assess whether one's sample size is sufficiently large? Did he recommend that researchers use specific values for small, medium, and large effects? Did these values remain consistent over time? How did he come up with these values? Let's consider the cited sources.

Cohen (1992) noted that "researchers find specifying ES [effect size] the most difficult part of power analysis" (p. 156). To address this issue, Cohen, Cohen, West, and Aiken (2003; based largely on Cohen & Cohen, 1983, pp. 59–60) outlined the following three strategies for identifying an appropriate effect size in power analysis:

1. To the extent that studies that have been carried out by the current investigator or others are closely similar to the present investigation, the ESs found in these studies reflect the magnitude that can be expected.
2. In some research areas an investigator may posit some minimum population effect that would have either practical or theoretical significance.
3. A third strategy is deciding what ES values to use in determining the power of a study is to use certain suggested conventional definitions of *small*, *medium*, and *large* effects...This option should be looked upon as the default option only if the earlier noted strategies are not feasible. (p. 52).

Consider the history behind the conventional definitions of small, medium, and large effect, which should be used *only if the other strategies are not feasible.* As described by Aguinis, Beaty, Boik, and Pierce (2005), Cohen's first published description of specific magnitudes for effects appeared in his 1962 *Journal of Abnormal and Social Psychology* article. In this article, Cohen reported results of a review and content analysis of articles published in the 1960 volume of this same journal. In the Method section of his article, when describing the effect sizes he used for his power analysis, Cohen (1962, p. 147) noted that "the level of average population proportion at which the power of the test was computed was the average of the sample proportions

found" and "the sample values were used to approximate the level of population correlation of the test." For the correlation coefficient, Cohen defined .40 as medium because this seems to have been close to the average observed value he found in his review. Then, he chose the value of .20 as small and .60 as large. In other words, Cohen's definitions of small, medium, and large effect sizes are based in part on observed values as reported in the articles published in the 1960 volume of *Journal of Abnormal and Social Psychology*, and in part on his own subjective opinion. A few years later, Cohen (1988) decided to lower these values to .10 (small), .30 (medium), and .50 (large) because the originally defined values seemed a bit too high. Given the history behind the conventional values for small, medium, and large effects, it is not surprising that Cohen (1992) himself acknowledged that these definitions "were made subjectively" (p. 156).

In sum, numerous researchers conduct a power analysis to determine whether a study's sample size is sufficiently large to detect an effect using Cohen's conventional definitions of effect sizes. A critical analysis of this practice in light of the sources invoked to support its use leads to the following conclusions. First, Cohen mentioned that using his admittedly conventional values is only one of three procedures for identifying a targeted effect size to be used in a power analysis to assess whether a study's sample size is sufficiently large (Cohen & Cohen, 1983, pp. 59–60). In fact, this strategy should be used only as a last resort and only if the other two preferred strategies are not feasible. However, many researchers seem to focus on this procedure to the exclusion of the other two. Second, Cohen himself noted that his values for small, medium, and large effects are subjective. In fact, he changed the values for small, medium, and large effects over time with no apparent reason but his subjective opinion that these values should be modified downward.

Discussion

Statistical power is the probability of detecting an effect that indeed exists in the population. Sample size is one of the key factors that affect statistical power. If statistical power is not sufficient, one risks the possibility of erroneously concluding that there is no effect in the population. Thus, when an effect is not found, journal reviewers usually request that a power analysis be conducted to assess whether

a study's sample size was sufficiently large. At that point, a researcher must make a decision about what targeted effect size to use because choosing a large effect may lead to the conclusion that a particular *N* was sufficiently large, but this same *N* may not be sufficiently large to detect a smaller effect. In short, a particular sample size may be seen as adequate or not depending on the targeted effect size used in the power analysis.

Although Cohen suggested three strategies for identifying the effect size to be used in a power analysis, most researchers use the effects that Cohen labeled small, medium, and large. Per Cohen's own admission, these values are largely subjective. As our review indicates, they were initially derived from a very narrow literature review of articles published in the 1960 volume of the *Journal of Abnormal and Social Psychology*. However, using these values is a pervasive practice, perhaps because it is more convenient to do so as compared to using the other two preferred strategies for identifying targeted effect sizes (i.e., an effect size derived from previous literature or an effect size that is scientifically or practically significant).

The two preferred strategies for identifying a targeted effect size used in a power analysis point to the need to take into account the specific research context and domain in question and to not rely on broad-based conventions. For example, Cohen (1988) wrote that, for the f^2 effect size, .02 is a "small effect." However, Aguinis et al. (2005) conducted a 30-year review of all articles in *Academy of Management Journal*, *Journal of Applied Psychology*, and *Personnel Psychology* that used moderated regression to test hypotheses about categorical moderator variables and found that the median effect size is $f^2 = .002$ (i.e., 10 times smaller than what Cohen labeled as a small effect). Cohen (1988) himself recommended that context be taken into account in choosing a targeted effect size in a power analysis when he wrote that effect sizes are relative not only to each other but also "to the area of behavioral science or even more particularly to the specific content and research methods being employed in any given investigation" (p. 25).

Finally, also related to the importance of placing a particular effect within its context, it is generally not appropriate to equate Cohen's "small" (which requires a large *N* to be detected) effect with "unimportant effect" and Cohen's "large" effect (which requires a smaller *N* to be detected) with "important effect." In some contexts, what seems to be a small effect can actually have important

consequences. For example, Martell, Lane, and Emrich (1996) found that an effect size of 1% regarding male-female differences in performance appraisal scores led to only 35% of the highest-level positions being filled by women. Accordingly, Martell et al. (1996) concluded that "relatively small sex bias effects in performance ratings led to substantially lower promotion rates for women, resulting in proportionately fewer women than men at the top levels of the organization" (p. 158). Aguinis (2004) and Aguinis et al. (2005) described several additional illustrations of how, in some contexts, effects that are labeled as "small" based on Cohen's definitions actually have very significant consequences for both theory and practice.

Summary: The rule of thumb: Researchers determine the appropriateness of a particular sample size by conducting a power analysis using Cohen's definitions of small, medium, and large effect size. The kernel of truth: The use of Cohen's small, medium, and large effect size is only one of three methods that he recommended, and the least preferred of the three, to determine sample size via a power analysis. The inappropriate application of the rule of thumb: The definitions of small, medium, and large effect size are believed to have been determined objectively and can be used regardless of research context and domain. The follow-up: Future research is needed to understand the size of minimally meaningful targeted effect sizes in various research contexts and research domains.

Increase the A Priori Type I Error Rate to .10 Because of Your Small Sample Size

Recall that statistical power is $1 - \beta$, β is the Type II error rate, and β is inversely related to α (i.e., Type I error rate). In the presence of what is seen as a small *N*, many authors decide to increase the a priori α from the usual .01 and .05 values to .10 or even .20 to decrease β and increase statistical power. Our review revealed that this is a fairly common rule of thumb. For example, Brown (2003) noted that "[g]iven that the sample was now relatively small (i.e., 41 teams), an α level of .10 was used for all hypothesis testing following the recommendations of Kervin (1992)" (p. 951). Likewise, Garg, Walters, and Priem (2003) argued that "our sample size is not overly large; it is appropriate to use a less conservative criterion for statistical significance (Sauley & Bedeian, 1989; Skipper, Guenther, & Nass, 1967). We

therefore selected .1, a priori, as the appropriate level of significance for testing our hypotheses" (p. 734). As another illustration, Boland, Singh, Salipante, Aram, Fay, and Kanawattanachai (2001) increased their a priori α to .20 using the justification that their sample was small. Specifically, they stated that "[t]he small sample required that we balance Type I and Type II error rates in statistical testing. At a traditional 95 percent confidence level, the power is only .20 (Cohen, 1977), given an average cell size of 12. Stevens (1996: 172) recommended a more 'lenient' alpha level as a way to improve power. We chose an 80 percent confidence level to ensure at least a power of 0.50. Thus, we set the Type I error rate at 20 percent" (p. 399).

As shown by the above illustrations, the practice of relaxing the a priori α level to .10 or even .20 is a methodological practice often implemented when a study includes a small sample. Increasing the α level increases statistical power and the chances of detecting an existing effect. However, is this practice really justified by the cited sources? In other words, do the cited sources actually suggest increasing alpha to the specific value of .10 or even .20? Why not .15? Or .40, for that matter? Let's consider the evidence.

Sauley and Bedeian (1989) is often invoked as a source in support for the increase of α to .10. In discussing research studies with small samples, Sauley and Bedeian noted that

> when either sample size or anticipated effect size are small, a researcher should typically select a less conservative level of significance (e.g., .10 vs. .05). (p. 340)

However, these authors also noted that

> there is no right or wrong level of significance. Blind adherence to the .05 level of significance as the crucial value for differentiating publishable from unpublishable research cannot be justified. As Skipper et al. (1967) suggest, the selection of a significance level by a researcher should be treated as one more research parameter. Rather than being set at a priori levels of .05, .01, or whatever, the appropriateness of specific level of significance should be based upon considerations such as... sample size, effect size, measurement error, practical consequences of rejecting the null hypothesis, coherence of the underlying theory, degree of experimental control, and robustness. (p. 339)

Kervin (1992), which is another source used in support of the use of an increase α to .10, noted that

> [s]ince sampling error is larger with smaller samples, you may want to be more lenient (larger alpha) with smaller samples, other matters being equal, in order to avoid low research power. (p. 557)

Finally, in another one of the sources cited in support of an increase in the a priori α to .10, Stevens (1996) argued that when one has a small sample,

> it might be prudent to abandon the traditional α levels of .01 or .05 to a more liberal α level to improve power sharply. Of course, one does not get something for nothing. We are taking a greater risk of rejecting falsely, but that increased risk is *more than balanced* by the increase in power. (p. 137)

In sum, the recommendation that we increase our a priori α level to .10 is fairly common in the literature as a means to increase statistical power in the presence of a small sample size. However, a careful examination of this recommendation in light of the sources used to support this practice leads to the following conclusions. First, the practice of increasing the a priori α is reasonable and leads to increased statistical power. Second, however, the practice to increase α to the specific value of .10 or even .20 is subject to the criticism that these values are arbitrary, much like the values of .05 and .01 are also arbitrary. Moreover, without taking into account the research context (e.g., negative consequences of incorrectly concluding there is an effect as a consequence of a Type I error), the practice of increasing the α level to an arbitrarily selected greater value may be equally as, or even more, detrimental to theory development and practice than having a small sample size, insufficient statistical power, and erroneously concluding that there is no effect.

Discussion

In the organizational and social sciences, researchers usually adopt the conventional .05 and .01 values for the a priori α (i.e., probability of erroneously concluding that there is an effect). As noted above, many authors choose to increase α to .10 or .20. However, this choice is seldom justified and no discussion is usually provided regarding the trade-offs involved (i.e., increase in the probability of committing a Type I error). If one wishes to increase power by increasing α, one should make an informed decision about the specific

trade-off between Type I and Type II errors rather than choosing an arbitrarily larger value for α.

Murphy and Myors (1998) suggested a useful way to weigh the pros and cons of increasing the Type I error rate for a specific research situation. The appropriate balance between Type I and Type II error rates can be achieved by using a preset Type I error rate that takes into account the Desired Relative Seriousness (DRS) of making a Type I versus a Type II error. Because Type II error = 1 – power, this strategy is also useful for choosing an appropriate Type I error in relation to statistical power.

Instead of increasing α to an arbitrary value, researchers can make a more informed decision regarding the specific value to give to α. Consider the following situation described by Aguinis (2004, pp. 86–87). A researcher is interested in testing the hypothesis that the effectiveness of a training program for unemployed individuals varies by region such that the training program is more effective in regions where the unemployment rate is higher than 6%. Assume this researcher decides that the probability of making a Type II error (i.e., β, incorrectly concluding that unemployment rate in a region is not a moderator) should not be greater than .15. The researcher also decides that the seriousness of making a Type I error (i.e., incorrectly concluding that percentage of unemployment in a region is a moderator) is twice as serious as making a Type II error (i.e., DRS = 2). Assume the researcher makes the decision that DRS = 2 because a Type I error means that different versions of the training program would be needlessly developed for various regions and this would represent a waste of the limited resources available. The desired preset Type I error can be computed as follows (Murphy & Myors, 1998):

$$\alpha_{\text{desired}} = \left[\frac{p(H_1)\beta}{1-p(H_1)}\right]\left(\frac{1}{\text{DRS}}\right) \tag{11.1}$$

where $p(H_1)$ is the estimated probability that the alternative hypothesis is true (i.e., there is a moderating effect), β is the Type II error rate, and DRS is a judgment of the seriousness of a Type I error vis-à-vis the seriousness of a Type II error.

For this example, assume that based on a strong theory-based rationale and previous experience with similar training programs,

the researcher estimates that the probability that the moderator hypothesis is correct is $p(H_1) = .6$. Solving Equation 11.1 yields

$$\alpha_{desired} = \left[\frac{(.6)(.15)}{1-.6}\right]\left(\frac{1}{2}\right) = .11.$$

Thus, in this particular example, using a nominal Type I error rate of .11 would yield the desired level of balance between Type I and Type II statistical errors.

Implementing this procedure for choosing the specific a priori Type I error rate provides a more informed and better justification than using any arbitrary value such as .10 or .20 without carefully considering the trade-offs and consequences of this choice. Also, implementing this more informed strategy for selecting an a priori α is less likely to raise concerns among journal editors and reviewers as compared to selecting any arbitrary value.

Summary: The rule of thumb: When faced with a small sample, researchers increase the a priori Type I error rate to .10 or even .20 as a means to increase statistical power. The kernel of truth: Increasing Type I error will increase statistical power (i.e., probability of detecting existing effects). The inappropriate application of the rule of thumb: Increasing Type I error rate to .10, .20, or any other arbitrarily selected value is assumed to be beneficial regardless of research context and research domain. The follow-up: Future research is needed to understand the trade-offs involved in making Type I in relation to Type II errors in various research contexts and research domains.

Sample Size Should Include at Least 5 Observations per Estimated Parameter in Covariance Structure Analyses

It seems to be common knowledge that a factor analysis should include 5 observations per estimated parameter. This 5:1 ratio seems to be a common recommendations and is followed not only in the context of factor analysis but also in testing the fit of a measurement model before testing a substantive structural model in structural equation modeling, path analysis, and other types of analyses based on covariance structures (e.g., Pierce, Aguinis, & Adams, 2000; Pierce, Broberg, McClure, & Aguinis, 2004). The 5:1 ratio rule is also

used by authors in referring to structural models, not just measurement models.

Bentler's work is a source often cited in support of the 5:1 ratio rule of thumb. For example, Kinicki, Prussia, Wu, and McKee-Ryan (2004) stated that "Bentler (1990) recommends a minimum of five cases for each estimated parameter in structural models" (p. 1061). Likewise, Epitropaki and Martin (2004) cautioned that their results should be interpreted with caution because "the minimum 5:1 cases per parameter (Bentler, 1995) is still not met in those six groups" (p. 304). Additionally, Takeuchi, Yun, and Tesluk (2002) cited Bentler and Chou (1987) when stating that "[i]t is recommended that in SEM, the ratio of respondents to parameters estimated should be at least 5:1" (p. 660). Finally, as an additional illustration, Sturman and Short (2000) noted that "although strict guidelines for minimum sample sizes do not exist (Anderson & Gerbing, 1988), our sample of 416 exceeds the minimum of 200 recommended by Boomsma (1982), and our sample size to parameter ratios of at least 8:1 exceed the suggested minimum of 5:1 for reliable maximum likelihood estimation (Bentler, 1985)" (p. 685).

What is the origin of the 5:1 ratio rule? Did Bentler (1985) really say that we need 5 observations per parameter estimated in a covariance structure analysis to obtain trustworthy estimates? Let's consider the evidence.

In a frequently cited source used to invoke this rule of thumb, Bentler (1985) noted the following:

> An over-simplified guideline regarding the trustworthiness of solutions and parameter estimates might be the following. The ratio of sample size to number of free parameters to be estimated may be able to go as low as 5:1 under normal elliptical theory. Although there is little experience on which to base a recommendation, a ratio of at least 10:1 may be more appropriate for arbitrary distributions. (p. 3)

> These ratios need to be larger to obtain trustworthy *z*-tests on the significance of parameters, and still larger to yield correct model evaluation chi-square probabilities. (p. 3)

Two years later, Bentler and Chou (1987, p. 90) identified "large" sample size as one of the statistical requirements of structural equation modeling because "the statistical theory is based on 'asymptotic' theory, that is, the theory that describes the behavior of statistics as the sample size becomes arbitrarily large (goes to infinity). In practice, samples

can be small to moderate in size, and the question arises whether large sample statistical theory is appropriate in such situations."

Bentler and Chou provided a virtually verbatim "oversimplified guideline" from Bentler (1985) to serve as a rule of thumb regarding the ratio of number of observations per parameters estimated in a model:

> The ratio of sample size to number of free parameters may be able to go as low as 5:1 under normal and elliptical theory, especially when there are many indicators of latent variables and the associated factor loadings are large. Although there is even less experience on which to base a recommendation, a ratio of at least 10:1 may be more appropriate for arbitrary distributions. These ratios need to be larger to obtain trustworthy *z*-tests on the significance of parameters, and still larger to yield correct model evaluation chi-square probabilities. (p. 91)

In sum, having an appropriate number of observations per estimated parameter in a factor analysis, as in any covariance structure analyses, is obviously an important issue. Not having a sufficient number of observations will lead to unstable and untrustworthy parameter estimates. However, a closer examination of the 5:1 ratio as described in the cited sources leads to the following conclusions. First, this is a lower-bound value and an oversimplified rule of thumb and not necessarily a desirable value. Invoking the 5:1 rule of thumb in support of the conclusion that a particular sample size is ideal is misleading. Second, this ratio applies to situations in which multivariate normality has been observed, which is an unusual situation in the organizational and social sciences. In fact, when multivariate normality is not present, a ratio of at least 10 observations per estimated parameter is recommended for obtaining trustworthy estimates of parameters. Moreover, an even larger number of observations is required to obtain trustworthy estimates of the statistical significance of parameters.

Discussion

Researchers seem to focus on what is an "oversimplified" guideline of 5 observations per parameter. Moreover, this guideline applies to situations in which the data follow a multivariate normal distribution only, which is not typical in the organizational and social sciences. This oversimplified guideline of 5 observations per estimated parameter should be seen as a lower-bound value and not

necessarily a desirable value, particularly when the multivariate normality assumption is violated. Invoking the 5:1 rule of thumb to claim that a particular study has the ideal sample size and follows best practices is misleading.

Summary: The rule of thumb: Sample size should be such that there are at least 5 observations per estimated parameter in a factor analysis and other covariance structure analyses. The kernel of truth: This oversimplified guideline seems appropriate in the presence of multivariate normality. The inappropriate application of the rule of thumb: The 5:1 ratio is believed to be an ideal and best-practice research scenario. The follow-up: Future research is needed to understand the appropriateness of the 5:1 ratio in the presence of multivariate normality and for various degrees of model complexity.

Discussion

In this chapter, we have discussed three rules of thumb related to sample size that, based on a review of articles published from 2000 to 2006 in some of the most prestigious journals in management, are invoked quite commonly. Table 11.1 summarizes each rule of thumb, the kernel of truth, the inappropriate application of each rule of thumb, and the research needed regarding each of these rules of thumb.

Why are these rules of thumb used? We can only speculate on the reasons, but we suspect that some authors may invoke these rules of thumb as a preemptive strike to counter a potential criticism from a reviewer when results do not turn out as predicted (e.g., there is lack of support for a hypothesized effect). Others may invoke these rules as a response to a criticism from a reviewer (i.e., "your sample size is not sufficient for a covariance structure analysis," "your small sample size led to insufficient statistical power to detect population effects") or even at the direction of a reviewer or a journal editor (i.e., "given your small sample size, you must conduct a power analysis using Cohen's definitions of effect size"). Regardless of the reason for invoking these rules, we emphasize that our focus is on a critical analysis of these rules and not on specific authors who have used them. It is not our intention to point fingers and blame specific authors. In fact, we are ourselves guilty of using some of the rules of thumb we critically analyzed in this chapter (e.g., Aguinis & Stone-Romero, 1997, used Cohen's definitions of small, medium, and large effect sizes).

TABLE 11.1 Critical Analysis Summary for the Three Rules of Thumb Related to Sample Size

The rule of thumb	We should determine the appropriateness of N by conducting a power analysis using Cohen's definitions of small, medium, and large effect size.	When faced with a small N, we should increase the a priori Type I error rate to .10 or even .20 as a means to increase statistical power.	N should include at least 5 observations per estimated parameter in a factor analysis and other covariance structure analyses.
The kernel of truth	The use of Cohen's small, medium, and large effect sizes is only one of three methods (but the least preferred) that can be used to determine the appropriateness of N via a power analysis.	The increase of Type I error will increase statistical power (i.e., probability of detecting existing effects).	This oversimplified guideline seems appropriate in the presence of multivariate normality.
The inappropriate application of the rule of thumb	The definitions of small, medium, and large effect size are believed to have been determined objectively and can be used regardless of research context and domain.	The increase of Type I error rate to .10, .20, or any other arbitrarily selected value is assumed to be beneficial regardless of research context and research domain.	The 5:1 ratio is assumed to be an ideal and best-practice research scenario.
Research needed	What is the size of minimally meaningful targeted effect sizes in various research contexts and research domains?	What are the trade-offs involved in making Type I in relation to Type II errors in various research contexts and research domains?	What is the appropriateness of the 5:1 ratio in the presence of multivariate normality and for various degrees of model complexity?

The first question we discussed is, Should we determine sample size by conducting a power analysis using Cohen's conventional definitions of small, medium, and large effect sizes? The answer to this question is no. First, Cohen's values are, by his own admission, largely subjective and may not be relevant in many research domains in the organizational and social sciences. Second, one should take context into account in choosing a targeted effect size for a power analysis. In many situations, what is commonly labeled as a small effect can have great significance for science and practice. Finally, rather than using Cohen's definitions, there are two preferred strategies for identifying a targeted effect size in a power analysis: (a) derive it from previous literature or (b) choose an effect size that will have significant implications for theory and practice. Unfortunately, using Cohen's definitions of effect size to conduct a power analysis is often used as a rationalization for concluding that a specific sample size is sufficiently large. In many cases, this argument is used inappropriately to avoid facing the inconvenient fact that a particular study's sample size is not sufficiently large to detect effect sizes that are practically or scientifically significant.

The second question we addressed is, When one has a small sample, is it advisable to increase the a priori Type I error rate to .10 or .20 to increase statistical power? The answer to this question is "it depends." If the increased α value is chosen arbitrarily, then the answer is no. However, if the increased value is chosen after a careful examination of the trade-offs involved between Type I and Type II error, then the answer is yes. Overall, an increase in the a priori Type I error rate is justified if the resulting value is chosen via an informed balancing of the trade-offs involved. Increasing the Type I error and choosing a value based on an informed decision is also likely to be more readily accepted by journal editors and reviewers as compared to choosing an arbitrarily larger value (e.g., .10 or .20). Unfortunately, arbitrarily increasing the a priori Type I error rate to .10 or .20 is often used as a rationalization for ignoring the result that the hypothesized effect is not statistically significant at the more traditional .05 or .01 levels. In many cases, as when Cohen's definitions of effect size are used, this argument is used inappropriately to avoid facing the inconvenient fact that a particular study's sample size is not sufficiently large to detect effect sizes that are practically or scientifically significant.

The final question we discussed is, Is it true that a sample size that includes 5 observations per estimated parameter in a covariance structure analysis leads to trustworthy estimates? The answer to this question is "it depends." In most situations in the organizational and social sciences in which the data do not follow a multivariate normality pattern, at least 10 observations per parameter estimated are needed. On the other hand, 5 observations per parameter estimated may suffice when the data are multivariate normal (which is not a frequent situation). Nevertheless, this is an oversimplified rule and a lower-bound value for the number of observations. Thus, researchers should not invoke the 5:1 rule of thumb to support a statement that the sample size is ideal. Unfortunately, using the 5:1 rule of thumb is often used as a rationalization for using a sample size that may be too small. In many cases, this argument is used inappropriately to avoid facing the inconvenient fact that a particular study's sample size is not sufficiently large, resulting in large standard errors and difficulties in replicating the findings in future studies.

In closing, the phrase *rule of thumb* has many purported origins. One of them is that is that the phrase originates from some of the many ways that thumbs have been used to draw inferences regarding the alignment or distance of an object by holding the thumb in one's eye-line, the temperature of brews of beer, or the estimated inch from the joint to the nail. We hope our critical analysis of the three rules of thumb regarding sample size will improve the way organizational and social scientists draw inferences from their own research.

Author Note

An abbreviated version of this manuscript was presented at the annual conference of the Society for Industrial and Organizational Psychology, New York, New York, April 2007. We thank Bob Vandenberg, Chuck Lance, Gilad Chen, Hank Sims, Larry James, and the Management & Organization doctoral students at the Robert H. Smith School of Business (University of Maryland) for constructive feedback on earlier versions of this manuscript.

This research was conducted, in part, while Herman Aguinis was on sabbatical leave from the University of Colorado Denver and holding visiting appointments at the University of Salamanca (Spain) and University of Puerto Rico.

Correspondence and requests for reprints should be addressed to Herman Aguinis, Mehalchin Term Professor of Management, The Business School, University of Colorado, Campis box 165, P.O. Box 173364, Denver, CO 80217–3364, http://carbon.cudenver.edu\~haguinis

References

Aguinis, H. (2004). *Regression analysis for categorical moderators.* New York: Guilford.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90,* 94–107.

Aguinis, H., Boik, R. J., & Pierce, C. A. (2001). A generalized solution for approximating the power to detect effects of categorical moderator variables using multiple regression. *Organizational Research Methods, 4,* 291–323.

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82,* 192–206.

Bentler, P. M. (1985). *Theory and implementation of EQS: A structural equations program.* Los Angeles, CA: BMDP Statistical Software.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bentler, P. M., & Chou, C. H. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16,* 78–117.

Boland, R. J., Singh, J., Salipante, P., Aram, J. D., Fay, S. Y., & Kanawattanachai, P. (2001). Knowledge representations and knowledge transfer. *Academy of Management Journal, 44,* 393–417.

Brews, P. J., & Tucci, C. L. (2004). Exploring the structural effects of internetworking. *Strategic Management Journal, 25,* 429–451.

Brown, K. G. (2001). Using computers to deliver training: Which employees learn and why? *Personnel Psychology, 54,* 271–296.

Brown, T. C. (2003). The effect of verbal self-guidance training on collective efficacy and team performance. *Personnel Psychology, 56,* 935–964.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

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

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

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/ Correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Epitropaki, O., & Martin, R. (2004). Implicit leadership theories in applied settings: Factor structure, generalizability and stability over time. *Journal of Applied Psychology, 89,* 293–310.

Garg, V. K., Walters, B. A., & Priem, R. L. (2003). Chief executive scanning emphases, environmental dynamism, and manufacturing firm performance. *Strategic Management Journal, 24,* 725–744.

Kervin J. B. (1992). *Methods for business research.* New York: Harper Collins.

Kim, H., Hoskisson, R. E., & Wan, W. P. (2004). Power dependence, diversification strategy, and performance in keiretsu member firms. *Strategic Management Journal, 25,* 613–636.

Kinicki, A. J., Prussia, G. E., Wu, J., & McKee-Ryan, F. M. (2004). Employee response to performance feedback: A covariance structure analysis using Ilgen, Fisher, and Taylor's (1979) model. *Journal of Applied Psychology, 89,* 1057–1069.

Lane, P. J., Cannella, A. A., & Lubatkin, M. H. (1998). Agency problems as antecedents to unrelated mergers and diversification: Amihud and Lev reconsidered. *Strategic Management Journal, 19,* 555–578.

Martel, R. F., Lane, D. M., & Emrich, C. (1996). Male-female differences: A computer simulation. *American Psychologist, 51,* 157–158.

Morgeson, F. P., & Campion, M. A. (2002). Minimizing tradeoffs when redesigning work: Evidence from a longitudinal quasi-experiment. *Personnel Psychology, 55,* 589–612.

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests.* Mahwah, NJ: Lawrence Erlbaum.

Pierce, C. A., Aguinis, H., & Adams, S. K. R. (2000). Effects of a dissolved workplace romance and rater characteristics on responses to a sexual harassment accusation. *Academy of Management Journal, 43,* 869–880.

Pierce, C. A., Broberg, B. J., McClure, J. R., & Aguinis, H. (2004). Responding to sexual harassment complaints: Effects of a dissolved workplace romance on decision-making standards. *Organizational Behavior and Human Decision Processes, 95*, 66–82.

Raver, J. L., & Gelfand, M. J. (2005). Beyond the individual victim: Linking sexual harassment, team processes, and team performance. *Academy of Management Journal, 48*, 387–400.

Sauley, K. S., & Bedeian, A. G. (1989). 05: A case of the tail wagging the distribution. *Journal of Management, 15*, 335–344.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Sturman, M. C., & Short, J. C. (2000). Lump-sum bonus satisfaction: Testing the construct validity of a new pay satisfaction dimension. *Personnel Psychology, 53*, 673–700.

Takeuchi, R., Yun, S., & Tesluk, P. E. (2002). An examination of crossover and spillover effects of spousal and expatriate cross-cultural adjustment on expatriate outcomes. *Journal of Applied Psychology, 87*, 655–666.

Vandenberg, R. J., & Grelle, D. M. (2009). Alternate model specifications in structural equation modeling: Facts, fictions, and truth. In C. E. Lance & R. J. Vanderberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 165–191). New York: Routledge/Psychology Press.