© *Academy of Management Journal*
2010, Vol. 53, No. 1, 9–13.

FROM THE EDITORS

BIG SAMPLES AND SMALL EFFECTS: LET'S NOT TRADE RELEVANCE AND RIGOR FOR POWER

I began work on my Ph.D. almost 20 years ago, and I have noticed two interconnected trends over the years—one positive and one potentially negative—that seem worthy of attention. Scholars' ever-increasing ability to gather larger samples is the positive trend. I can recall thinking as I began designing my dissertation research: "If I can just get a sample over 100 then I should be able to publish my results in a good journal." I highly doubt such a low target is acceptable to many dissertation committees or journals today. Fears of low statistical power and concerns over representativeness would, I think, send most students back to the drawing board. An increase in remarkably small reported effects among large-sample studies is the corresponding and potentially negative trend that I have observed. As an *AMJ* reviewer and now as an associate editor, I see more and more studies in which correlations and standardized regression coefficients of .05 or less receive the prized label "highly significant."

Together, these two trends make me think of a line by Robin Williams as the voice of the genie in Disney's movie *Aladdin*. Williams describes the experience of being a genie as: "Phenomenal cosmic powers! . . . Itty-bitty living space." Although the increased availability of data and sophisticated statistical tools for handling them are major contributors to the advancement of our understanding of organizations, I wonder whether the corresponding phenomenal statistical power might mask other shortcomings of our research designs and leave us with itty-bitty effect sizes that limit the relevance of our research. In essence, I wonder whether large samples might contribute to our learning more and more about less and less. As management scholars, can we really suggest that managers should change their decision calculus on the basis of knowledge that some new variable explains .0025 percent of the variance in organizational performance? My purpose here is to explore the veracity of my observations and offer some suggestions for how we as scholars might go about maintaining methodological rigor and managerial relevance even as we increase the sizes of our samples.

INCREASING POWER, DECREASING EFFECTS

My premise about trends toward larger samples and smaller effects was based on anecdotal evidence and my idiosyncratic experiences, so it seemed reasonable to apply some data. Thus, I gathered all of the correlations from the quantitative studies published over the last two full calendar years in the *Academy of Management Journal* (2007 and 2008) and over the 2 years occurring 20 years earlier (1987 and 1988). This procedure resulted in correlations from 106 independent samples from recent studies and, because of increased publishing volume over time, only 57 studies from the older period. To gain some balance, I added studies from 1989, which resulted in a total of 86 older studies.

The data appear to support my anecdotal observations. The older studies averaged only 300 observations, in contrast to 7,578 for the newer studies ($p < .05$). If I removed three very large samples with over 75,000 observations (two over 150,000), the average among the new studies dropped to 3,423, but the statistical difference between the groups was even greater than when the outlier samples were included ($p < .001$). Comparing the same time periods, average effect sizes as measured by correlations (r's) fell from .22 to .17, a 23 percent drop that held when I removed large correlations (i.e., $> .80$) that presumably depict relationships among measures of the same constructs.

With the three outliers removed, I ran some regressions in an attempt to better understand what might be causing effects to shrink. I found that whereas surveys generated larger effects, newer studies, studies with large samples, and studies conducted at the firm level of analysis[1] generated smaller effects. Upon further investigation, how-

I offer my thanks to Ryan Bowen and Sean Normand for their help collecting data. I also benefited greatly from the insights of Peter Bamberger, Russell Crook, Duane Ireland, Micki Kacmar, Dave Ketchen, Elizabeth Morrison, and Jeremy Short.

[1] I excluded 16 studies (5 old, 11 new) pairwise that had units of analysis such as processes, grievances, and expenditures. Including 7 new studies with teams as the unit of analysis did not materially impact results.

ever, the reason that firm-level studies reported smaller effects appeared to be their heavy reliance on secondary data rather than surveys. Although large-sample studies appeared to have always reported smaller effects—the overall correlation between sample size and effect size was $-.26$ ($p < .001$)—the negative impact of sample size was significantly larger among the new studies. Admittedly, this was a fairly coarse analysis in that I did not limit it to those relationships of central interest to the researchers. Still, it suggests to me that as our standards for what constitutes a desirable sample increase, the sizes of the relationships that we report appear to shrink.

The question is, Why? If we assume that the relationships we investigate are not slowly disappearing, it must be that only *reported* effects are in decline. In which case, statistical theory offers the most obvious explanation. With small samples, it is difficult to know whether an effect is "real" (i.e., not zero) or simply the result of random sampling error, so statistical theory asserts that we need fairly large effects before we can claim confidently (i.e., type I error is kept to $p < .05$) that we have indeed found something (Cohen, 1988). As sample size increases, however, random sampling error decreases, and we can have the same level of confidence with smaller effects. At the extreme, the largest sample among those I captured was 212,014 observations (Miller, Fern, & Cardinal, 2007). Correlations of .0043, which round to .00, are statistically significant with a sample size that large. Thus, the most obvious explanation for why reported effects are shrinking is that our newfound statistical power is allowing us to claim significance for smaller and smaller effects. Given reviewers' and editors' strong preference for publishing statistically significant results, both at *AMJ* and elsewhere, these newly significant and smaller effects are seemingly being reported in greater numbers than was the case 20 years ago.

THE RESEARCH IMPLICATIONS OF LARGE SAMPLES

One could reasonably argue that the trends I have observed reflect an important step forward in that our large samples have brought about a notable increase in our ability to identify small, but important and real, relationships that we could not otherwise detect. Rapidly increasing the pool of known relationships has allowed for the development and support of increasingly complex and interesting theories that explain organizational phenomena. In the large-sample example noted above, Miller et al. (2007) proposed a theory that explains how knowledge transfer among corporate divisions leads to more impactful innovations than those based on existing divisional knowledge or knowledge transfer between firms. It is difficult to see how one might test such a theory without a large sample of patent-level data.

Are there dangers, though, in developing ever more complex theories and testing them with increasingly large samples? Two concerns come to mind. The first is that it is possible that our collective infatuation with large samples might cause us to relax our vigilance regarding construct validity. In essence, it might be tempting to view phenomenal statistical power as an effective substitute for accurate measurement. The second concern is that we might fool ourselves into believing that statistical significance is equivalent to theoretical or managerial significance. Effect size matters; managers and researchers alike should be concerned not only with whether a theory has support, but also with the strength of the support (Eden, 2002). Each of these concerns has implications for how we can best conduct and report our research.

Implications of Large Samples for Construct Measurement

There is a well-known systematic positive relationship between construct validity and effect size. If two measures have perfect validity and there is no sampling error, then their sample correlation will equal the population effect. Deviations from perfect validity, however, lower effects in such a way that if we could quantify precisely how much each measure deviates from perfect validity, then the correlation between the two imperfect measures could be "corrected" to arrive back at the population effect (Hunter & Schmidt, 2004). The negative impact of poor construct validity can be quite dramatic (Schmidt, Hunter, & Urry, 1976); thus, with relatively small samples, researchers must pay close attention to construct validity. Failure to do so reduces effect sizes, the probability of finding significant results and, consequently, the probability of publication. With large samples, however, increased statistical power means that even poorly measured constructs often will find significance.

One could reasonably argue, I think, that relaxing a bit on measurement standards is a luxury afforded by larger samples, but it is one that overlooks the potential cost in terms of knowing whether our theories are truly supported. By way of example, Bromiley and Johnson (2005) observed that R&D intensity (i.e., R&D/total sales) has been used as a measure of asset specificity to test trans-

action cost theory and as a measure of research capability to test resource-based theory. They argue that R&D intensity depicts neither the specificity of funded R&D projects nor how productively those resources are used. It measures only the relative amount of financial resources devoted to R&D and, as Ketchen, Boyd, and Bergh (2008) pointed out, if resource expenditures were a good measure of capability, New York Yankee baseball fans might not have had to wait the past nine years before enjoying a World Series win. Using R&D as a proxy because that is what is available in a large database might lead to a statistically significant effect; however, because of poor construct validity, we cannot be certain that the effect represents support for the relevant theory (Ketchen et al., 2008).

Poor construct validity potentially obstructs the advance of knowledge in another way as well. A key goal of meta-analysis is to assess the size of a relationship depicted in a body of research (Eden, 2002). Most aggregation formulae weight studies by sample size, so large samples influence effect size estimates more than small samples (e.g., Hunter & Schmidt, 2004). If large-sample studies use less valid measures and consequently report smaller effects, then future attempts to assess the level of support for important theoretical relationships will underreport effect size estimates.

Whether researchers are actually relaxing standards when samples are large is an open question; however, at least in some areas of management research, the prospect that poor measurement is reducing effect sizes seems quite high (e.g., Boyd et al., 2005). Effects deflated by measurement error are less likely to be significant and therefore published unless a sample is large and offers enough statistical power to make poorly measured relationships significant. In this way, larger samples shift some of the burden for assessing construct validity from statistical theory onto authors, reviewers, editors, and research consumers. Small effects from poor measures will not be "kicked out" as nonsignificant if a sample is large, so statistical significance does not necessarily signal good measurement in a large-sample study. Consequently, authors incur an increased burden to argue that measures correspond to their theoretical construct definitions, and those who evaluate the research have a greater responsibility to assess the clarity of those arguments.

Implications of Small Effects for Theoretical and Managerial Relevance

If we researchers are maintaining, and perhaps even improving, our measurement practices over time, then the logical explanation for shrinking effect sizes is that we are increasingly capable of detecting smaller and smaller effects. Such effects are real, but we have not previously had the means to confirm them statistically. By all accounts, this capability represents a scientific advance. Increasing our ability to claim smaller effects as statistically significant does not, however, change their theoretical or managerial relevance.

Miller et al. (2007) can again serve as an example. Drawing on theory about how divisionalized structures create and distribute knowledge, they predicted that knowledge transferred among an organization's divisions would be more impactful than knowledge developed within a division or from outside the organization. They found that each additional interdivisional patent citation led to a .018 ($p < .005$) increase in the number of subsequent citations, all else being equal. This effect is "real" in the sense that it is statistically different from zero, and it shows clear support for their theory.

The important question for future researchers, however, is whether the effects are large enough to pursue further theoretical development. Is it worth our time and effort to build theory that identifies boundary conditions for where interdivisional knowledge transfer will *not* be effective? What about moderators, such as organizational decentralization, that might impact the effectiveness of such knowledge transfer? Such questions only make sense in the context of understanding how much support we can show for the theory. A theory might find support, but its explanatory power—that is, the effect size observed—is so weak that further efforts to develop the theory might not be warranted.

Small effects also raise questions about managerial relevance. It is an important requirement that articles published in *AMJ* build, extend, or test theories that help explain organizational phenomena that are relevant to managers. However, when samples become large and effects become small, the risk increases that the managerial relevance we collectively seek will remain elusive. Managers need to have some confidence that acting upon the theories that we present is likely to have a noticeable impact on their organizations. Organizations are complex, and managers are not able to "hold constant" important confounds as we do in our statistical analyses. Thus, if managers begin to act on theories that are supported by small effects, they are not likely to notice positive results even when they occur. Such a development would surely further damage the dubious reputation that our scholarship has among some managers (Hambrick, 1994).

RECOGNIZING POWER AND HIGHLIGHTING RELEVANCE

There are a couple of simple steps that we can take to recognize when statistical power is so great that authors and readers must be extraordinarily vigilant about construct validity and relevance. A first step is to report statistical power. Power is rarely reported because power analyses are typically conducted in the planning stages of a research project, so that researchers know what sample size is needed for them to have a specific probability—(typically 80 percent) of finding small, medium, or large population effects. (See Cohen [1988] for commonly used conventions for what constitutes small, medium, and large for different effect size statistics.) Knowing that you need a sample of 785 to have an 80 percent chance of finding a small population effect is important when you plan to invest a great deal of labor into an investigation that you hope will return interesting and publishable findings, but it might not appear important to readers. At 3,423 observations, however, the average sample today has nearly 100 percent power with respect to small effects, meaning that small population effects of $r = .10$ will almost certainly be found and that even population effects as small as $r = .044$, which rounds to zero, have a 73 percent chance of being found. A line in the description of the sample that tells readers that this "average" sample has an 80 percent chance of finding population effects as small as $r = .048$ should help put readers on the alert that statistical significance should not be the only—or even the primary—criterion for evaluating construct validity or relevance.

Power analysis tells readers that they must be vigilant, but it does not say anything about the extent to which statistically significant effects are theoretically or managerially relevant. Researchers can take two actions to demonstrate relevance. They can address the subject of effect size in their Results sections by reporting standardized regression coefficients (betas) where possible. When the dependent variable is on a large scale (e.g., profits, firm sales) relative to the independent variables (e.g., R&D intensity, Likert scales), unstandardized coefficients can look large even when they are not particularly meaningful. Where prediction is the central focus of the research, unstandardized coefficients are essential. Macro economists and business practitioners, for example, use unstandardized regression coefficients to make predictions about next year's GDP or unit sales. When the purpose is testing theory, however, standardized coefficients will give readers an intuitive idea as to the relative size of the relationships under investigation. An alternative to reporting standardized coefficients is to calculate and report confidence intervals. A confidence interval's width offers an easy-to-understand view of an effect estimate's accuracy, and the lower bound (the upper, in the case of a negative effect) shows proximity to zero. Either way, if you report standardized regression coefficients or confidence intervals, readers will not need a calculator in hand to know whether your results are theoretically and managerially significant and not just statistically significant.

Another action to take to demonstrate relevance is to use space in your Discussion section to describe the impact of the independent variables on the dependent variable in plain English. Exactly how much would sales, profits, employee retention, or job satisfaction, for example, change if managers were to make a one unit, or one standardized unit, change in the independent variable? If a moderator is hypothesized, how much exactly does the focal relationship change when the moderator is one standard deviation above versus below average? Indeed, if the effects under investigation are meaningful, they can often be translated into dollar amounts that highlight their potential impact. A savings of only $100 per employee from a work practice intervention, for example, might be quite dramatic in a company with several thousand employees.

As examples, in a study of management faculty, Gomez-Mejia and Balkin (1992) reported "a 37 percent annual pay differential between the highest and lowest producers of top-tier publications," and Huselid (1995) found that a one standard deviation increase in the use of high-performance work practices related to "$27,044 more in sales and $18,641 and $3,814 more in market value and profits, respectively." Though it might also be helpful to have a clear idea of what a one standard deviation change looks like, these statements are clear to both researchers and managers. I should note that the examples I am offering here are both from articles that received *AMJ*'s annual Best Article Award, which suggests to me that there is merit in following the approaches that these scholars used in their Discussion sections. Yes, savvy readers can make these calculations, but why should they? Describing the real impact of a relationship will not take much space, and it will offer readers an intuitive understanding of the potential impact of your research.

CONCLUSION

My conclusion, obviously, is that sample and effect size matters a great deal. Rephrasing Robin Williams's genie, large samples potentially give us "phenomenal statistical power! . . . itty-bitty effect

sizes." Certainly, this does not need to be the case if underlying relationships are large and we researchers pay close attention to construct validity. The statistical power rendered by large samples does, however, give us the ability to find small effects, and it seems to me that such power increases our responsibility to insure that other research design problems are not masked, and that we look less to statistical significance and more to theoretical and managerial significance in evaluating our research.

Returning to my ambitions for my own dissertation research almost 20 years ago, perhaps a sample of 100 is not so bad if construct validity is high. If we are going to send our students back to the drawing board, let us show balance in our concern for sample and effect size. Having an adequate sample is obviously important, but the long-term impact of our research will be judged more by whether we can show strong evidence that our theories are correct and of real benefit to managers.

James G. Combs
Florida State University

REFERENCES

Boyd, B. K., Gove, S., & Hitt, M. A. 2005. Construct measurement in strategic management research: Illusion or reality? ***Strategic Management Journal,*** 26: 239–258.

Bromiley, P., & Johnson, S. 2005. Mechanisms and empirical research. In D. J. Ketchen & D. D. Bergh (Eds.), ***Research methodology in strategy and management***: 15–30. Oxford, U.K.: Elsevier.

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

Eden, D. 2002. Replication, meta-analysis, scientific progress, and *AMJ*'s publication policy. ***Academy of Management Journal,*** 45: 834–846.

Gomez-Mejia, L. R., & Balkin, D. B. 1992. Determinants of faculty pay: An agency theory perspective. ***Academy of Management Journal***, 33: 921–955.

Hambrick, D. C. 1994. Presidential address: What if the Academy actually mattered? ***Academy of Management Review***, 19: 11–16.

Hunter, J. E., & Schmidt, F. L. 2004. ***Methods of meta-analysis*** (2nd ed.). Thousand Oaks, CA: Sage.

Huselid, M. A. 1995. The impact of human resource management practices on turnover, productivity, and corporate financial performance. ***Academy of Management Journal,*** 38: 635–672.

Ketchen, D. J., Boyd, B. K., & Bergh, D. D. 2008. Research methodology in strategic management: Past accomplishments and future challenges. ***Organizational Research Methods***, 11: 643–358.

Miller, D. J., Fern, M. J., & Cardinal, L. B. 2007. The use of knowledge for technological innovation within diversified firms. ***Academy of Management Journal***, 50: 308–326.

Schmidt, F. L., Hunter, J. E., & Urry, V. E. 1976. Statistical power in criterion-related validation studies. ***Journal of Applied Psychology***, 61: 473–485.