Statistical Power and the Testing of Null Hypotheses: A Review of Contemporary Management Research and Recommendations for Future Studies

LUKE H. CASHEN
Louisiana State University

SCOTT W. GEIGER
University of South Florida St. Petersburg

The purpose of this study is to determine how well contemporary management research fares on the issue of statistical power with regard to studies specifically predicting null relationships between phenomena of interest. This power assessment differs from traditional power studies because it focuses solely on studies that offered and tested null hypotheses. A sample of studies containing hypothesized null relationships was taken from five mainstream management journals over the 1990 to 1999 time period. Results of the power assessment suggest that management researchers' abilities to affirm null hypotheses are low. On average, the power assessment revealed that for those studies that found nonsignificance of results and consequently affirmed their null hypotheses, the actual Type II error rate was nearly 15 times greater than what is advocated in the literature when failing to reject a false null hypothesis. Recommendations for researchers proposing and testing formal null hypotheses are also discussed.

Keywords: *hypothesis testing; null hypothesis; statistical power; power analysis*

Following Cohen's (1962) study on the statistical power of abnormal and social psychological research, a multitude of power assessment articles were published that investigated power issues in a variety of disciplines. For instance, power surveys have been performed in areas such as management (e.g., Ferguson & Ketchen, 1999; Mazen, Graf, Kellogg, & Hemmasi, 1987; Mazen, Hemmasi, & Lewis, 1987), management and applied psychology (e.g., Mone, Mueller, & Mauland, 1996), psychology (e.g., Rossi, 1990; Sedlmeier & Gigerenzer, 1989), management information systems (e.g., Baroudi & Orlikowski, 1989), education (e.g., Brewer, 1972), commu-

Authors' Note: Both authors contributed equally to this article. The authors wish to thank Marcia Simmering and Michael Sturman for their comments on prior drafts.

Organizational Research Methods, Vol. 7 No. 2, April 2004 151-167
DOI: 10.1177/1094428104263676

nication (e.g., Chase & Tucker, 1975; Katzer & Sodt, 1973), and marketing (e.g., Sawyer & Ball, 1981). In general, the common finding of these power assessments is that insufficient statistical power plagues research in these diverse areas of study.

Such findings could be attributable to the fact that power issues and power analyses tend to receive inadequate attention by researchers. Cohen (1992) addressed this issue by stating, "It is not at all clear why researchers continue to ignore power analysis. The passive acceptance of this state of affairs by editors and reviewers is even more of a mystery" (p. 155). Equal frustrations were noted by Sedlmeier and Gigerenzer (1989), who suggested that concerns about power in research are almost nonexistent, at least in print. This conclusion was supported by the fact that fewer than 5% of the studies in their power assessment mentioned power, and no studies estimated the power levels of their tests. Such a finding is in congruence with Cohen (1977), who stated that in reports of research in which power issues are obviously important, the issue is frequently not addressed. Nickerson (2000) suggested that such inattention might be attributable to statistical power not frequently being understood and, as a result, not often employed in research. Additional evidence was offered by Mone et al. (1996), who noted that the impact of past power assessment surveys has been minimal and that calls for greater statistical power levels and usage have gone unheeded. In their study, Mone et al. surveyed the authors of a sample of studies contained in top-tier journals and found that almost two thirds of the respondents never used power analysis. Furthermore, the respondents stated that there is little call for greater usage of power analysis by journal editors or reviewers.

Echoing these aforementioned concerns, authors of power assessment articles argue that insufficient statistical power may leave researchers unable to detect or reject false null hypotheses. In other words, researchers may actually overlook meaningful differences or effects as a result of low power. Cohen (1977) noted that such an occurrence is highly unfavorable to behavioral scientists because it is then reasonable to suggest that there is not an equitable chance of rejecting the null hypothesis, and, in general, behavioral scientists "typically hope to 'reject' [the null] hypothesis and thus 'prove' that the phenomena in question is in fact present" (p. 1). Cohen and other authors of statistical power assessments basically conclude that a failure to reject a null hypothesis leaves readers wondering whether it is due to insufficient statistical power or truly due to the absence of the phenomenon. Additional concerns about designing studies with low power are noted by Howard, Maxwell, and Fleming (2000) who suggested that such actions "tend to lead to a body of literature in which results appear to contradict one another" (p. 316).

It is common knowledge that researchers usually wish to demonstrate that the phenomenon in question is present (i.e., reject the null hypothesis in favor of the alternative hypothesis). However, there are instances in which researchers do have a priori, theoretically justified reasons to hypothesize formal, statistical null relationships (Cohen, 1990; Cortina & Dunlap, 1997; Cortina & Folger, 1998; Greenwald, 1975, 1993). Support for positing and testing null relationships between variables of interest is offered by Greenwald (1993), who noted that "scientific advance is often most powerfully achieved by rejecting theories. A major strategy for doing this is to demonstrate that relationships predicted by a theory are not obtained, and this would often require acceptance of a null hypothesis" (p. 421). Others have argued that the tenability of the null hypothesis is as legitimate a goal of research as is demonstrating the tenability of any alternative hypothesis (Chow, 1996; Cortina & Folger, 1998; Frick, 1995;

Nickerson, 2000). We do believe that theoretically based arguments that lead researchers to predict null relationships between their research variables of interest are justified. Furthermore, data collected in this study suggest that management journals have accepted the practice of formulating and testing null hypotheses.

We use the term *null hypothesis* to mean a hypothesis of no nontrivial effect, no nontrivial correlation, or no nontrivial difference—a distinction that is in line with previous research (e.g., Cohen, 1977; Cortina & Dunlap, 1997; Cortina & Folger, 1998). It is important to note that a null hypothesis is different from a nil hypothesis (also known as a point-null hypothesis), which is a hypothesis for which the value to be nullified is precisely zero (Cohen, 1977; Cortina & Dunlap, 1997). It is our contention that when management researchers formally hypothesize a null relationship between constructs of interest, it is not their desire to demonstrate that the true value of a statistic equals zero but rather that the effect or relationship to be tested is so small as not to be usefully distinguishable from zero (Cohen, 1977; Cortina & Dunlap, 1997; Cortina & Folger, 1998). Such an argument is reasonable because a vast majority of situations indicative of management research are likely to produce effects or relationships that are not precisely zero—a situation similar to psychological research as reported by Cortina and Folger (1998). Thus, to reject a nil hypothesis (i.e., effect size equals zero) is not a great accomplishment for management researchers. Instead, the focus should be directed at determining whether the effect size is negligible or trivial (Cohen, 1977, 1990; Greenwald, 1993). Thus, our use of the term *null hypothesis* represents a hypothesis of a trivial effect instead of a hypothesis of nil effect.

Having taken notice of the fairly discouraging findings of previous power assessment studies, we are interested in how well contemporary management research fares on the issue of statistical power solely in relation to formal null hypotheses presented and tested in published research. As a result, the purpose of this study is (a) to assess the statistical power of a sample of management research studies that specifically hypothesize null relationships and (b) to discuss the implications of these findings for management researchers and offer recommendations for authors presenting and testing null hypotheses in future research papers.

This investigation is important for two main reasons. First, of all the power assessment studies to date, none have focused solely on analyzing the power of studies that offer formal null hypotheses. As such, this article serves to determine whether failures to reject null hypotheses, as reported in the findings of our sample, can be made with a high degree of confidence. If power levels are not sufficiently high, then reaching conclusions of "no effect" may not be realistic and potentially lead to conflicting and/or invalid findings in the literature. Only if power levels are high can one come close to inferring that the null hypothesis is true when there is a failure of a test to establish statistical significance (Nickerson, 2000). Thus, power is particularly important for those testing null hypotheses because lack of power may in fact lead researchers to incorrectly affirm null hypotheses. By conducting power analyses on our sample of research articles, we can come one step closer to addressing this issue. Second, this article consolidates previous work that addresses null hypothesis testing (e.g., Cohen, 1977; Cortina & Folger, 1998; Sedlmeier & Gigerenzer, 1989; Rossi, 1990) with the intent of providing management researchers who specifically offer theoretically based null hypotheses in their research, with specific guidelines for testing such relationships. Such guidelines are important because predicting a null relationship (e.g., there is no relationship between X and Y) requires an emphasis on Type II errors versus

predicting a directional relationship (e.g., X and Y are positively related), which requires guarding against Type I errors.

Components of Statistical Power and the Null Hypothesis as a Research Hypothesis of Interest

There are three main components that determine the level of statistical power of an inference test: the significance level (α), the sample size, and the effect size. The relationship between power and its three determinants is such that if one of the four elements (i.e., power, significance level, sample size, or effect size) is unknown, it can be calculated using the known values of the other three elements. Hence, researchers are often able to a priori determine statistical power levels of their tests.

Researchers investigating a phenomenon typically hypothesize that a relationship between the investigated variables exists. Classical statistical inference tests posit a null hypothesis (H_o: the phenomenon under investigation is absent, or there is no—or at best a trivial—difference between the parameters being tested), which researchers contrast against the alternative hypothesis (H_a: the phenomenon is present, or there is a difference in the parameters being tested). Because researchers typically hope to reject the null hypothesis, they normally report the probabilities associated with the likelihood that such a conclusion is erroneous (i.e., α). However, when such tests are not significant or when one expects the null hypothesis to be upheld, it is critical to discuss the likelihood of rejecting the null hypothesis in favor of the alternative hypothesis if the alternative hypothesis is in fact true. Such a probability is better known as statistical power. Frequently, power is represented as $1 - \beta$, where β is the probability of failing to reject the null hypothesis when it is actually false. Such an error is commonly referred to as a Type II error.

The following section describes each of the power determinants and emphasizes the need to consider these issues when testing null hypotheses. In other words, the specifics of how these determinants relate to testing null relationships are the focus of this section rather than a mere overview of the well-known components of statistical power.

Significance Level

Interpreting statistical inferences mandates that researchers specify acceptable levels of statistical error. The most common approach is to specify the level of Type I error, generally represented as α. Formally defined, a Type I error is the probability of rejecting the null hypothesis when it is actually true. On average, the attention paid by researchers to the two types of statistical inference errors (Type I and Type II) is by far not equal. The belief is that the consequences of a false positive (Type I error) claim are more serious than those of a false negative (Type II error) claim. As a result, Type I errors are usually focused on more frequently and guarded against more stringently by researchers (Baroudi & Orlikowski, 1989; Brewer, 1972; Chase & Chase, 1976; Cohen, 1977; Cowles & Davis, 1982; Greenwald, 1993; Mazen, Graf, et al., 1987; Myers & Melcher, 1969; Mone et al., 1996; Nickerson, 2000; Rossi, 1990; Sedlmeier & Gigerenzer, 1989). However, it is important to note that several authors (e.g., Mazen, Graf, et al., 1987; Sauley & Bedeian, 1989) advocate distributing the risk between Type I and Type II errors because the risks/consequences associated with

committing a Type II error may be extremely costly. In other words, the ratio of β:α should be sensitive to the relative risks inherent in Type I and Type II errors for every test conclusion (Mazen, Graf, et al., 1987; Sauley & Bedeian, 1989).

Echoing Cohen (1977) and Sedlmeier and Gigerenzer (1989), we suggest that researchers set their level of β to correspond to the traditional level of α when testing a non-null (i.e., alternative) hypothesis, which typically is set at the .05 level. Thus, when null hypotheses serve as the research hypotheses of interest, the researcher should opt for a β level of .05, which corresponds to a .95 power level; otherwise, statistical insignificance of the tests has no real significance. Because power is $1 - \beta$, then at a power level of .80, $\beta = .20$, which means that there is a .20 probability of sustaining a false null. We argue that this represents a power level too low and a probability of Type II error too high to confidently affirm the null hypothesis. Such arguments are supported by Rossi (1990) who suggested,

> If power was high, then failure to reject the null can, within limits, be considered as an affirmation of the null hypothesis, because the probability of a Type II error must be low. Thus, in the same way that a statistically significant test result permits the rejection of the null hypothesis with only a small probability of error (alpha, the Type I error rate), high power permits the rejection of the alternative hypothesis with a relatively small probability of error (beta, the Type II error rate). (p. 646)

We fully understand that some advocate setting β levels according to each situation and the overall cost of the error, but we firmly advocate a minimal power level of .95 ($\beta = .05$), if possible, for researchers to have confidence in their results and to guard against building a literature of contradictory results. This is particularly important for those testing null hypotheses.

Sample Size

As the number of observations in the sample increases, the reliability (i.e., precision) of the sample value approximating the population value also increases (Cohen, 1977). As a result of this greater reliability, a researcher has a higher probability of rejecting a false null hypothesis. Thus, as the sample size increases, so does the power of the study. Ideally, researchers should specify α, effect size, and the desired level of power and then determine the sample size needed in the study so that more valid conclusions can be drawn from the results of testing null hypotheses.

Effect Size

The effect size represents the magnitude or strength of the relationship between the variables in the population (Cohen, 1977). As previously argued, researchers can fail to reject the null hypothesis when the true relationship between two events of interest is determined to be trivial or inconsequential. In other words, affirmation of null hypotheses does not occur when the true value of a statistic equals exactly zero, but rather the relationship between variables of interest is so small as not to be usefully distinct from zero. Cohen (1977) and Sedlmeier and Gigerenzer (1989) argued that determination of a trivial effect is made when power ($1 - \beta$) is set at a high value and the sample size used is large enough so that the risk of Type II error (β) is relatively small and similar to that of the risk of Type I error, which is commonly set at the .05 level.

When conducting a power analysis as part of testing the null hypothesis, it is important to determine when an effect is large enough to be considered nontrivial. Lane, Cannella, and Lubatkin (1998) notes that, conceptually, a trivial effect implies a small effect size, as defined by the conventional values set forth by Cohen (1977, 1992). Cohen (1977, 1990) demonstrated that if a researcher considers an effect size of $r = .10$ (a small effect size for a correlation according to Cohen) as negligible and wishes to test the null hypothesis ($\alpha = .05$, power = .95, and $\beta = .05$), then a sample size of 1,308 is required. It is obvious that the use of small effect sizes places great demands on the sample sizes of studies. From this example, it appears that it takes an impractically large sample size to fail to reject the null hypothesis; however, "the procedure makes clear what it takes to say or imply from a failure to reject the null hypothesis that there is no nontrivial effect" (Cohen, 1990, p. 1309).

So, with a small value for the effect size (i) and power set at a high value (so that β is relatively small), nonsignificance of results allows the researcher to properly conclude that the population effect size is no more than i (i.e., negligible), a conclusion significant at the specified level of β. Thus, drawing on the logic with which we reject the null hypothesis with risk equal to α, Cohen (1977) stated,

> The null hypothesis can be accepted in preference to that which holds that the effect size equals i with risk equal to β. Since i is negligible, the conclusion that the population effect size is not as large as i is equivalent to concluding that there is "no" (nontrivial) effect. (p. 16)

For this power assessment, we felt that drawing on general approximations of small effect sizes for the statistical tests covered by Cohen (1977, 1992) was appropriate.

Statistical Power in Contemporary Management Research

In this section, we report on the methods and results of a survey we undertook to determine the attention paid to the issue of statistical power in contemporary management research and to determine the level of statistical power characterizing many published management studies. Specifically, this survey focused only on those studies that offered and empirically tested null hypotheses. We conclude this section by discussing some of the implications of our findings for progress in, and interpretation of, management research.

Method

To assess the level of statistical power in studies proposing null hypotheses, we reviewed five major journals that published management research over the 10-year period of January 1990 to December 1999. Previous surveys of statistical power in abnormal and social psychology, education, sociology, applied psychology, social work, and marketing have traditionally reviewed studies published in one or a few volumes of one journal in their respective disciplines. Notable exceptions to this trend include Mazen, Graf, et al. (1987), Rossi (1990), Mone et al. (1996), and Ferguson and Ketchen (1999), who incorporated multiple journals and/or multiple years in their power assessments. Our study followed the lead of these studies by drawing on pub-

lished research in the *Academy of Management Journal*, *Administrative Science Quarterly*, *Strategic Management Journal*, *Journal of Management*, and *Journal of Management Studies*.

These journals were chosen for this study because they reflect current research in the management discipline, contain empirical studies, are peer reviewed, have high rejection rates, and have scholarly orientations. In addition, these 5 journals served as the sampling domain of this power assessment because they are consistently identified as having large impacts on the management discipline (Gomez-Mejia & Balkin, 1992; Sharplin & Mabry, 1985). Specifically, Sharplin and Mabry (1985) included these 5 journals in their listing of the top 10 management journals according to their impact on the management discipline. Furthermore, all 5 of the chosen journals were included in a top-tier management journal ranking as developed by Gomez-Mejia and Balkin (1992). These authors noted that articles published in these journals are to be "considered as premier publications" (p. 934).

It is important to note that this power assessment examined studies from management journals that exclusively published macro- and/or micro-management research. As such, leading journals that publish a mix of management and/or applied psychology research were excluded from the sample. Such a distinction is in line with previous power assessments, whose authors have chosen to include strictly management journals in their assessments of management research (e.g., Mazen, Graf, et al., 1987; Mazen, Hemmasi, et al., 1987; Schwenk & Dalton, 1991). This is in contrast to the power assessment by Mone et al. (1996), which included journals that published a mix applied psychology and management research.

The decision to review empirical studies published over a 10-year time frame (the 1990s) was made for two reasons. First, the most recent assessments of statistical power of management research are now dated (e.g., Mazen, Graf, et al., 1987; Mazen, Hemmasi, et al., 1987), but whether their general findings of insufficient power in management research are indicative of current practices are unknown. As such, starting our review of the journals in 1990 left ample time for management researchers to absorb these findings. Second, although a vast majority of power assessments in a variety of disciplines have opted for a very limited time frame for examining data (i.e., usually 3 years or fewer), we decided that an adequate data set would include 10 years of data.

Our search uncovered 53 articles formally hypothesizing 99 null relationships. We did not include null relationships that were predicted and tested (e.g., moderator regression) as part of a proposed interaction between variables (e.g., Hypothesis 1a: At high levels of Z, there will be a significant positive relationship between X and Y; Hypothesis 1b: At low levels of Z, there will be no relationship between X and Y). Table 1 shows the distribution of the chosen studies and the null hypotheses listed according to the journals in which they were published.

Specific examples of null hypotheses that were included in this study are as follows. In an effort to examine the linkage between executive succession and the extent of corporate strategic change across a sample of *Fortune* 1000 diversified firms, Wiersema (1992) hypothesized that "there will be no association between the degree of strategic change prior to succession and the nature of executive succession" (p. 80). Another study, conducted by Hubbard, Vetter, and Little (1998), investigated the issues associated with publishing replication and extension research in management journals. Specifically, these authors hypothesized that "there is no difference in the

Table 1
Distribution of Studies Containing Formal Null Hypotheses

Journal	*Number of Studies With Null Hypotheses*	*Percentage of Studies in the Sample*	*Number of Null Hypotheses*	*Percentage of Null Hypotheses in the Sample*
Academy of Management Journal	17	32.1	24	24.3
Strategic Management Journal	17	32.1	39	39.4
Journal of Management	6	11.3	10	10.1
Administrative Science Quarterly	7	13.2	12	12.1
Journal of Management Studies	6	11.3	14	14.1
Total	53	100	99	100

timeliness of replication articles published in the first, second, and third tiers of the journal hierarchy" (p. 245). Investigating the impact of top management team heterogeneity on a firm's competitive moves, Hambrick, Cho, and Chen (1996) theoretically argued that "top management team heterogeneity is unrelated to the firm's response propensity" (p. 667). Last, in their investigations of the effects of innovation and market complexity on firm performance, Lawless and Anderson (1996) argued that "a firm's average distance in resource space from firms in other niches has no significant effect on its performance" (p. 1196).

To remain consistent with previous power assessments, we determined power for each null hypothesis statistical test reported in the articles providing that it was a major statistical test for which Cohen's (1977) power analysis tables and formulae were available (i.e., R^2, β, F, t, r, and p). All studies in this sample used parametric tests; thus the use of equivalent parametric tests in lieu of nonparametric tests was not necessary as it has been in previous power assessment articles (e.g., Sedlmeier & Gigerenzer, 1989). In addition, following Mazen, Graf, et al. (1987) and Mone et al. (1996), we omitted secondary tests such as manipulation checks and peripheral reliability estimates.

Of the 53 articles selected, 5 studies hypothesizing 9 null relationships did not contain sufficient information for us to determine the correct level of statistical power associated with the testing of the null hypothesis and as a result were not included in the final sample. On the surface, it appears that there was low power in 2 of the studies; however, without the actual power levels, no conclusions could be made. In addition, 5 other studies hypothesizing 13 null relationships were dropped from this study because they used statistical techniques for which no conventional power analysis tests were available. As a result, our final sample consisted of 43 studies hypothesizing 77 statistical null hypotheses.

As in past surveys of statistical power in management and other disciplines, the power of each test was determined by using the study's given sample size, setting $\alpha = .05$, and, as noted earlier, specifying a small effect size.[1] Following such a procedure, we calculated the necessary sample size for each test using the study's given sample

size, setting α = .05, power = .95, which makes β = .05, and choosing the nondirectional critical region for all power calculations.

Results

Before performing quantitative analyses of power within the sampled studies, a qualitative assessment of the treatment of power reveals an interesting pattern. Of the 43 studies in our sample, 4 (9.3%) discussed the statistical power associated with the testing of the null hypotheses. Of these studies, 3 elaborated on the specific procedures for determining the statistical power of the tests. These 3 studies reported power levels ranging from .48 to .98. The fourth study did not contain the exact procedures for determining the statistical power of the tests yet mentioned that low levels of power, in general, were a contributing factor to nonsignificant findings. Thus, overall, 90.7% of the sampled studies did not reference the statistical power of their hypothesis tests, at least in print. These findings are similar to those reported by Mazen, Graf, et al. (1987) who suggested that "an important finding of this study was that explicit consideration of power issues was almost nonexistent" (p. 376) and Sedlmeier and Gigerenzer (1989) who reported that out of 54 articles in their sample (64 experiments), remarks on power were found in only 2 cases and power estimates were absent in all cases.

Our empirical investigation revealed that the average statistical power of all 77 null hypothesis tests across all studies was .29. Thus, as a group, these studies had a .71 probability of failing to reject the null hypothesis when the null hypothesis was actually false (i.e., a 71% chance of committing a Type II error).[2] The implication of this finding is that within these studies, a sizeable chance existed for researchers to erroneously conclude that their null hypotheses were affirmed and the phenomena of interest were trivial/absent.[3] This finding is in congruence with the findings reported by other power assessments in the management discipline. For example, Mazen, Graf, et al. (1987); Mazen, Hemmasi, et al. (1987); and Mone et al. (1996) each reported mean power levels of .31, .23, and .27 for their samples, respectively. In addition, our results are in line with Rossi (1990), who summarized the findings of 25 power surveys of behavioral research and noted that the mean power level across these 25 studies to detect a small effect was .26.

The frequencies of hypotheses, cumulative frequencies, cumulative percentages, and central tendency measures are reported in Table 2. The table shows that 13 (17%) of the hypotheses in our sample reached or exceeded the .95 level of power. This finding is superior to other power assessments of management research as shown by Mazen, Graf, et al. (1987); Mazen, Hemmasi, et al. (1987b); and Mone et al. (1996), who reported that 4%, 0%, and 6% of the studies in their samples reached or exceeded the .95 level of power, respectively. Similar contrasts are seen in relation to power assessments of psychology research as reported by Sedlmeier and Gigerenzer (1989) and Rossi (1990), who reported that 2% and 1% of the studies in their samples reached or exceeded the .95 level of power, respectively. What is discouraging is that out of the 77 hypotheses in our sample, 60 tests of hypotheses (78%) did not even have a 50% chance of detecting a small effect size (i.e., power .50). This finding appears to be similar to Mazen, Graf, et al. (1987), who reported that 83% of the studies faced the same odds. The results of other power assessments in relation to this issue were a bit grimmer. Mazen, Hemmasi, et al. (1987); Mone et al. (1996); Sedlmeier and Gigerenzer (1989); and Rossi (1990) reported that 89%, 85%, 92%, and 96% of the studies in their

Table 2
Frequency, Cumulative Frequencies, and
Cumulative Percentage Distribution of Statistical Power (n = 77)

Power	*Frequencies (Number of Hypotheses)*	*Cumulative Frequencies*	*Cumulative Percentages*
.99+	12	12	15.6
.95-.98	1	13	16.9
.90-.94	1	14	18.2
.85-.89	1	15	19.5
.80-.84	0	15	19.5
.70-.79	0	15	19.5
.60-.69	1	16	20.8
.50-.59	1	17	22.1
.40-.49	2	19	24.7
.30-.39	10	29	37.7
.20-.29	10	39	50.6
.10-.19	29	68	88.3
.05-.09	9	77	100

Note. Results are based on power analyses using small effect sizes. Mean power level = .29; median power level = .22.

samples did not have a 50% chance of detecting a small effect size, respectively. Overall, only 4 studies in our sample had sufficient power to draw sound conclusions from the results of their null hypothesis tests.

In addition, our power assessment revealed that out of the 77 null hypotheses in our sample, 46 were affirmed because their respective tests failed to return statistically significant results. For these 46 hypotheses, the average power was .25, which translates into an actual level of β (Type II error rate) equal to .75. As such, researchers' failure to reject the null hypothesis on the basis of insignificant findings might have been done so erroneously 75% of the time (i.e., 35 out of the 46 hypotheses). In comparison to the advocated Type II error rate of .05, this represents an actual error rate that is 15 times greater. This finding supports one of the conclusions reached by Sedlmeier and Gigerenzer (1989) who, based on their power assessment findings, suggested that researchers tend to act as if they believe that mistakenly rejecting the null hypothesis is 11 to 14 times more serious than accepting it. In other words, the power levels of the studies sampled were considerably low, thus translating into fairly large values for β, and in relation to the commonly accepted value of $\alpha = .05$ for hypothesis testing (Cohen, 1977; Sauley & Bedeian, 1989) this leads to a large discrepancy between β and α, namely, 15:1 for this study. It is important to note that this is a comparison between the two types of errors, not between the mean level of power for the sample of articles (i.e., .29) and the desired level of power for the testing of a null hypothesis (i.e., .95). This comparison translates into a desired level of power that is approximately 3.1 times larger than the actual mean level of power for our sample.

Consistent with prior research (Cohen, 1977; Cortina & Dunlap, 1997; Cortina & Folger, 1998), we previously suggested that management researchers formally proposing and testing null hypotheses do not seek to demonstrate a nil effect but rather a relationship that is not practically distinguishable from zero. A review of the data revealed that out of our total sample, 47% of the hypotheses clearly specified null relationships by employing language such as "no significant difference," "a negligible

effect," and "no significant effect." This is not to say that the other 53% of the hypotheses were nil relationships; it merely draws a conclusion based solely on the wording of the hypotheses. In addition, we assessed the procedures researchers employed to test their hypotheses. This assessment revealed that none of the authors tested their hypotheses as nil relationships according to the methods and results sections of the articles. In other words, we did not encounter a single instance in which the researchers sought or discussed an effect size equal to zero.

Discussion

Implications for Interpreting Research

An important finding of this power assessment study was that explicit consideration of the power issue was almost nonexistent among researchers testing null hypotheses. This is consistent with other power assessment studies (e.g., Brewer, 1972; Cohen, 1962; Mazen, Graf, et al., 1987). As previously mentioned, 9.3% of the studies in our sample mentioned statistical power in relation to their testing of null hypotheses. Of this 9.3%, three-fourths actually reported specific power statistics. With such a small number of studies reporting power levels, it appears that considerations of statistical power, at least in print, tended to be underemphasized. It might be possible that such a conclusion is not limited to the journals that served as the sample in this study.

Post hoc power analyses suggested that overall, the studies examined had below desired power. Thus, for those tests reporting nonsignificance of findings, readers must understand that the results may be due to trivial effect sizes or actually insufficient statistical power. As one of the recommendations below suggests, reporting power statistics with the results of statistical tests allows readers to better make the determination that the phenomena of interest are actually absent or that the probability of failing to reject a false null hypothesis when it is actually true is too great. It is particularly important to note that the risk of falsely affirming a null hypothesis could be the product of poor research design. For researchers seeking to reject a null hypothesis, low power may make it more difficult to find support for their predictions. However, in instances in which researchers formally propose and test null hypotheses, low power may make it easier to find a lack of significance. If the consequences of finding false positives are more serious than false negatives, then researchers must be acutely aware of power when testing null hypotheses. Thus, although researches should always strive to guard against Type I and Type II errors, we believe it is vital that researchers testing null hypotheses protect against Type II errors (i.e., supporting the null when it is in fact false).

Recommendations for Future Research

Based on this power assessment, several recommendations can be made to researchers who hypothesize null relationships in their manuscripts (see Table 3). It is important to note that several of our recommendations are not novel, yet it does not minimize their importance. Since Cohen's (1962) work, many recommendations have been offered by a multitude of individuals with regard to improving statistical power

Table 3
Recommendations for Testing Null Hypotheses

Recommendation 1	Researchers should not only calculate but also report power for every standard statistical test. This allows the reader to understand the risks associated with nonsignificant findings.
Recommendation 2	Researchers should establish a β level of .05 or lower to confidently conclude that a trivial effect exists between variables of interest (Cohen, 1977; Sedlmeier & Gigerenzer, 1989; Rossi, 1990). Thus, the risk of a Type II error will at least parallel to generally accepted levels for Type I errors.
Recommendation 3	Researchers should include confidence intervals in their findings to provide further detail that the hypothesized null effect is not trivial due to sampling error (Cortina & Folger, 1998; Nickerson, 2000).
Recommendation 4	To the extent possible, conduct as many experiments or include as many measures (based on relevant operationalizations of variables) for each combination of constructs of interest (Cortina & Folger, 1998). Conducting multiple experiments in a lab setting using different operationalizations (or using several different variables for one construct in a nonlab setting) of the variables of interest provides for greater confidence if a null hypothesis is supported consistently using several different variables in several tests.
Recommendation 5	Incorporate in each experiment and analysis an additional independent variable that is recognized as having a relationship with the dependent variable (Cortina & Folger, 1998). Next, evaluate the relationship in each experiment/analysis to illustrate that the independent variable of interest has a zero or trivially nonzero relationship with the dependent variable while the additional independent variable has a significant nonzero relationship with the dependent variable.

and the proper implementation and use of power analyses. However, the vast majority of these recommendations/insights has been, as previously mentioned, deemphasized by researchers. Refer to the works of Baroudi and Orlikowski (1989); Brewer (1972); Ferguson and Ketchen (1999); Katzer and Sodt (1973); Mazen, Graf, et al. (1987); Mazen, Hemmasi, et al. (1987); Mone et al. (1996); Rossi (1990); and Sedlmeier and Gigerenzer (1989) for empirical support of this argument. Based on the results of this study and the above research, we feel it is imperative to reiterate prior recommendations in the literature regarding power as well as to make recommendations specific to researchers testing null hypotheses.

First, although management scholars for the most part agree that the power of a statistical test is important, what is not universally accepted is that power should be calculated and reported for every standard statistical test. Meehl (1991) suggested that a power analysis should be mandatory in accepting null hypotheses, which would allow individuals who read and use management research to understand the risks associated with nonsignificant findings. In other words, "studies with high power that find insignificance provide strong support for the decision not to reject the null hypothesis, while studies with low power provide little support for either the null or alternative hypotheses" (Baroudi & Orlikowski, 1989, p. 89). Thus, reporting the power of a particular test provides us with at least some interpretation of the results and guards against premature conclusions of "no effect."

Second, when testing a null hypothesis, we recommend that researchers should establish a β level of .05 or lower to confidently conclude that a trivial effect exists between variables of interest and thus not reject the null hypothesis. As previously mentioned, Sedlmeier and Gigerenzer (1989) and Rossi (1990), echoing Cohen

(1977), argued that the reason for setting β at such a low level is that determination of a trivial effect is made when power (1 – β) is set at a high value and the sample size used is large enough so that the risk of Type II error (β) is relatively small and similar to the risk of a Type I error, which is commonly set at the .05 level. In other words, at a β level of .05 or lower (power = .95 or higher), the probability of failing to reject the null hypothesis when the hypothesis is false is low enough to interpret statistical findings with confidence. Thus, we firmly advocate that all tests of null hypotheses should be conducted at power levels of .95 or greater (β = .05 or lower), which parallels common statistical practices associated with α. At such power levels, readers of statistically nonsignificant results will minimize the chances of interpreting the results as also being scientifically insignificant. However, our results suggest that researchers testing formal null hypotheses are far from adopting what we feel is an appropriate level of β. Thus, it appears that the "prejudice" against the null hypothesis, as suggested by Greenwald (1975), still permeates research in the 1990s. Greenwald reached such a conclusion by surveying authors and reviewers and asking them, among other things, to indicate the level of α (β) they would regard as a satisfactory basis for rejecting (accepting) the null hypothesis. His results revealed that the acceptable levels of α and β among these authors and reviewers were .046 and .274, respectively.

It is important to note that the actual choice of β does not have to default to the .05 level (Camilleri, 1962; Sauley & Bedeian, 1989). We feel that this should serve as a minimum level, yet the actual level ought to be dictated by the consequences of committing a Type II error. As Rudner (1953) noted, "How sure we need to be before we accept a hypothesis will depend on how serious a mistake would be" (p. 2), yet in scientific research, the calculation of costs is complicated as compared to cost calculations of practical applications, where the various costs can be expressed in terms of money or time (Camilleri, 1962).

Popper (1959) argued that the strongest confirmatory evidence for a scientific hypothesis is failure of concerted efforts of competent researchers to falsify it. As such, this leads to our third recommendation, the use of confidence intervals, because they represent a good effort on behalf of the researcher to find a lack of effect. Establishing a confidence interval of small range is valuable because confidence intervals take into account sample size and variability (Frick, 1995). As such, the smaller the confidence interval that includes zero, the stronger the evidence for the null hypothesis (Nickerson, 2000). Unfortunately, it is virtually impossible to judge whether the resulting size of the confidence interval is large or small. Frick (1995) suggested that researchers may have a subjective sense of the interval's size based on past experience, the size of typical confidence intervals, and the scale of the measures employed by the researcher. Although the judgment concerning the size of the confidence interval represents a subjective evaluation, when the confidence interval is obviously too large, the experiment is not good evidence for the null hypothesis (Frick, 1995).

Fourth, as advocated by Cortina and Folger (1998), we recommend that researchers conduct as many experiments or include as many measures as possible (based on relevant operationalizations of variables) for each combination of constructs of interest. As such, in a lab setting, a researcher would conduct a separate experiment for each operationalization of the variables of interest. If not in a lab setting, researchers can use multiple operationalizations for the variables of interest. For example, if examining firm performance, several measures such as ROE, ROA, or ROI could be used separately to determine if the null finding remains consistent among all operational-

izations of the variable. The use of multiple operationalizations reduces any criticism that lack of findings is a result of measurement error (Cortina & Folger, 1998).

Fifth, and also advocated by Cortina and Folger (1998), we recommend the inclusion of an additional variable known to have a relationship with the dependent variable. If the additional variable is found to have the expected significant relationship with the dependent variable, then the criticism that measurement error or confounding variables prevented a significant relationship between the independent variable of interest and the dependent variable can be minimized.

Last, we recommend that journal editors and reviewers pay closer attention to the issue of statistical power. It is important for journals to require that authors present evidence of the power of their tests so that readers may make informed decisions about the validity of the results. If statistical power associated with the testing of null hypotheses is less than desirable and as a result produces evidence favoring the null hypothesis, then replication studies with greater power can be performed to assess the validity of previous findings.

Because of the bias against publishing results yielding p values greater than .05 (Greenwald, 1993), results that are actually Type II errors, which could include potentially important and interesting findings, are often buried along with true null effects without ever being published (Nickerson, 2000). An approach to minimize such a risk is to publish the results of all experiments, whether they reach statistical significance or not, and rely on meta-analysis to draw conclusions based on large bodies of results in the aggregate (Hunter & Schmidt, 1990; Nickerson, 2000; Schmidt & Hunter, 1997). The understanding and acceptance of null hypothesis testing is a step in this direction. Thus, we agree with Greenwald (1993), who "hoped that journal editors will base publication decisions on criteria of importance and methodological soundness, uninfluenced by whether a result supports or rejects a null hypothesis" (p. 446).

In summary, we find that the aforementioned recommendations are crucial for researchers to consider when seeking support for their null hypotheses. These steps allow for greater confidence of results because they represent a sound attempt of testing the null hypothesis. This idea parallels what Frick (1995) called the "good-effort criterion." He believed that

> the only way to provide evidence supporting the null hypothesis is to try, but fail, to demonstrate a statistically significant effect. The inclination to accept the null hypothesis thus depends on the quality of the attempt to find an effect. A good effort is an attempt that was likely to find an effect if one existed, and is good evidence for the null hypothesis; a bad effort is not. (p. 135)

Conclusion

Our goal in conducting this study was to determine how well contemporary management research fares on the issue of statistical power and null hypothesis testing in published research. Such an investigation is paramount because "knowledge of the power of a statistical test indicates the likelihood of obtaining a statistically significant result. Presumably, most researchers would not want to conduct an investigation of low statistical power" (Rossi, 1990, p. 646). Furthermore, we wanted to determine whether researchers made power considerations when concluding support for a null hypothesis based on nonsignificant findings. Because presenting statistical null

hypotheses in research is generally frowned upon by a majority of scholars, any time researchers present null hypotheses, they must do their utmost to demonstrate that sufficient statistical power is present to confidently affirm or disaffirm the hypothesis. In addition, it is important to note that relegating the null hypothesis to a secondary status is unwarranted (Atkinson, Furlong, & Wampold, 1982; Cortina & Folger, 1998; Frick, 1995; Greenwald, 1993). As such, we, and others, believe that it is acceptable for a null hypothesis (i.e., a hypothesis of trivial effect or difference) to be offered on its own, if theoretically justified. In other words, "there is no suggestion that a null hypothesis must be used as a comparison value" (Cortina & Folger, 1998, p. 335; Greenwald, 1993).

Although our study is the first power assessment that specifically investigates formal null hypotheses in management research, its findings echo those of previous power assessments in that research suffers from less than desirable levels of statistical power. Specifically, this study suggests that the probability of failing to reject false null hypotheses is greater than what is advocated in the power literature as it pertains to testing formal null hypotheses. It is hoped that the recommendations offered for future management research can prove to be of assistance for others testing null hypotheses.

Notes

1. The general approximations of small effect sizes (ES) for the statistical tests used in this study are the approximations offered by Cohen (1977, 1992). They are *t* test on the means of two independent sample, ES (*d*) = .20; significance of a product moment correlation coefficient, ES (*r*) = .10; difference between correlation coefficients, ES (*q*) = .10; test that a proportion is .50 and the sign test, ES (*g*) = .05; differences between population proportions, ES (*h*) = .20; chi-square tests for goodness of fit and contingency tests, ES (*w*) = .10; analysis of variance and covariance, ES (*f*) = .10; multiple regression and correlation analysis, ES (f^2) = .02.

2. We also evaluated the power of these studies using medium effect sizes. The results revealed that the average statistical power of the 77 null hypothesis tests was .81. Thus, as a group, these studies had a 19% chance of committing a Type II error. In addition, the 45 hypotheses that were affirmed (i.e., the tests failed to return statistically significant results) had an average power equal to .76. Last, only 49.4% of the sample null hypotheses managed to meet the .95 power level.

3. The analyses were conducted in a manner such that the empirical test served as the unit of analysis. Other studies (e.g., Cohen, 1962; Mazen, Graf, et al., 1987; Mazen, Hemmasi, et al., 1987; Mone et al., 1996) have chosen to average the power levels for each study (i.e., no matter how many tests were contained in an article, all articles counted equally in the analysis). As such, we also calculated the average power level for our sample this way to observe any differences in the results. The mean level of power for our sample using an "all articles count equally" approach was .293 versus the mean power level of .290 using the hypothesis as the unit analysis, a difference we feel is insignificant.

References

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology*, *29*, 189-194.

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

Brewer, J. K. (1972). On the power of statistical tests in the *American Educational Research Journal*. *American Educational Research Journal*, *9*, 391-401.

Camilleri, S. F. (1962). Theory, probability, and induction in social research. *American Sociological Review*, *27*, 170-178.

Chase, L. J., & Chase, R. B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology*, *61*, 234-237.

Chase, L. J., & Tucker, R. K. (1975). A power-analytic examination of contemporary communication research. *Speech Monographs*, *42*, 29-41.

Chow, S. L. (1996). *Statistical significance: Rationale, validity, and utility.* Thousand Oaks, CA: Sage.

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

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

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

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161-172.

Cortina, J. M., & Folger, R. G. (1998). When is it acceptable to accept a null hypothesis: No way, Jose? *Organizational Research Methods*, *1*, 334-350.

Cowles, M. P., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, *37*, 553-558.

Ferguson, T. D., & Ketchen, D. J., Jr. (1999). Organizational configurations and performance: The role of statistical power in extant research. *Strategic Management Journal*, *20*, 385-395.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition*, *23*, 132-138.

Gomez-Mejia, L. R., & Balkin, D. B. (1992). Determinants of faculty pay: An agency theory perspective. *Academy of Management Journal*, *35*, 921-955.

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

Greenwald, A. G. (1993). Consequences of prejudice against the null hypothesis. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 419-448). Hillsdale, NJ: Lawrence Erlbaum.

Hambrick, D. C., Cho, T. S., & Chen, M. J. (1996). The influence of top management team heterogeneity on firms' competitive moves. *Administrative Science Quarterly*, *41*, 659-684.

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods*, *5*, 315-332.

Hubbard, R., Vetter, D. E., & Little, E. L. (1998). Replication in strategic management: Scientific testing for validity, generalizability, and usefulness. *Strategic Management Journal*, *19*, 243-254.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Katzer, J., & Sodt, J. (1973). An analysis of the use of statistical testing in communication research. *Journal of Communication*, *23*, 251-265.

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.

Lawless, M. W., & Anderson, P. C. (1996). Generational technological change: Effects of innovation and local rivalry on performance. *Academy of Management Journal*, *39*, 1185-1217.

Mazen, A. M., Graf, L. A., Kellogg, C. E., & Hemmasi, M. (1987). Statistical power in contemporary management research. *Academy of Management Journal*, *30*, 369-380.

Mazen, A. M., Hemmasi, M., & Lewis, M. F. (1987). Assessment of statistical power in contemporary strategy research. *Strategic Management Journal*, *8*, 403-410.

Meehl, P. E. (1991). Why summaries of research on psychological theories are often uninterpretable. In R. E. Snow & D. E. Wilet (Eds.), *Improving inquiry in social science* (pp. 13-59). Hillsdale, NJ: Lawrence Erlbaum.

Mone, M. A., Mueller, G. C., & Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology*, *49*, 103-120.

Myers, B. L., & Melcher, A. J. (1969). On the choice of risk levels in managerial decision-making. *Management Science*, *16*, B31-B39.

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

Popper, K. (1959). *The logic of scientific discovery.* New York: Basic Books.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*, 646-656.

Rudner, R. (1953). The scientist *qua* scientist makes value judgments. *Philosophy of Science*, *20*, 1-6.

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

Sawyer, A. G., & Ball, D. (1981). Statistical power and effect size in marketing research. *Journal of Marketing Research*, *18*, 275-290.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Hillsdale, NJ: Lawrence Erlbaum.

Schwenk, C. R., & Dalton, D. R. (1991). The changing shape of strategic management research. In P. Shrivastava, A. Huff, & J. Dutton (Eds.), *Advances in strategic management* (Vol. 7, pp. 277-300). Greenwich, CT: JAI.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309-316.

Sharplin, A. D., & Mabry, R. H. (1985). The relative importance of journals used in management research: An alternative ranking. *Human Relations*, *38*, 139-149.

Wiersema, M. F. (1992). Strategic consequences of executive succession within diversified firms. *Journal of Management Studies*, *29*, 73-94.

Luke H. Cashen is a doctoral candidate in the Rucks Department of Management at Louisiana State University. His research interests include corporate diversification, portfolio restructuring, and governance.

Scott W. Geiger, PhD, is an assistant professor of management at the University of South Florida St. Petersburg. His research interests include corporate diversification, technology and innovation, and business ethics. His research has appeared in journals such as Organizational Behavior and Human Decision Processes*, the* Journal of Business Ethics*, and the* Journal of Business Research*.*