Assessment Of Statistical Power In Contemporary Accounting Information Systems Research

By: **Dwayne N. McSwain**

Abstract

The purpose of this study is to provide a current, representative assessment of statistical power in accounting information systems (AIS) research. This study empirically investigates whether the statistical power of extant AIS research has been strong enough to detect important relationships that may exist. A power analysis of 45 articles from the most recent, complete five years (1996-2000) of Journal of Information Systems and Journal of Management Information Systems shows that, on the average, 56 percent of empirical studies do not have high power levels. This suggests that, on average across all effect sizes, more than half the time AIS researchers risk not being able to detect significant effects when, in fact, they exist. This risk increases greatly as the effect size decreases. Current findings suggest the need for more statistical power planning in AIS research designs. Statistical power is important to AIS research because it increases the probability of making correct decisions about empirical studies. Without adequate statistical power, AIS research may fail to identify statistically significant results and viable research streams might be abandoned prematurely. Statistical power will also become increasingly important as empirical studies in AIS study relatively smaller effects.

McSwain, D. (2004). "Assessment of Statistical Power in Contemporary Accounting Information Systems Research," *Journal of Accounting and Finance Research,* Winter II, 2004. 100-108. NC Docks permission to re-print granted by author.

ASSESSMENT OF STATISTICAL POWER IN CONTEMPORARY ACCOUNTING INFORMATION SYSTEMS RESEARCH

Dwayne N. McSwain, Middle Tennessee State University

ABSTRACT

The purpose of this study is to provide a current, representative assessment of statistical power in accounting information systems (AIS) research. This study empirically investigates whether the statistical power of extant AIS research has been strong enough to detect important relationships that may exist. A power analysis of 45 articles from the most recent, complete five years (1996–2000) of *Journal of Information Systems* and *Journal of Management Information Systems* shows that, on the average, 56 percent of empirical studies do not have high power levels. This suggests that, on average across all effect sizes, more than half the time AIS researchers risk not being able to detect significant effects when, in fact, they exist. This risk increases greatly as the effect size decreases. Current findings suggest the need for more statistical power planning in AIS research designs. Statistical power is important to AIS research because it increases the probability of making correct decisions about empirical studies. Without adequate statistical power, AIS research may fail to identify statistically significant results and viable research streams might be abandoned prematurely. Statistical power will also become increasingly important as empirical studies in AIS study relatively smaller effects.

INTRODUCTION

Of the challenges and opportunities facing researchers, perhaps the most pressing is the need to conduct and publish quality research. Similar to research in other social sciences, the quest to conduct and publish quality research in accounting has placed a major emphasis on statistical significance when testing hypotheses (Lindsay 1993). One unintended consequence of this focus on statistical significance has been a lack of acknowledgment by researchers of statistical power and effect size (Mazen et al. 1987a, 1987b; Cohen 1977, 1988; Baroudi and Orlikowski 1989; Lindsay 1993; Mone et al. 1996). Although the formal concept of statistical power has been around since 1933 (Cohen 1988), its usage has not been prominent in extant accounting research (Lindsay 1993).

There have been many studies of the use of statistical power in other research fields (Mazen et al. 1987a, 1987b; Baroudi and Orlikowski 1989; Lindsay 1993; Mone et al. 1996); however, only recently has the investigation into the impact of statistical power on accounting research begun (Lindsay 1993). Understanding statistical power is important to accounting researchers because studies with low levels of statistical power often result in inconclusive findings, even if the study is well designed (Semon 1990). Researchers invest time, effort, and money into studies and want to detect significant findings, if they exist (Baroudi and Orlikowski 1989). Without adequate statistical power, accounting research may fail to identify statistically significant results. Thus, sufficient statistical power may help prevent the premature abandonment of viable research streams (Ferguson and Ketchen 1999).

The purpose of this study is to provide a current, representative assessment of statistical power in accounting information

systems (AIS) research. This study will empirically investigate whether the statistical power of extant AIS research has been strong enough to detect important relationships that may exist.

This paper is organized as follows. The first section describes the components of statistical power and the relationships among these components. In the next section, an assessment of current statistical power levels in leading AIS journals is described. The third section presents the results of this investigation. The implications of these results are discussed in the fourth section. This paper concludes with a summary of the findings, the limitations of this study, and the possibilities for future research.

COMPONENTS OF STATISTICAL POWER

By definition, the probability of committing a Type I error is alpha (α) (Larsen and Marx 1986). Type I errors occur when researchers mistakenly reject the null hypothesis (Mone et al. 1996). But what about a Type II error (β)? What are the chances that the observed sample will take on values that will "deceive" the researcher into failing to reject the null hypothesis when it should be rejected (Larsen and Marx 1986)? This question is addressed by the concept of power.

Vogt (1999) defines statistical power as a gauge of the sensitivity of a statistical test to detect significant effects. In other words, statistical power is the probability that an empirical test will detect a relationship when a relationship, in fact, exists (Vogt 1999). Statistical power is determined by three interacting components: effect size (δ), significance level (α), and sample size (*n*) (Cohen 1977, 1988). Effect size is the magnitude of findings (e.g., a correlation between two variables, the difference between two means), alpha is the level of risk of rejecting a true null hypothesis, and sample size is the number of observations used in a test (Lindsay 1993). The power of a statistical test is calculated by subtracting the probability of a Type II error from 1 ($1 - \beta$). Power can range from a minimum of 0 to a maximum of 1, with 80 percent often considered an acceptable level (Vogt 1999). Thus, a common acceptable probability of correctly rejecting the null hypothesis when it should be rejected is 80 percent.

The relationships among effect size, alpha, sample size, and power are quite complicated, and a number of sources of guidance are available (Hair et al. 1998). However, power and its three components are so closely related that when any of the three are known, the fourth can be calculated easily (Cohen 1988). A brief introduction of each component follows.

Effect Size (δ). According to Cohen (1977, p 9), effect size is "the degree to which a phenomenon is present in the population," or "the degree to which the null hypothesis is false." *Ceteris paribus* (α and *n*), power increases as effect size increases. Thus, the larger the effect size of the phenomenon being studied, the greater the probability the researcher has for rejecting the null (Lindsay 1993).

Of the three power components, effect size is probably the most important determinant of statistical power and the least understood (Baroudi and Orlikowski 1989). Additionally, effect size is also the most difficult parameter to estimate (Mazen et al. 1987b). According to Pedhazur and Schmelkin (1991), the ambiguity surrounding the meaning of effect size contributes to the problem of determining effect size, because the term is often used interchangeably to refer to magnitude, importance, or meaningfulness. Difficulties in defining magnitude, importance, and meaningfulness further confound the problem of effect size determination (Pedhazur and Schmelkin 1991). Pedhazur and Schmelkin (1991) postulate that knowledge of the subject matter, the properties of the measures used, and hard thinking are the most important ingredients for making informed decisions about effect size.

Cohen (1977) posits conventional, operational definitions for small, medium, and large effect sizes for different statistical tests. Although Cohen (1977) himself admits that these conventional definitions of effect size are somewhat arbitrary, they do provide a standard index and sufficient guidance for researchers to use in determining effect size. According to Lindsay (1993), Cohen defines small, medium, and large effect sizes as being approximately equal to an *r* of .10, .30, and .50, respectively.

Alpha (α). Cohen (1977) posits that statistical power is an increasing function of alpha, holding other things equal (*n* and δ). A small alpha level results in a relatively small power value (Cohen 1977). As α decreases, it becomes less probable that the null hypothesis will be rejected because it requires an increasingly larger δ (Lindsay 1993). Because there is also an inverse relationship between α and β, an increase in α decreases β and therefore increases power (Mazen et al. 1987a). Power not only increases with larger α, but also with directional hypothesis tests (Baroudi and Orlikowski 1989). However, Cohen (1977, 1988) has expressed serious reservations about using directional testing in behavioral science research in all but relatively limited circumstances. The reason for Cohen's (1977, 1988) reservations is that most behavioral science studies are concerned with proof that some phenomenon exists or does not exist. In this case, the researcher is normally comparing some parameter (e.g., mean, proportion, correlation) for two populations and no direction of the difference is specified, because either direction from the null hypothesis constitutes evidence against the null (Cohen 1977, 1988).

Sample Size (n). Generally, the larger the sample size, the smaller the error and the more accurate the measure of the phenomenon under investigation (Mazen et al. 1987a). In other words, the precision of sample estimates increases as *n* increases (Cohen 1977). The larger the *n*, *ceteris paribus* (α and δ), the greater the probability of rejecting a false null hypothesis (Baroudi and Orlikowski 1989). Although increasing sample size is a simple concept, factors beyond the control of researchers often limit the size of samples included in their studies. For example, the following are all possibilities for constrained sample sizes: a limited amount of money, a limited amount of time, or a limited number of qualified participants (participants possessing or not possessing some characteristic).

Power (1 – β). Statistical power can be increased by increasing any one of its components (Cohen 1988). The larger the effect size, *ceteris paribus*, the more likely it is that a statistical test will detect the effect. Because an inverse relationship exists between α and β, statistical power can be increased by increasing α. Increasing sample size can also increase power, because statistical power increases monotonically with increases in sample size (Lindsay 1993).

Although Cohen (1977, 1988) recommends statistical power be assessed *a priori*, Baroudi and Orlikowski (1989) suggests power analysis is useful before, during, and after the research process. In each case, power analysis allows researchers to take appropriate action and get the most out of their study (Baroudi and Orlikowski 1989). When statistical power is inadequate, researchers may not be able to detect meaningful differences or effects (Lindsay 1993). Thus, time and effort may be wasted or a research stream may be abandoned prematurely. On the other hand, when statistical power is excessive (usually the result of large sample sizes), the test may be oversensitive and small effects may appear significant. Although such findings are statistically significant, they may not have practical significance (Hair et al. 1998).

METHOD

Although there are only two leading journals devoted to AIS research (*Journal of Information Systems* and *International Journal of Accounting Information Systems*), there are several leading management

TABLE 1
Distribution of AIS Studies Employing
Statistical Inference Testing: 1996-2000

Journal	Number	Percent
Journal of Information Systems	18	40%
Journal of Management Information Systems	27	60%
Total	45	100%

information systems journals that also publish AIS research, including *MIS Quarterly, Information Systems Research,* and *Journal of Management Information Systems.* To assess the current level of statistical power in AIS research, this study examined empirical articles in two of the leading journals, *Journal of Information Systems (JIS)* and *Journal of Management Information Systems (JMIS).*

The unit of analysis was the journal article. For the most recent complete five years (1996–2000), all articles from *JIS* were selected and ten articles per volume from *JMIS* were selected randomly. This resulted in an initial sample of 89 articles, and represents 100% and 26% of all articles published over these five years for *JIS* and *JMIS*, respectively. The initial sample was screened for articles containing ANOVA/ ANCOVA, *t* test, multiple regression, correlation, and chi-square statistical tests. According to Cohen (1977, 1988), these statistical tests lend themselves to power analyses. This resulted in a usable sample of 45 articles, as shown in Table 1.

As discussed above, the power of a statistical test is a function of three parameters: alpha, effect size, and sample size. Because the sample size is given in each particular article, only the first two parameters need to be identified for the purposes of the current study. Following past power research (Mazen et al. 1987a, 1987b; Cohen 1977, 1988; Baroudi and Orlikowski 1989; Lindsay 1993; Mone et al. 1996), this study assumed non-directional (two-tailed) tests with an alpha level of 0.05. Cohen's (1977) operational definitions of small, medium, and large effect sizes were used for each type of statistic analyzed in this study. These three parameters were used to determine power from Cohen's (1977) power analysis tables for each full-model ANOVA/ANCOVA, *t* test, multiple regression, correlation, and chi-square statistical test reported in the articles. As recommended by Cohen (1977, 1988), linear interpolation was used to determine power values for parameters that fell between table values.

Only tests of the major hypotheses are included in the analyses. Power calculations were not made for reliability tests, tests of statistical assumptions, or manipulation checks. An average power figure was calculated for each article, and each article carried equal weight in the analysis. Power was calculated for a total of 175 tests.

RESULTS

The frequency and ascending cumulative percentage distributions of the average power of AIS studies to detect small, medium, and large effects and the related descriptive statistics are presented in Table 2. Average power across type of statistical test was 0.22 for small, 0.74 for medium, and 0.92 for large effect sizes. More importantly, none of the

TABLE 2

Frequency and Ascending Cumulative Percentage Distribution of Statistical Power*

	Small Effect Size		Medium Effect Size		Large Effect Size	
Power	Frequency	Cumulative Percent	Frequency	Cumulative Percent	Frequency	Cumulative Percent
.99+	0	0	9	20	26	58
.95—.98	0	0	7	36	6	71
.90—.94	0	0	4	45	**4**	**80**
.80—.89	0	0	2	49	2	84
.70—.79	0	0	**8**	**67**	4	93
.60—.69	1	2	2	71	0	93
.50—.59	3	9	3	78	1	96
.40—.49	2	13	4	87	1	98
.30—.39	7	29	3	94	1	100
.20—.29	**7**	**45**	2	98	0	100
.10—.19	16	80	1	100	0	100
.01—.09	9	100	0	100	0	100
n	45		45		45	
Mean	0.22		0.74		0.92	
sd	0.14		0.25		0.15	

Note: Power assessments assumed .05 alpha and two-tailed tests.

The corresponding power level range for each mean effect size is shown in bold.

*Frequencies above the dotted line achieved the conventional power of .80.

studies reviewed in the current study met the conventional power level of 0.80 for detecting small effect sizes, and only 49 percent of these studies met that standard for detecting medium effect sizes. Although 84 percent of the studies exhibited sufficient power for detecting large effect sizes, large effect sizes are inherently much easier to detect (Cohen 1988). This fact, coupled with the weak power levels for detecting small and medium effects, may imply that this outcome is generally not preplanned.

Across all effect sizes, the recommended power level was achieved in less than half of the articles (44%). Put differently, 56 percent of these studies had low power levels. This suggests that, on average across all effect sizes, more than half the time AIS researchers risk not being able to detect significant effects when, in fact, they exist. This risk increases greatly as the effect size decreases.

Table 3 contains descriptive information concerning power levels for each type of statistic and effect size. To determine which type of statistics had significantly different power levels at $\alpha = 0.05$, Tukey's studentized range (HSD) tests (an ANOVA procedure) were conducted. The purpose of this analysis is to assess whether there are significant differences in the actual power levels of various types of statistical tests commonly used in leading AIS journals.

Analyses of variance (ANOVAs) of power level by type of statistic used in each of the 175 analyses revealed significant main effect differences for small (df = 4, 170; $F = 3.45$; $p < .0097$), medium (df = 4, 170; $F = 9.13$; $p < .0001$), and large (df = 4, 170; $F = 4.84$; $p < .0010$) effect sizes. For small effects, only the difference between the power of correlation and chi-square is significantly different, with the power level of correlation being greater. For medium effects, both the power of correlation and multiple regression are significantly greater than the power of ANOVA/ANCOVA, chi-square, and *t* test.

TABLE 3

Means and Standard Deviations of Statistical Power for Small, Medium, and Large Effect Sizes by Type of Statistic

Type of Statistic	*N* of Analyses	Percent of Tests	Effect Size		
			Small	Medium	Large
ANOVA/ANCOVA	64	37%	.23	.74	.95
			(.18)	(.22)	(.08)
t Test	26	15%	.17	.60	.85
			(.09)	(.27)	(.16)
Multiple Regression	26	15%	.26	.89	.96
			(.13)	(.20)	(.11)
Correlation	29	16%	.28	.89	.98
			(.13)	(.18)	(.11)
Chi-Square	30	17%	.17	.66	.92
			(.13)	(.29)	(.17)
Total	175	100%			

Note: Standard deviations are in parentheses. Power assessments assumed .05 alpha and two-tailed tests.

Also for medium effects, the power of ANOVA/ANCOVA is significantly greater than chi-square and *t* test. For large effects, the power of correlation, multiple regression, and ANOVA/ANCOVA are significantly greater than the power of *t* test.

Table 4 contains descriptive information concerning power levels for each journal and effect size. Tukey's studentized range (HSD) tests were conducted to determine which journal had significantly different power levels at $\alpha = 0.05$.

ANOVAs of power level by journal only revealed significant main effect differences for large (df = 1, 173; $F = 8.53$; $p < .0040$) effect sizes. Small (df = 1, 173; $F = .75$; $p < .3866$) and medium (df = 1, 173; $F = 3.69$; $p < .0565$) effect sizes did not reveal a significant difference of power by journal. For both small and medium effects, the power levels are not significantly different. For large effects, the power level of *JIS* is significantly greater than that found in *JMIS*.

Descriptive information concerning power levels for each year (1996–2000) of publication and effect size is presented in Table 5. Tukey's studentized range (HSD) tests were conducted to determine which years had significantly different power levels at $\alpha = 0.05$.

ANOVAs of power level by year of publication revealed significant main effect differences for small (df = 4, 170; $F = 4.65$; $p < .0014$), medium (df = 4, 170; $F = 5.32$; $p < .0005$), and large (df = 4, 170; $F = 2.63$; $p < .0361$) effect sizes. For both small and medium effects, the power levels of 2000 and 1996 are significantly greater than the power of 1998 tests. Also for medium effects, the power of 2000 statistics is significantly greater than 1999. For large effects, only the difference between the power of 2000 and 1998 is significantly different, with the power of 2000 being greater.

DISCUSSION

Statistical power should be a topic of interest to any researcher using statistical inference testing (Baroudi and Orlikowski 1989). According to Lindsay (1993), the formal inclusion of statistical power permits scientific significance to be attached to failures to reject the null, allows researchers to remedy designs that have too little power, and helps prevent trivial results from being declared significant.

TABLE 4

Means and Standard Deviations of Statistical Power for Small, Medium, and Large Effect Sizes by Journal

Journal	*N* of Analyses	Percent of Tests	Effect Size		
			Small	Medium	Large
Journal of Information Systems	80	46%	.23	.79	.97
			(.16)	(.21)	(.07)
Journal of Management Information Systems	95	54%	.21	.72	.91
			(.14)	(.28)	(.15)
Total	175	100%			

Note: Standard deviations are in parentheses. Power assessments assumed .05 alpha and two-tailed tests.

However, current findings suggest that AIS researchers do not always conduct research studies with enough power to detect small and medium effect sizes. Overall, the average power of ANOVA/ANCOVA, *t* test, multiple regression, correlation, and chi-square statistics reported in *JIS* and *JMIS* to detect small, medium, and large effect sizes was 0.22, 0.74, and 0.92, respectively (from Table 2). Low power levels for tests with medium and small effect sizes are a reason for concern.

Across all effect sizes, the recommended power level was achieved in less than half of the articles (44%). This means that a researcher who investigated unknown differences between means had, on the average, more than a 50-50 chance to erroneously sustain the null hypothesis. It would seem foolish to conduct studies in which the probability of failure is greater than half at the outset. Clearly, current findings suggest the need for more statistical power planning in AIS research designs.

This study also examined differences in power levels by type of statistic, journal, and year. Regarding type of statistic, *t* test was found to have significantly smaller power levels than other types of statistics. This suggests that AIS researchers who use *t* test may want to use larger sample sizes to avoid a greater risk of overlooking significant effects. For small and medium effects, there was no significant difference between *JIS* and *JMIS* power levels. *JIS* had a significantly higher power level for large effects. Although articles published in 2000 contained the highest power levels for all effect sizes, this difference may not be practically significant because of the relatively larger number of analyses in the year 2000. On the other hand, this result could be indicative of an increase in power awareness by *JIS* researchers, editors, and reviewers, but there was no consistent trend.

Verma and Goodale (1995) posit that young disciplines typically start by studying large effects, and, as these disciplines mature, more and more research is undertaken that explores smaller effects. Pedhazur and Schmelkin (1991) point out, however, that large effect sizes are not generally encountered in sociobehavioral research fields. This lack of large effect size could be a problem for AIS research, because current findings indicate that AIS research is not very powerful. Recall that statistical power generally suffered in current AIS studies as effect size decreased. Hence, the researcher believes that power levels will become increasingly important in future empirical AIS research.

Although the statistical power of a test can be improved by increasing one of its three components (α, δ, and n), effect size can be considered to be more-or-less fixed and acceptable α levels are set by norms of the field of study (Verma and Goodale 1995).

TABLE 5

Means and Standard Deviations of Statistical Power for Small, Medium, and Large Effect Sizes by Year of Publication

Year of Publication	*N* of Analyses	Percent of Tests	Effect Size		
			Small	Medium	Large
1996	41	24%	.26	.80	.94
			(.16)	(.25)	(.13)
1997	23	13%	.20	.69	.90
			(.15)	(.27)	(.20)
1998	25	14%	.14	.61	.89
			(.07)	(.24)	(.14)
1999	33	19%	.18	.70	.94
			(.12)	(.25)	(.07)
2000	53	30%	.27	.85	.97
			(.17)	(.21)	(.09)
Total	175	100%			

Note: Standard deviations are in parentheses. Power assessments assumed .05 alpha and two-tailed tests.

Thus, researchers are left with sample size as the controlling factor for generating acceptable power levels (Verma and Goodale 1995). This information and *a priori* assumptions allow researchers to design more sensitive, powerful, and economical studies.

CONCLUSION

The conclusion of this paper is that AIS researchers need to consider conducting more statistical power planning. Generally, the average study in this analysis did not achieve the recommended 0.80 level of statistical power when the effect size was small or medium. Only AIS research for large effect size was powerful enough to detect the phenomena under analysis. Thus, one can conclude that AIS research is, on the average, statistically powerful only if the effect size is large. Statistical power is important to AIS research because it increases the probability of making correct decisions about empirical studies.

This study is not without limitation. First, this analysis dealt entirely with published research. This method ignores unpublished effect sizes, which may or may not be smaller than that of published studies. Second, this analysis gave the same weight to each article, and, in general, it would seem that good studies are more powerful than bad ones. Finally, this study examined only two leading AIS journals. This limits the generalizability of the current findings.

Future research in this area should consider including more AIS journals. Future research could also compare the power level of AIS research to that of different areas of the social sciences. Finally, a more accurate comparison may be made by holding time constant, to account for the advancement of methodological rigor over time.

REFERENCES

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

Cohen, J. 1977. *Statistical Power Analysis for the Behavioral Sciences*, Revised Ed. New York, NY: Academic Press.

———. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed. Hillsdale, NJ: Erlbaum.

Ferguson, T. D., and D. J. Ketchen, Jr. 1999. Organizational configurations and

performance: The role of statistical power in extant research. *Strategic Management Journal* 20 (April): 385-395.

Hair, J. F., R. E. Anderson, R. L. Tatham, and W. C. Black. 1998. *Multivariate Data Analysis*, 5th Ed. Upper Saddle River, NJ: Prentice-Hall.

Larsen, R. J., and M. L. Marx. 1986. *An Introduction to Mathematical Statistics and Its Applications*, 2nd Ed. Englewood Cliffs, NJ: Prentice-Hall.

Lindsay, R. M. 1993. Incorporating statistical power into the test of significance procedure: A methodological and empirical inquiry. *Behavioral Research in Accounting* 5: 211-236.

Mazen, A. M. M., M. Hemmasi, and M. F. Lewis. 1987a. Assessment of statistical power in contemporary strategy research. *Strategic Management Journal* 8 (July/August): 403-410.

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

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

Pedhazur, E. J., and L. P. Schmelkin. 1991. *Measurement, Design, and Analysis: An Integrated Approach*. Hillsdale, NJ: Erlbaum.

Semon, T. 1990. Keep ignoring statistical power at your own risk. *Marketing News* 24 (September): 21.

Verma, R., and J. C. Goodale. 1995. Statistical power in operations management research. *Journal of Operations Management* 13 (August): 135-152.

Vogt, W. P. 1999. *Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences*, 2nd Ed. Thousand Oaks, CA: Sage.