Methodological Note

Statistical Power in Operations Management Research

Rohit Verma

David Eccles School of Business, University of Utah, Salt Lake City, UT 84112, USA

John C. Goodale

David Eccles School of Business, University of Utah, Salt Lake City, UT 84112, USA

This paper discusses the need and importance of statistical power analysis in field-based empirical research in Production and Operations Management (POM) and related disciplines. The concept of statistical power analysis is explained in detail and its relevance in designing and conducting empirical experiments is discussed. Statistical power reflects the degree to which differences in sample data in a statistical test can be detected. A high power is required to reduce the probability of failing to detect an effect when it is present. This paper also examines the relationship between statistical power, significance level, sample size and effect size. A probability tree analysis further explains the importance of statistical power by showing the relationship between Type 11 errors and the probability of making wrong decisions in statistical analysis. A power analysis of 28 articles (524 statistical tests) in the Journal of Operations Management and in Decision Sciences shows that 60% of empirical studies do not have high power levels. This means that several of these tests will have a low degree of repeatability. This and other similar issues involving statistical power will become increasingly important as empirical studies in POM study relatively smaller effects.

Introduction

A number of articles published over the last ten years in leading management journals have argued that research in Production and Operations Management (POM) is often not very useful to operations managers and lags practice because it does not take into account the applied nature of operations management (Adam and Swamidass, 1989; Amoako-Gyampah and Meredith, 1989; Chase, 1980). It has also been pointed out that the nature and scope of POM as a discipline of science has changed dramatically over the last decade or so (Chase and Prentis, 1987). Some examples of the changing nature and complexity of operations in the 1990s are the productivity crisis in western economies, issues in quality management, issues in service operations management, and international issues in POM. In recent responses to the

above concerns, several authors have stressed the need and importance of field-based empirical research (Flynn et al., 1990; Meredith et al., 1989; Swamidass, 1991). In fact, the editors of the *Journal of Operations Management* have explicitly encouraged POM researchers to submit field-based empirical research papers (Ebert, 1990).

Information derived from field-based empirical research provides a systematic knowledge of actual practice in manufacturing and service operations, which can be used to identify relevant research problems and to provide a baseline for longitudinal studies. Data collected from empirical research can be used to validate the findings of simulation and mathematical programming-based research. Empirical research is useful in expanding the scope of research in several fuzzy and under-researched topics, e.g., operations strategy, international issues, etc. In addition, it can add new perspectives to over-researched topics, e,g., operations planning and control, location planning, etc. A/so, empirical studies open up opportunities for conducting interdisciplinary research. Finally, empirical research can be used to build and verify theories.

Recently, Flynn et al. (1990) published a detailed review paper which explains how empirical research can be done in Operations Management and related areas. This paper describes a systematic approach to empirical research drawn from Organizational Behavior, Psychology, Marketing, Anthropology, Sociology and other social sciences. In brief, the paper elaborates on the different stages in conducting empirical research, theory building versus theory verification, research designs, etc. Swamidass (1991) has identified empirical theory building as a methodological void in Operations Management. He addresses two questions in his article - What is empirical science? and How can empirical theory building be nurtured in POM?

The above articles are very valuable resources for all empirical researchers in POM and related disciplines. However, both fall short in communicating the importance of the *sensitivity* of an empirical experiment. An *empirical experiment* can be broadly defined as a study based on empirical data that intends to test hypotheses, identify relationships, and/or infer causation. The issue of sensitivity of an empirical experiment has received considerable attention in other, more mature, social sciences (Chase and Chase, 1976; Cohen, 1962, 1977, 1988, 1992; Keppel, 1991).

A quantitative index of the sensitivity of an experiment is measured by its *statistical power* (Cohen, 1977, 1988). *The statistical power represents the probability of rejecting the null hypothesis when the alternative hypothesis is true.* Power can be interprefer as the probability of making a correct decision when the null hypothesis is false. Therefore, statistical power represents the probability that a statistical test of the null hypothesis will conclude that the phenomenon under study exists. In other words, power reflects the degree to which differences in sample data in a statistical test can be detected. A high power is required to reduce the probability of failing to detect an effect when it is present.

The purpose of this paper is to address the issue of sensitivity of an empirical experiment. Specifically, we will (1) present the concept and discuss the need and relevance of statistical power analysis, (2) present a summary of power calculations from a number of articles in the *Journal of Operations Management* and in *Decision Sciences,* and (3) discuss strategic implications of power analysis. These issues are of interest to all empirical researchers in POM and related areas where empirical studies are just beginning to make inroads. In addition, the information presented here is fairly general in nature and will be useful to all researchers who write or study empirical research articles.

Any science is known to be built on a large body of facts and information. Statistics is a common tool used in most social sciences to uncover facts and to enhance the understanding of a particular phenomenon. A statistical analysis provides a way of determining the *repeatability* of any differences observed in an empirical study. A well designed study also permits the inference of *causation.* In most of the social sciences, POM included, researchers are not primarily interested in just describing summary statistics of the sample. In general, the goal is to make *inferences* about the whole population. Therefore, an empirical experiment begins by formulating a number of research hypotheses. These hypotheses may represent deductions or derivations from a formal theoretical explanation of a phenomenon of interest, or they may simply represent speculations concerning the phenomenon. Research hypotheses are the questions which researchers hope to answer by conducting an empirical study. This article emphasizes the importance of power analysis in designing and conducting an empirical experiment and shows that results with low statistical power have a low degree of repeatability. The power calculations, presented later in this paper, reveal that some of the published research articles in POM and related areas are very deficient in statistical power. Hence, this is an appropriate time to understand the fundamentals of power analysis.

In general, an empirical research project involves a lot of time, effort and money. Hence, it becomes important to design and conduct sensitive projects - those that are sufficiently powerful to detect any differences that might be present in the population. An important implication of power analysis is in the design of empirical experiments. Researchers can conduct a power analysis before the actual experiment and avoid undertaking a study which is expected to have low power. Hence, power analysis has important implications for planning of an empirical research project. Another reason for conducting a power analysis is to avoid wasting resources by performing studies with too much power. Studies with excessive power often have much larger than necessary sample sizes, and therefore waste time and money.

Post-hoe power analysis of an empirical study might help researchers in arriving at a conclusion. For instance, a post-hoe power analysis might add insight to a nonsignificant F-test. It can suggest if the nonsignificant F occurs because there is *no actual effect* (the phenomenon of interest was not present) or because the power of the study was insufficient to detect the effect (Keppel, 1991).

Statistical power of an experiment also reflects the degree to which the results might be duplicated when an experiment is repeated. If the power of an experiment is 0.50, it means that 50% of the repeated empirical experiments will not yield a significant result, even though the phenomenon exists (Keppei, 1991).

The Fundamentals of Statistical Power Analysis

This section of the paper reviews the factors that affect the statistical power of an empirical experiment. We explain the relationship between these factors and suggest the conditions under which high power can be achieved. *Statistical power analysis* describes the relationship among the four variables involved in statistical inference: the *significance level* (α)*, sample size* (N)*, effect size* (ES) and *statistical power.* For any statistical model, these variables are such that each is a function of the other three (Cohen, 1977, 1988). Therefore,

$$Statistical\ power = f(\alpha, N, ES) \quad (1)$$

The following sections discuss how statistical power is related to the other three parameters in Eq. (1) (significance level, sample size and effect size). The relationship between statistical power and the probability of making a wrong decision in statistical tests is also discussed.

Statistical Power and Significance Level

In general, an empirical experiment begins with specification of statistical hypotheses (null and alternative hypotheses) which consist of a set of precise statements about the parameters of interest. The statistical hypothesis under test is known as the *null hypothesis* (H_0) and is analogous to saying that the phenomenon of interest does not exist. The *alternative hypothesis* (H_a) specifies values for the parameter that are *incompatible* with the null hypothesis. In other words, the alternative hypothesis states that the phenomenon under study exists. Next, relevant data are collected and then either the null hypothesis is rejected or it is retained. However, the crux of the problem is the fact that a portion of differences observed among the experimental conditions are random variations.

The procedure followed in hypothesis testing does not guarantee that a correct inference will always be drawn. On the contrary, there will always be a probability of making an error. Depending on the actual state of the populations under study, the researcher can make either a *Type I error* or a *Type 11 error.* A Type I error occurs when H_0 is rejected when it is true. A Type II error occurs when H_0 is not rejected when H_a is true. The probabilities of a Type I error and Type II error are known as α and β, respectively. The α is also known as the *significance level. Statistical power* is defined according to the following equation (Cohen, 1977, 1988):

$$Statistical\ power = 1 - \beta \quad (2)$$

Table 1 presents the possible conclusions and errors in a statistical test and their

relation to each other. Consider, $\alpha = 0.05$ and $\beta = 0.20$ for a study in which H_0 is rejected. Even though H_0 is rejected, there is a finite probability of making an incorrect decision (significance level $= \alpha = 0.05$) and a finite probability of making the correct decision (statistical power $= 1 - \beta = 0.80$), because it is impossible to know if H_0 is indeed false. If the researcher did know that H_0 is false, then there would have been no need for the hypothesis test. This simple example shows that the chances of making a correct decision in hypothesis testing increase with higher statistical power.

Decision	Reality	
	H_0 True H_a False	H_0 False H_a True
Reject H_0 and Accept H_a	Decision: Incorrect Error: Type I Probability [a]: α Example: $\alpha = 0.05$	Decision: Correct Error: None Probability [b]: $1 - \beta$ Example: $1 - \beta = 0.80$
Retain H_0 and Do not accept H_a	Decision: Correct Error: None Probability [a]: $1 - \alpha$ Example: $1 - \alpha = 0.95$	Decision: Incorrect Error: Type II Probability: β Example: $\beta = 0.20$

[a] Probability of Type I error α referred to as the significance level.
[b] Probability $1 - \beta$ is referred to as the power of the study.

Table 1 Illustration of Type I and Type II errors in hypothesis testing

Consider an empirical study in which a researcher wants to test if two population means are equal to each other. The researcher can choose to perform an F-test (ANOVA). This means that the researcher must decide if the F-ratio obtained from the experimental data is consistent with the actual sampling distribution of F when the null hypothesis is true. Because of the probabilistic nature of this analysis, the researcher can never be certain if the calculated F-ratio corresponds to H_0 or H_a.

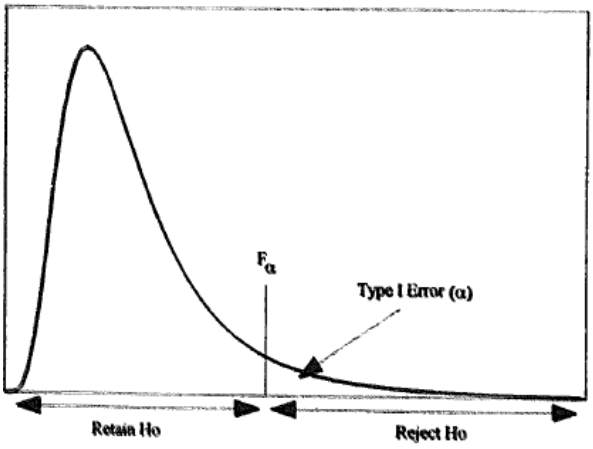

Fig. 1. Sampling distribution of F-ratio when the null hypothesis is true.

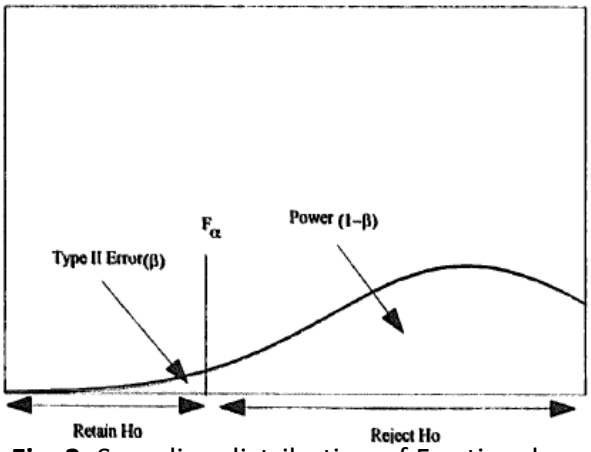

Fig. 2. Sampling distribution of F-ratio when the alternative hypothesis is true.

Fig. 1 represents the theoretical distribution of F when the null hypothesis is true. The region of rejection is specified to the right of f_α, which represents the magnitude of α. Hence, if the F-ratio calculated from the sample data falls in this region, then the null hypothesis will be rejected and an α error will be made. Fig. 2 represents the theoretical distribution of F when the alternative hypothesis is true. The region of rejection is again specified to the right of F_α. The area to the left of F_α is the probability of making an incorrect decision. The critical value of $F(F_\alpha)$ is the same in Figs. 1 and 2. This happens because, in general, F_α is set to a fixed value with the null hypothesis in mind. However, the value of β (and hence statistical power) is not fixed. Fig. 3 combines Figs. 1 and 2. The reciprocity of α and β errors is clear in Fig. 3. Any change in the size of the rejection region $F_\alpha$will produce opposite changes in the two types of errors. It is also clear from Fig. 3 that there is always a finite probability of making an error whether He is retained or rejected. Therefore, it is important in any statistical analysis to control both α and β errors. A balance between the two types of errors is needed because reducing any one type of error increases the probability of increasing the other type of error.

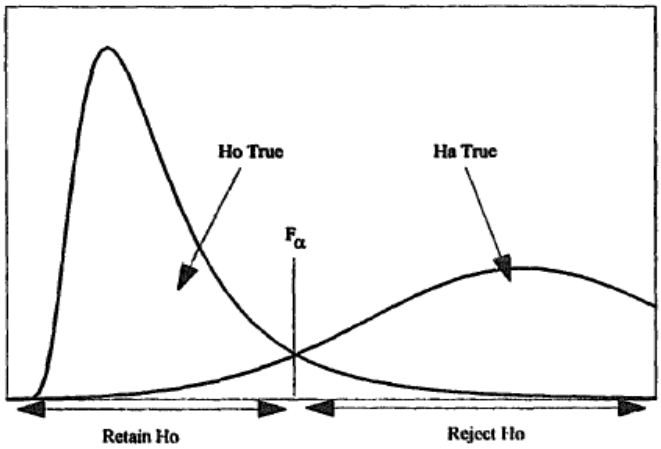

Fig. 3. Sampling distributions of F-ratio under null and alternative hypothesis

Typically, α is taken to 0.05 in most of the social/behavioral sciences. In several cases, however, $\alpha = 0.01$, 0.001 and 0.0001 are also used. Since the α and β errors are dependent on each other, a more stringent α level leads to a lower power for a given sample size. Often there is a tendency to report tests according to the "$p - value$" obtained on the computer printouts and to give more importance to" more significant" results. Doing this, however, reduces the power of those tests. In other words, if two statistical tests have different $p - values$ (say 0.05 and 0.001) but the same power (say 0.60) then both tests have the same probability (60%) of getting the same results if the statistical tests are repeated with new data sets under otherwise identical experimental conditions. Therefore, a better approach might be to treat all the tests which are significant at a fixed a value (say, $\alpha = 0.05$), with the same importance (Cohen, 1977, 1988).

There is no agreement among researchers on the issue of what defines a reasonable level of power. Certainly a power of 0.50 is too low. At the same time a power of 0.90 requires a very large sample size (Dallal, 1986). Recently, however, methodologists are beginning to

agree that a power of about 0.80 represents a reasonable and realistic value for research in social/behavioral sciences (Cohen, 1977, 1988; Hinlde and Oliver, 1983; Kirk, 1982). A power of 0.80 is reasonable in the sense that it reflects a general sentiment among researchers that α errors are more serious than β errors and that a 4:1 ratio of α to β error is probably appropriate.

Statistical Power and Effect Size

The shape and location of the non-central F distribution shown in Figs. 2 and 3 depend on several factors, one of which is the actual difference between the populations. *Effect size* is an index which measures the strength of association between the populations of interest (Cohen, 1977, 1988). Several measures of the effect size have been proposed. The reader is referred to Camp and Maxwell's (1983) article and Cohen's (1977, 1988) text for a detailed analysis of these effect size indices. Here we will briefly review some of the more popular approaches. One commonly used approach is known as *omega-square* (ω^2) (Keppel, 1991). When applied to a single factor experimental design, ω^2 is based on two variances derived from the populations, one is the differences *among* the population means $({\sigma_a}^2)$ and the other is the variability *within* the populations $({\sigma_{s/a}}^2)$.

$$\omega^2 = \frac{{\sigma_a}^2}{{\sigma_a}^2 + {\sigma_{s/a}}^2} \tag{3}$$

If there is no difference between the populations then $\omega^2 = 0$. The value of ω^2 varies between 0 and 1.0 when there are actual differences between the populations. Effect size, as measured by ω^2, is a *relative* measure, reflecting the proportional amount of the total population variance that is attributed to the variation among the populations under study. Cohen (1977, 1988), Camp and Maxwell (1983), and Vaughan and Corballis (1969) have developed a number of useful equations for calculating ω^2 for simple statistical tests. For example, for a one-way ANOVA with the same sample sizes, ω^2 can be estimated in the following manner (Keppel, 1991):

$$\widehat{\omega}^2 = \frac{(a-1)(F_{ob}-1)}{(a-1)(F_{ob}-1)+an} \tag{4}$$

where a represents the number of groups, n represents the sample size in each group and F_{ob} represents the observed F-ratio.

Although ω^2 is the most commonly reported measure of relative magnitude in experimental studies, other indices have also been used to estimate effect size. For example, Pearson and Hartley (1951, 1972) charts use a ϕ^2 statistic (based on an estimate of expected minimum effect size) to calculate the sample size required to get a reasonable value of power for a given significance level. Estimation of Pearson and Hartley's ϕ^2 requires sample size (n) for all groups (a), means of different populations (μ_i), the grand mean (μ_T) and the common within group variance in different populations $({\sigma_{s/a}}^2)$.

$$\phi^2 = n\frac{\frac{\sum(\mu_i-\mu_T)^2}{a}}{\sigma_{s/a}{}^2} \quad (5)$$

Effect size can also be estimated by Cohen's (Cohen, 1977, 1988) f statistic. Cohen uses the range of means of populations divided by the common within-group standard deviation as an index in effect size calculation. All measures of effect sizes however, are interrelated. For example, the following two equations show how sample size (n), ω^2, Pearson and Hartley's ϕ^2 and Cohen's f parameters are related to each other (Keppel, 1991):

$$\phi^2 = n\frac{\hat{\omega}^2}{1-\hat{\omega}^2}, \quad (6)$$

$$\phi^2 = nf^2. \quad (7)$$

Effect size indices are important because they help researchers distinguish between a meaningful effect and a trivial one and between the relative magnitude of effects. A small but significant F-ratio for a statistical test might suggest the presence of a *trivial effect* that was detected by a particularly powerful study whereas a medium but non-significant F-ratio might suggest the possible presence of an *important effect* that was not detected because of a serious lack of power. Cohen (1977, 1988) suggests the following rough guidelines to describe the size of an effect in the social/behavioral sciences:

- A "small" effect is an experiment that produces an ω^2 of 0.01 (approximately $\omega^2 < 0.03$).
- A "medium" effect is an experiment that produces an ω^2 of 0.06 (approximately ω^2 is 0.03 to 0.11).
- A "large" effect is an experiment that produces an ω^2 of 0.15 or greater (approximately $\omega^2 = 0.11$ or more).

Statistical Power and Sample Size

Even though the power of a statistical test depends on three factors, from a practical point of view only the sample size is used to control power. This is because the α level is effectively fixed at 0.05 (or some other value). Effect size can also be assumed to be fixed at some unknown value because generally researchers cannot change the effect of a particular phenomenon. Therefore sample size remains the only parameter that can be used to design empirical studies with high statistical power.

In general, bigger sample sizes are needed for higher statistical power. Increasingly larger sample sizes are needed to continuously increase power by a fixed amount (Kraemer, 1985, Kraemer and Theimann, 1987). Also, relatively small expected effect sizes require substantial sample sizes to achieve a reasonable power.

Table 2 demonstrates the relationship between statistical power, sample size and effect size for two different significance levels (Dallal, 1986). It clearly shows that for a given effect

Power value	Large effect		Medium effect		Small effect	
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
Power = 0.6	18	12	45	30	274	179
Power = 0.7	20	14	53	36	323	219
Power = 0.8	24	17	62	44	385	271
Power = 0.9	29	22	67	57	478	354

[a] Source: (Dallal, 1986).

Table 2 Sample size as a function of power, significance level and effect size

size, bigger sample sizes are needed to maintain the same power level if statistical significance is increased. For example, the required sample size increases from 44 to 62 for a medium effect size if α is reduced from 0.05 to 0.01 to maintain power at 0.80. Table 2 also shows that bigger sample sizes are required to get higher power and smaller effect sizes require relatively bigger samples to obtain reasonable power levels.

Pearson and Hartley (1951, 1972) have constructed some very helpful charts from which sample sizes can be estimated for a given power level. The Pearson and Hartley Charts present statistical power for two α levels (0.05 and 0.01), different degrees of freedom, and for different ϕ^2 values. Eq. (5) can be used to calculate ϕ^2 by conducting a pilot study with a small sample size. Then, the Pearson and Hartley chart can be used to estimate "sample size required" to get the "calculated ϕ^2" at a reasonable power level. Cohen (1977, 1988) developed another useful set of tables for calculating sample size required to get reasonable power levels at three different α levels (0.10, 0.05 and 0.01). Software programs that greatly facilitate the estimation of power and sample size are now available. In this research, we have used the *Statistical Power Analysis* program developed by Borenstein and Cohen (1988).

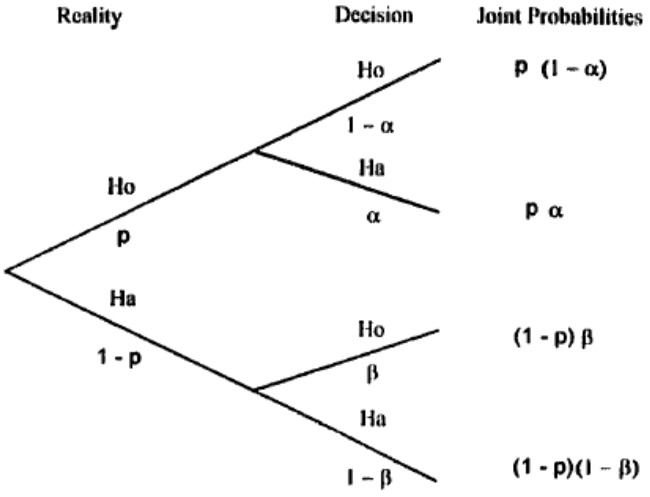

Fig. 4. Statistical Power: A Bayesian Analysis

Information about other programs is available in (Brecht et al., 1988; Dallal, 1986; Anderson, 1981; Goldstein, 1989; Stoloff and Couch, 1988). Minno (1991) reports that a PC version of SPSS can also conduct power analysis. A number of journals (for example *Behavior Research*

Methods, Instruments and Computers, Educational and Psychological Measurements, American Statistician, British Journal of Mathematical and Statistical Psychology, Multivariate Behavioral Research) and magazines *(PC Magazine, Byte)* periodically publish reviews of new statistical packages. Please refer to these publications for information regarding new software programs.

Statistical Power and Decision Making

The importance of statistical power can be further understood by a simple probability tree analysis. The probability tree in Fig. 4 shows the two states of nature in hypothesis testing (H_0 true or H_a true) and the two decisions (accept H_0 or reject H_a) made by the researchers for those states of nature. The researcher either accepts H o or rejects H 0 without knowing the true state of nature. From this probability tree it is clear that there is a finite probability of making an error no matter what decision the researcher makes. Therefore, the following four conditional probabilities are of interest.

If H_0 is accepted then two possibilities exist:

(A) Probability of H_0 being true in reality given that the researcher accepts H_0, P(RHoIDH0). This represents the conditional probability of making a correct decision when H o is accepted. Using Bayes' theorem, the above conditional probability can be calculated in the following manner (Watson et al., 1993):

$$P(RH_0|DH_0) = \frac{p(1-\alpha)}{p(1-\alpha)+(1-p)\beta} \quad (8)$$

(B) Probability of H_a being true in reality given that the researcher accepts H_0, $P(RH_0|DH_0)$. This represents the conditional probability of making a wrong decision when H o is accepted. Hence,

$$P(RH_a|DH_0) = \frac{(1-p)\beta}{p(1-\alpha)+(1-p)\beta} \quad (9)$$

If H_0 is rejected and H_0 is accepted then two possibilities exist:

(A) Probability of H_0 being true in reality given that the researcher accepts H_a, $P(RH_0|DH_a)$. This represents the conditional probability of making a wrong decision when H_0 is rejected. Therefore,

$$P(RH_0|DH_a) = \frac{p\alpha}{p\alpha+(1-p)(1-\beta)} \quad (10)$$

(B) Probability of H_a being true in reality given that the researcher accepts H_a, $P(RH_a|DH_a)$. This represents the conditional probability of making the correct decision when H_0 is rejected. Hence,

$$P(RH_a|DH_a) = \frac{(1-p)(1-\beta)}{p\alpha+(1-p)(1-\beta)} \quad (11)$$

Eqs. (8) and (11) represent the conditional probabilities of making the correct decisions and Eqs. (9) and (10) represent the conditional probabilities of making the wrong decisions. In an actual experiment, the value of α is often fixed (in advance). Eqs. (8)-(11) contain "p", which is defined as the probability of H_0 being true in reality. The objective of this analysis is to show the general effect of β error (and hence statistical power) on the likelihood of making right and wrong decisions. We chose 5 different levels of p (0.1, 0.3, 0.5, 0.7, 0.9) to represent different scenarios. When $p = 0.1$, it represents a very high probability that H_a, is true in reality. Similarly, $p = 0.0$ represents a high probability of H_0 being true. The value of α was fixed to 0.05 in this analysis.

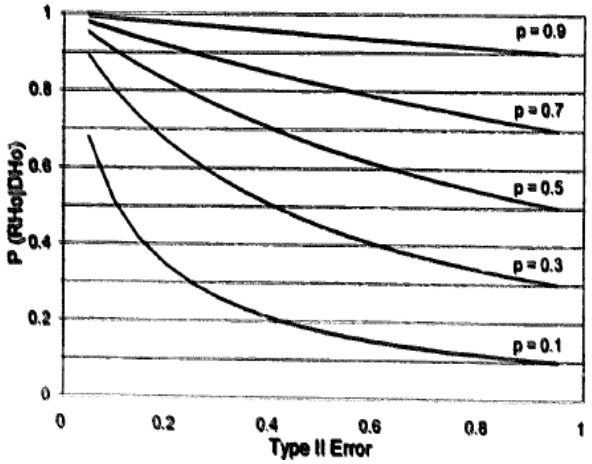

Fig. 5. Probability of making a correct decision when H_0 is accepted.

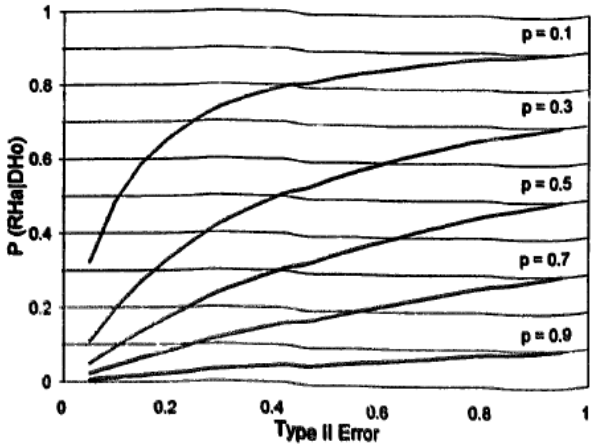

Fig. 6. Probability of making a wrong decision when H_0 is accepted.

Figs. 5, 6, 7 and 8 show the changes in the four conditional probabilities with respect to changes in the β value, for different "p" levels. It can be clearly seen from these plots that the conditional probabilities of making the correct decision decrease rapidly as β increases from a

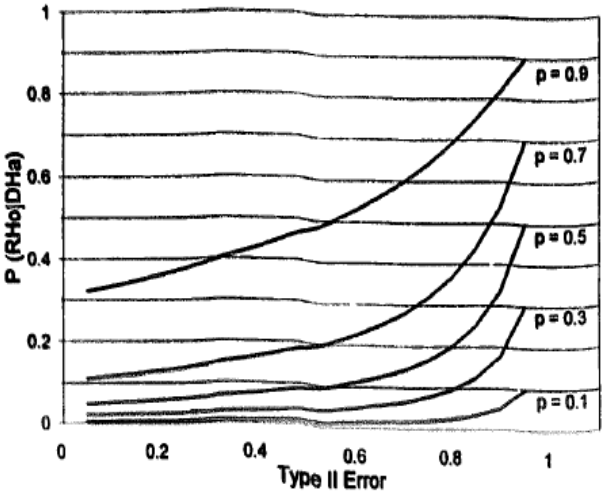

Fig. 7. Probability of making a wrong decision when H_a is accepted.

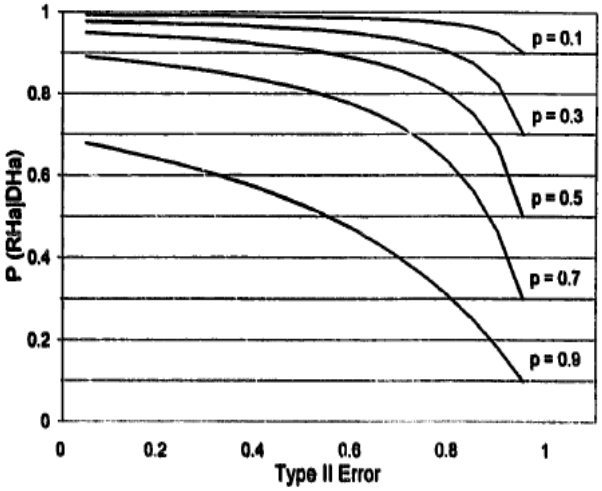

Fig. 8. Probability of making a correct decision when H_a is accepted.

small (0.2) to a large (0.8) value (Figs. 5 and 7). It is also clear from Figs. 6 and 8 that the probability of making the wrong decision increases as the β value increases.

The above analysis suggests that it is in the best interest of the research project to keep the value of as low (or statistical power as high) as possible. Figs. 7 and 8 should be of special interest to most empirical researchers because more often they are interested in rejecting H_0 and accepting H_a.

Previous Studies of Statistical Power

Although statistical power is very important in empirical research, traditionally most of the emphasis has been given to statistical significance. The next section of the paper presents a brief literature review of the power analysis of published research in various social/behavioral sciences.

In spite of the arguments in the previous section, most researchers pay little attention to power and, in fact, most of the empirical studies in behavioral/social sciences are lacking in power. Jacob Cohen (1962), in a classic study on the statistical power analysis of the articles published in the 1960 volume of the *Journal of Abnormal and Social Psychology,* found that the average power of the articles was only 0.48. This means that the significant effects reported in this volume of the journal would have, on average, about a 50-50 chance of being detected by others who try to duplicate these findings. Recently, Sedlmeier and Gigerenzer (1989) duplicated Cohen's study and analyzed the articles published in the 1984 volume of the *Journal of Abnormal Psychology,* and came to almost the same conclusion - the average power for detecting the median effect was found to be 0.50. Recently, several other studies were conducted in other social/behavioral sciences which also came to the conclusion that most of the research articles were seriously lacking in power. For example, Brewer (1972) found that the power of the studies with medium effect size as 0.58, 0.71, and 0.52, respectively, in *American Journal of Educational Research, Journal of Research in Teaching,* and *The Research Quarterly.* Chase and Chase (1976) studied the articles published in the *Journal of Applied Psychology* and found the average power for experiments with medium effect size to be 0.67. Table 3 presents a summary of results from a number of power calculation studies.

Osbom (1990) conducted a survey of manufacturing organizations in the United States using statistical control charts. He found that 75% of the processes had less than 50% chance of detecting a critical shift on the first sample taken after the shift had occurred. Out of' the 61 respondents, he found that power levels for 13 processes were less than 10% and eight of those were actually less than 1%. Ashton (1990) concludes that statistical power might be the reason for not being able to detect superior investment performance. Baroudi and Orlikowski (1989) conducted a survey of published Management Information Systems literature for five years and concluded that there is about a 40% chance of failure to detect the phenomenon under study, even though it may exist.

Research discipline	Source	Small effect	Medium effect	Large effect
Applied psychology	(Chase and Chase, 1976)	0.25	0.67	0.86
Communication	(Chase and Tucker, 1976)	0.18	0.52	0.79
Education	(Brewer, 1972)	0.13	0.47	0.73
Evaluation Research	(Lipsey, 1990)	0.28	0.63	0.81
Gerontology	(Levenson, 1980)	0.37	0.88	0.96
Management information systems	(Baroudi and Orlikowski, 1989)	0.19	0.60	0.83
Mathematics education	(Clark, 1974)	0.24	0.62	0.83
Mass communication	(Chase and Baran, 1976)	0.34	0.76	0.91
Management	(Mazen et al., 1987a,b)	0.31	0.77	0.91
Marketing research	(Sawyer and Ball, 1981)	0.24	0.69	0.87
Medicine	(Reed and Slaichert, 1981)	0.14	0.39	0.61
Occupational therapy	(Ottenbacher, 1982)	0.37	0.65	0.93
Social psychology	(Cohen, 1962)	0.18	0.48	0.83
Sociology	(Spreitzer, 1974)	0.55	0.84	0.94
Speech Pathology	(Kroll and Chase, 1975)	0.16	0.44	0.73
Strategic management	(Mazen et al., 1987a,b)	0.23	0.59	0.83

Table 3 Average statistical power levels in empirical research in various disciplines

The unfortunate conclusion from all these findings is that research in social/behavioral sciences is woefully lacking in power. This statement implies that a substantial number of research projects have been undertaken and then discarded when they failed to produce results at acceptable significance levels. If the power is 0.50, then it suggests that half the research undertaken will not yield significant results even though there are real differences among the treatment conditions.

Statistical Power in the *Journal of Operations Management* and in *Decision Sciences*

A survey of 28 empirical research articles of the last five years in the *Journal of Operations Management* (vols. 7-10) and *Decision Sciences* (vols. 20-24) was undertaken. The reason for presenting the results of the power calculations is to show the general trend of statistical power in POM and related disciplines. The tests analyzed in this study were correlations, t-tests, and multiple regression. Other statistical tests were not considered because of the lack of the necessary information reported in most of the articles.

We used the *Statistical Power Analysis* program developed by Borenstein and Cohen (1988) to calculate the power of the statistical tests. This program is very easy to use and only requires summary statistics to calculate power. For example, the power calculation for a t-test requires knowing two sample means, sample standard deviations, sample sizes, and a level used. The program first estimates effect size (ω^2- Eq. (3)), and then calculates power as a function of effect size, sample sizes and significance level.

Table 4 presents the power levels found in the statistical tests for each journal with respect to High (0.80 to 1.0), Medium (0.60 to 0.80), Low (0.40 to 0.60), and Very Low Power (< 0.40). This shows that 317 out of the 524 tests (60%) do not have a high power level ($\geq$ 0.80). According to Table 3, *Decision Sciences* has a larger number of tests both at very low and high power. If we consider medium and high power levels to be acceptable, then the 73% of tests in

the *Journal of Operations Management* have acceptable power. About 59% of the tests in the *Decision Sciences* have acceptable power levels.

Table 4 also shows that regression analysis tends to have high power (92% of the tests) compared to correlation (36%) and t-tests (19%). For correlation, the number of tests classified as Low, Medium, and High Power are approximately uniformly distributed (30-35% in each). For t-tests, 64% of the tests have Very Low or Low power. One possible reason for relatively low power levels in correlation and t-tests might be their extensive: use in exploratory-type analysis.

Classification		Power level			
		Very low	Low	Medium	High
Journal	Decision sciences	34	84	50	124
	Journal of Operations Management	1	63	85	83
Method	Correlation	18	123	121	146
	t-test	17	23	11	12
	Regression	0	1	3	49

Table 4 Cross tabulation of journal and method by power level of statistical tests

With respect to the α level, Table 5 tends to support the contention that researchers focus on α levels and maintain the α levels at acceptable values; i.e., 98% have $\alpha \leq 0.05$ and 34% have $\alpha \leq 0.001$. Table 5 also looks at α level and power level. The results show that 58% of the tests with $\alpha = 0.001$ have high power but only approximately 30% of the tests with $\alpha \leq 0.01$ or 0.05 have high power. A possible reason for the high power of tests with $\alpha = 0.001$ might be bigger effect sizes and/or bigger sample sizes. Still, 42% of the tests with $\alpha = 0.001$ do not have high power. Additionally, approximately 70% of the tests with $\alpha = 0.01$ or $\alpha = 0.05$ have power levels that are not high.

Table 6 presents the results of the cross tabulation of effect size with journal, method, and power level. This shows that most of the tests reported in the *Journal of Operations Management and Decision Sciences are* analyzing effects considered large or medium. Small or subtle effects only account for 13% of the tests. A noticeable trend that can be identified in Table 6 is that tests with larger effect sizes achieve relatively higher power. In other words, 65% of the tests with large effect size have high power, 22% of the tests with medium effect size have high power, and only 10% of the tests with small effect size have high power.

Table 6 also shows that 77% of the regression tests analyzed large effects. On the other hand, 62% of the t-tests and 37% of the correlations analyzed large effects. Recall that Table 4 showed that most of the regression tests had much higher power than t-tests or correlations. Larger effect size might be a possible reason for this result. A detailed cross tabulation of the power analysis for each article is provided in Table 7 to show the relative contribution of each article to the aggregate measures used in the previous tables.

Classification		Effect size		
		Large	Medium	Small
Journal	Decision Sciences	128	100	64
	Journal of Operations Management	102	126	4
Method	Correlation	150	202	56
	t-test	39	13	11
	Regression	41	11	1
Power level	Very low	10	9	16
	Low	25	84	38
	Medium	45	83	7
	High	150	50	7

Table 6 Cross tabulation of journal by effect size of statistical tests

Article	Power level			
	Very low	Low	Medium	High
#1	0	0	0	2
#2	9	7	1	2
#3	1	1	3	10
#4	0	0	0	2
#5	0	11	2	21
#6	0	4	0	18
#7	0	0	0	7
#8	0	0	0	3
#9	9	18	6	11
#10	2	1	0	0
#11	0	8	1	3
#12	0	0	0	1
#13	2	8	9	4
#14	0	1	0	0
#15	0	1	0	0
#16	3	1	0	0
#17	5	9	14	4
#18	0	6	6	4
#19	3	0	1	5
#20	0	3	5	15
#21	0	0	0	1
#22	0	3	0	0
#23	0	0	2	10
#24	0	2	0	1
#25	0	4	2	4
#26	0	0	0	24
#27	1	50	60	7
#28	0	9	23	48
Total	35	147	135	207
Percent	6.7%	28.0%	25.8%	39.5%

Table 7 Cross tabulation of articles by power level of statistical tests

Concluding Remarks

The objective of this article was to stress the need and importance of statistical power in empirical research in POM and related fields. We have presented a detailed review of the concepts related to statistical power. A meta-analysis of empirical research articles published in the *Journal of Operations Management* and in *Decision Sciences* has also been presented. The meta-analysis suggests that our field does relatively better than several other social sciences. This is good news, especially because empirical research is relatively new to POM. At the same time, the meta-analysis also shows that several of the statistical tests had very low power.

Overall, the average power of correlation, t-tests and regression analysis reported in *Journal of Operations Management* and in *Decision Sciences* were approximately 0.71 and 0.72, respectively (from Table 4). The average power for tests with large, medium and small sizes were 0.85, 0.65 and 0.49, respectively. These numbers suggest that, overall, empirical articles in POM and related disciplines are doing better than several other disciplines. The literature of several other social/behavioral sciences show that the mean (or median) power observed was close to 0.50. Still, low power levels for tests with medium and small effect sizes is a reason for concern.

An interesting trend is prevalent for the α level. Researchers tend to focus on the α level and report lower α levels as more important, e.g., $p \leq 0.001$ is more meaningful than $p \leq 0.01$. However, both tests could have the same probability of making a Type II error. This means that if the studies are repeated, both tests have an equal probability of not finding the same significance level (same α) as was found in the original tests. Hence, a better strategy will be to treat all results significant at a predetermined or level with the same importance.

Currently, empirical studies in POM and related topics tend to focus on high and medium effects. This is not a surprising observation because empirical research is just beginning to make inroads in these fields. Other social/behavioral sciences have also gone through similar phases. A growing field starts by studying larger effects to "map the territory". As a field becomes more mature, more research is undertaken which explores smaller effects. Recall that statistical power generally suffers in current studies that analyze smaller effects. We feel that future empirical research will study increasingly smaller effects, so power levels will become increasingly important.

Recall that statistical power is a function of three factors: α level, effect size, and sample size. For a given study, effect size can be considered more-or-less fixed. Similarly, acceptable α levels are set by the norms of the field. Hence, only sample size is used as a controlling factor for generating acceptable power levels. With this information and some a priori assumptions, a more sensitive, powerful, and economical study can be designed.

A number of statistical power analysis tools (Pearson and Hartley, 1951, 1972; Cohen, 1977, 1988) and software programs (Borenstein and Cohen, 1988; Brecht et al., 1988; Dallal, 1986; Goldstein, 1989; Minno, 1991) are available for researchers planning to conduct power

analysis prior to conducting a large scale empirical study. We strongly recommend that researchers begin by conducting small pilot studies and then use their results to calculate the sample size required to get a reasonable power level in full-scale empirical studies. In fact, it might be possible to collect required data for a simple power analysis by using the results of pre-tests of empirical research instruments. Then, the study results should also note the power of the tests.

Generally, empirical studies are conducted to identify or verify relationships and/or infer cause-and-effect for the phenomenon of interest. Therefore, we feel that statistical power analysis is a very useful tool in meeting the goals of empirical studies because high power increases the probability of making correct decisions. For further reading, see (Cohen, 1988; Lipsey, 1990).

Acknowledge

We are grateful to Gary M. Thompson, Jack R. Meredith and two anonymous reviewers for many helpful comments.

References

Adam, E.E., Jr. and P.M. Swamidass, 1989. "Assessing operations management from a strategic perspective", Journal of Management, vol. 15, no. 2, pp. 181-203.

Amoako-Gyampah, K. and J.R. Meredith, 1989. "The operations management research agenda: An update", Journal of Operations Management, vol. 8, pp. 250-262.

Anderson, R.B., 1981. STATPOWER. An Apple Computer Program, ABT Associates, Cambridge, MA.

Ashton, D.J., 1990. "A problem in the detection of superior investment performance", Journal of Business Finance and Accounting, vol. 17, no. 3, pp. 337-350.

Baroudi, J.J. and W.J. Orlikowski, 1989. "The problem of statistical power in MIS research", MIS Quarterly, vol. 13, no. 1, pp. 87-106.

Borenstein, M. and J. Cohen, 1988. Statistical Power Analysis: A Computer Program, Erlbaum, Hillsdale, NJ.

Brecht, M.L., J.A. Woodward and D.G. Bonett, 1988. GANOVA4, Department of Psychology, University of California, Los Angeles, CA.

Brewer, J.K., 1972. "On the power of statistical tests in American educational research journal", American Educational Research Journal, vol. 9, pp. 391-401.

Camp, C.J. and S.E. Maxwell, 1983. "A comparison of various strengths of association measures commonly used in gerontological research", Journal of Gerontology, vol. 38, pp. 3-7.

Chase, R.B., 1980. "A classification and evaluation of research in operations management", Journal of Operations Management, vol. 1, no. 1, pp. 9-14.

Chase, L.J. and Chase, R.B., 1976. "A statistical power analysis of applied psychological research", Journal of Applied Psychology, vol. 61, PO. 23~-237.

Chase, L.J. and S.J. Baran, 1976. "An assessment of quantitative research in mass communication", Journalism Quarterly, vol. 53, pp. 308-311.

Chase, R.B. and E.L. Prentis, 1987. "Operations management: A field rediscovered", Journal of Management, vol. 13, pp. 351-366.

Chase, L.J. and R.K. Tucker, 1976. "Statistical power: Derivation, development, and data-analytic implications", The Psychological Record, vol. 26, pp. 473-486.

Clark, R., 1974. A study of the power of research as reported in the Journal of Research in Mathematics Education, Unpublished doctoral dissertation, University of Tennessee.

Cohen J., 1962. "The statistical power of abnormal-social psychological research: A review", Journal of Abnormal Psychology, vol. 65, pp. 145-153.

Cohen, J., 1977. Statistical Power Analysis for the Behavioral Sciences, Academic Press, New York.

Cohen, J., 1988. Statistical Power Analysis, Eribaum, Hillside, NJ.

Cohen, J., 1992. "A power premier", Psychological Bulletin, vol. 112, no. 1, pp. 155-159.

Dallai, G.E., 1986. "PC-SIZE: A program for sample size determinations", The American Statistician, vol. 40.

Ebert, R.J., 1990. "Announcement on empirical/field based methodologies in JOM", Journal of Operations Management, vol. 9, no. 1, pp. 135-137.

Flynn, B.B., S. Sakakibara, R.G. Schroeder, K.A. Bates and E.J. Flynn, 1990. "Empirical research methods in operations management", Journal of Operations Management, vol. 9, no. 2, pp. 250-284.

Goldstein, R., 1989. "Power and sample size via MS/PC-DOS computers", The American Statistician, vol. 43, pp. 253-260.

Hinkle, D.E. and J.D. Oliver, 1983. "How large should the sample size be? A question with no simple answer", Educational and Psychological Measurement, vol. 43, pp. 1051-1060.

Keppel, G., 1991. Design and Analysis: A Researcher's Handbook, Prentice Hall, Englewood Cliffs, 3rd Edition.

Kirk, R.E., 1982. Experimental Design: Procedures for the Behavioral Sciences, 2nd Edition, Brooks/Cole, Monterey, CA.

Kraemer, H.C., 1985. "A strategy to teach the concept and application of power of statistical tests", Journal of Educational Statistics, vol. 10, pp. 173-195.

Kraemer, H.C. and Theimann, S., 1987. How Many Subjects?: Statistical Power Analysis in Research, Sage, Newbury Park.

Kroll, R.M. and L.J. Chase, 1975. "Communication disorders: A power analytic assessment of recent research", Journal of Communication Disorders, vol. 8, pp. 237-247.

Levenson, R.I., 1980. "Statistical power analysts: Implications for researchers, planners, and practitioners in gerontology", The Gerontologist, vol. 20, pp. 494-498.

Lipsey, M.W., 1990. Design Sensitivity: Statistical Power for Experimental Research, Sage, Newbury Park.

Mazen, A.M., M. Hemmasi and M.F. Lewis, 1987a. "Assessment of statistical power in contemporary strategy research", Strategic Management Journal, vol. 8, no. 4, pp. 403-410.

Mazen, M.M., L.A. Graf, C.E. Kellogg and M., Hemmasi, 1987b. "Statistical power in contemporary management research", Academy of Management Journal, vol. 30, no. 2, pp. 369-380.

Meredith, J.R., A. Raturi, K. Amoako.Gyampah and B. Kaplan, 1989. "Alternative research paradigms in operations management", Journal of Operations Management, vol. 8, no. 4, pp. 297-320.

Minno, J., 1991. "Firm often true mainframe statistical power for PCs", Marketing News, vol. 25, no. 1.

Osbom, D.P., 1990. "Statistical power and sample size for control charts - Survey results and implications", Production and Inventory Management, vol. 31, no. 4, pp. 49-54.

Ottenbacher, K., 1982, "Statistical power and research in occupational therapy", Occupational Therapy Journal of Research, vol. 2, pp. 13-25.

Pearson, E.S. and H.O. Hartley, 1951. "Charts of the power function for analysis of variance tests, derived from the non-central F distribution", Biometrika, vol. 38, pp. 112-130.

Pearson, E.S. and H.O. Hartley, 1972. Btometrtka Tables for Statisticians, Cambridge University Press, London.

Reed, J.F. and W. Slaichert, 1981. "Statistical proof in inconcluslve "negative" trials", Archives of Internal Medicine, vol. 141, pp, 1307-1310.

Sawyer, A.G. and A.D. Ball, 1981. "Statistical power and effect size in marketing research", Journal of Marketing Research, vol. 18, pp. 275-290.

Sedlmeier, P. and G. Gigerenzer, 1989. "Do studies of statistical power have an effect on the power of studies", Psychological Bulletin, vol. 105, pp. 309.-316.

Spreitzer, E., 1974. Statistical power in sociological research: An examination of data-analytic strategies, Unpublished manuscript, Department of Sociology, Bowling Green State University.

Stoloff, M.U and J.V. Couch, 1988. Computer Use in Psychology: A Directory of Software, Second Edition, American Psychological Association, Washington, DC.

Swamidass, P.M., 1991. "Empirical science: New frontier in operations management research", Academy of Management Review, vol. 16, no. 4, pp. 793-814.

Vanghan, G.M. and M.C. Corballis, 1969. "Beyond tests of significance: Estimating strength of effects in selected ANOVA designs", Psychological Bulletin, vol. 72, pp. 384-386.

Watson, Billingsley, Croft and Huntsberger, 1993. Statistics for Management and Economics, 6th Edition, Allyn and Bacon, Newton, MA.

Authors Bios

Rohit Verma is a Ph.D. candidate in Operations Management at the University of Utah. He holds a B. Tech. in Metallurgical Engineering from the Indian Institute of Technology, Kanpur and M.S. in Metallurgical Engineering from the University of Utah, His dissertation integrates customer-based objectives, and operating constraints into production, process improvement and product-design related decisions of operations managers. He is a member of APICS, Academy of Management, Decision Sciences, INFORMS, and POM society. His previous research has appeared in *Powder Technology.*

John C. Goodale is a Ph.D. candidate in Operations Management at the University of Utah, He will be joining the management faculty at Bali State University in August 1995. He holds an B.S. in Mechanical Engineering from Michigan State University and an M.B.A. from the University of Utah. A member of Decision Sciences Institute, INFORMS, and the Academy of Management, his dissertation is titled "Accounting for individual productivity in labor tour scheduling". His previous research has appeared in the *International Journal of Production Research.*