Alpha, Null Hypothesis Statistical Testing and Confidence Intervals: Where Do We Go From Here?

Tamela D. Ferguson and William L. Ferguson
University of Louisiana at Lafayette

The choice of alpha = .05 to determine significance in null hypothesis statistical testing (NHST) has become ingrained in management research, though the choice of .05 level appears to have little scientific basis and is likely simple convenience. More appropriate approaches might be to choose an alpha level based on specific research design or stage of research stream development. Over the years, many have criticized NHST use, with some calling for its ban while advocating the use of confidence intervals instead. To this end, this paper presents a historical review, support for and criticisms of alpha and the related NHST, as well as discussion of issues concerning the use of confidence intervals as an alternative to NHST. The development of the organizational configurations-performance stream in strategic research is used to illustrate the critical importance of researchers making appropriate statistical testing choices.

Jacob Bernoulli (1654 -1705) first recognized problems with developing probabilities from sample data, and we still struggle with many related issues today. An approach with roots in this century, the choice of alpha = .05 to decide statistical significance in null hypothesis statistical testing (NHST) has become the primary statistical method of data analysis in much social science research. However, the choice of .05 as the alpha level has little scientific basis and is perhaps simply a matter of convenience. More appropriate approaches might be to choose an appropriate alpha level based on specific research design needs, especially relative to the importance of Type I and II errors, or even stage of research stream development. For instance, higher alphas may be more appropriate in younger research streams where large effects are commonly the focus of study. Whatever the alpha chosen, current NHST practices are a hybrid of the early works of Fisher (1925, 1926) and Neyman-Pearson (1928), implying a polemic history when original researcher intent is considered.

NHST has been the target of criticism over the last several decades, including 1) the notion that scientific inference and NHST address different questions, and 2) the use of dichotomous decisions of significance when a continuum of uncertainty is present. Of equal concern is debate pertaining to the usefulness and appropriateness of NHST, with the use of confidence intervals touted as perhaps more appropriate. For decades, alphas of .05 and NHST have been guiding lights for researchers determining statistical significance. However, their use is perhaps evolving into a practical significance approach in expanded reporting of results, which may include confidence intervals. Understanding the logic behind alpha selection and

the resultant impact on NHST usefulness should better position researchers to intelligently influence an appropriate course of action in the NHST/confidence interval debate.

Alpha (α) and Null Hypothesis Statistical Testing

While researchers have the choice of alpha levels (*i.e.*, acceptable Type I error), $\alpha = .05$ is by far the most common, even "sacred," in hypothesis testing (Skipper, Guenther & Nass, 1967). There is a long history of its use, with Fisher (1925, 1926) credited with establishing $\alpha = .05$ as a modern-day gauge of statistical significance. Even researchers attempting to summarize findings extensively rely on the .05 significance level as acceptable. For instance, review articles assessing statistical power in strategic management research (Magid, Mazen, Hemmasi & Lewis, 1987) as well as psychology and management research (Mone, Mueller & Mauland, 1996) both uniformly used non-directional null hypothesis testing at $\alpha = .05$ throughout, with no disclosure of rationale behind the researchers' choices. The researchers may have simply relied, as many have, on the .05 level to determine statistical significance without considering alternatives or potential ramifications of their choice.

Perhaps the extensive use of .05 as the alpha level of choice falls under Popper's (1959) label of conventionalism. In other words, rather than debating issues, researchers simply following some pattern (such as choosing $\alpha = .05$), then act as if their agreed upon behavior has scientific merit when in fact it is an *ad populum* argument—because we all believe, it must be true (Ferguson & Ketchen, 1999). However, some researchers assert the level of significance is one of the least important facets of research (*e.g.*, Lykken, 1968). In fact, as there are no right or wrong levels of significance, the choice should be considered as just another research parameter (Sauley & Bedeian, 1989; Skipper, Guenther & Nass, 1967). Thus, researchers are free and encouraged to choose the significance level they believe appropriate (Yule & Kendall, 1950).

The use of .05 significance appears to have little scientific basis and may be simply a matter of convenience (Winer, 1962; 1971). Thus, little reason may exist to arbitrarily accept $\alpha = .05$ when other alpha levels, particularly larger ones such as $\alpha = .10$, may be more appropriate for the research at hand. For instance, a higher alpha may be appropriate when the identification of large-effect relationships is of paramount interest to researchers, such as in relatively younger research streams. Sauley & Bedeian (1989) provide an extensive review of researcher considerations (*e.g.*, sample size, effect size, practical consequences) when selecting an appropriate level of statistical significance in null hypothesis statistical testing.

Still others assert that the arbitrary use of an alpha level of .05 as a point of dichotomy to determine the significance of research, along with the application of null hypothesis statistical testing (NHST), has limited the advancement of research (*e.g.*, Hunter, 1997; Loftus, 1996; Loftus & Masson, 1994; Schmidt, 1996). These same researchers generally argue that the use of confidence intervals may provide a more appealing method of testing and reporting statistical significance than the currently heavily relied upon alpha of .05 and NHST. In order to entertain the debate and present these issues in a logical order, a historical review of alpha and the related NHST, along with criticisms, as well as the use of confidence intervals as an alternative to NHST will be offered. This is followed by addressing the implications of these

statistical testing choices on the development of the configuration/strategic group research stream within the field of strategic management.

Historical Development of Null Hypothesis Statistical Testing

Problems inherent to developing probabilities from sample data have long been recognized. Jacob Bernoulli (1654-1705), of the noted Swiss family of mathematicians, is credited with developing the first treatise on sampling theory (Bernoulli, 1713), published posthumously by his nephew Nicholas. In so doing, Jacob Bernoulli advanced the focus of probability theory from rather abstract problems to those with greater real life application (*i.e.*, from predicting *a priori* outcomes in games of chance to *a posteriori* estimation and inference of various phenomena). However Jacob's standard of "moral certainty" (Bernstein, 1996), a concession to the unattainable standard of absolute certainty with regard to estimating the outcome of real events, was quite high (*i.e.*, he favored 1000/1001, implying p=0.001). Such moral certainty imposes severe computational and empirical burdens on researchers, which were even recognized by Nicholas ("The Slow"). Nicholas later expanded upon the work of Jacob through early application of the concept of confidence intervals, which were later generalized in the seminal development of the normal distribution by DeMoivre (1738). DeMoivre's normal distribution facilitated the determination of statistical measures of dispersion about a true mean (*e.g.*, standard deviation), the cornerstone of modern statistical inference.

The introduction of α into hypothesis testing occurred only after a succession of changes in methodological approach. Three modern approaches to hypothesis testing developed: Bayesian, null hypothesis significance testing, and competing hypotheses (see Loftus & Masson, 1994 for an extensive review). The Bayesian approach strives to determine the distribution of various population parameters or estimates the probability that a hypothesis is true given a particular data set. This approach is based on prior probability, and hence potential researcher opinion, and has not been widely influential in modern behavioral sciences.

Based on inductive inference, null hypothesis statistical testing (NHST) determines the likelihood of the given observations, if the null hypothesis is true (Fisher, 1925). Fisher contended the likelihood of observing some alternative hypothesis was unknown. Thus, his work revolved around a single (*i.e.*, null) hypothesis. Originally selecting one in 20 (*i.e.*, p=.05) as a "convenient" point to determine significance or not (Fisher, 1925), he later acknowledged p=.05 was a personal preference and that other significance levels were acceptable (Fisher, 1926).

Another account of the development of α =.05 as a measure of a significance benchmark is also credited to Fisher. His early work on quartile distributions resulted in introduction of the phrase "probable error" (PE), the value of which is equal to 0.67456, as a way to identify deviation from central measure. Fisher (1925) initiated the trend to express distribution values in terms of standard deviations instead of probable errors (Cowles & Davis, 1982), again building on the work of DeMoivre (1738). Fisher developed a probability integral table for which standard deviation could be used to find the portion of the population with a larger

value, thus making the standard deviation even more of a gauge (Fisher, 1925). He proposed that a deviation approximately three times the probable error is roughly twice the standard error—that is, a *Z* score of two in current vernacular, or 4.56% of the population. Historians of the origins of $\alpha = .05$ believe that this 4.56% level was simply rounded to 5% and utilized in some of Fisher's later work because understanding it was easier than *Z* scores (Cowles & Davis, 1982).

Whereas Fisher presented only the null hypothesis, Neyman and Pearson (1928) argued for including a competing (*i.e.*, alternative) hypothesis, and assumed one of the two was true. Type I and Type II error were introduced as rules were structured for assessing which hypothesis was correct. This made statistical power estimates possible—a parameter Fisher (1926) believed could not be assessed. In essence, α is the risk of committing a Type I error, and is directly related to statistical power: the lower the α, the lower the statistical power and the higher the probability of accepting a true null hypothesis. An inverse relationship exists between α and β (the probability of a Type II error). Thus, *ceteris paribus*, essentially only two ways to increase statistical power exist: 1) increase sample size, or 2) increase α, thereby reducing β. Researchers must recognize that they decide the level of acceptable risk of *both* Type I and Type II error when conducting a test. In many situations funding levels, time, and various other constraints limit the feasibility of increasing the sample size. Therefore, researchers should sometimes consider a higher α, thereby reducing β and increasing statistical power.

For example, increasing α is appropriate if overlooking a true departure from the null hypothesis is more costly than is false rejection of the null (*i.e.*, Type II errors cannot be ignored or are costly, as are Type I errors). Ferguson, Barrese and Levy (1998) provide a relevant discussion involving potential bias in widely publicized, commonly relied upon insurance company financial strength ratings issued by independent rating agencies (*e.g.*, A.M. Best, Standard & Poor's, Moody's). While rating agencies justifiably try to minimize errors for obvious business reasons, an insurer may be assigned a financial strength rating that may be either higher (Type I error) or lower (Type II) than appropriate relative to the true financial strength of the insurer. If financially strong insurers are incorrectly assigned lower than appropriate ratings, both society as a whole and individual insurers incur significant costs. To the extent lower ratings overestimate the true level of insurer risk, both the supply of insurance and the utility of insurance purchased will be lower than socially optimal levels. At the margin some insurance purchases will not be made and some risks will not be undertaken because of these constraints, resulting in lower output and utility than if the ratings had been correct.

Individual insurers with lower ratings also experience lower than expected demand for products from higher quality (lower risk) customers who rationally prefer greater security and certainty of potential claims payments. The resultant applicant pool will be relatively inferior, increasing the likelihood of adverse underwriting results and contributing to lower premium receipts, thereby increasing liquidity risk. These in turn may entice an insurer to accept higher than appropriate investment risk to try to offset the adverse underwriting. In sum, higher insurer insolvency risk and associated default costs necessarily result. Thus, incorrect

low ratings (the Type II error) can have important effects at both the firm and societal level. Therefore, in this type of situation an increased alpha level may be preferable in order to better recognize the impact of significant error effects.

Conversely, decreasing alpha may be appropriate where false positives are more costly or problematic than false negatives (*i.e.,* Type II errors are less a concern than Type I errors). For instance, if we are testing the effectiveness of equipment critical to the life safety of workers (*e.g.,* O-rings on NASA's *Challenger*), then Type I errors are likely to be of utmost importance. That is, an incorrect assumption regarding the suitability for use of the equipment could (and ultimately did) prove disastrous. In this case Type I errors would need to be virtually eliminated, necessitating a low alpha. Furthermore, if the null is in fact false, the probability of a Type I error is nonexistent because only two possible outcomes exist: 1) fail to reject the null (Type II error, probability = β), or 2) null is correctly rejected (probability represented by statistical power, 1 - β). Since most social science research is structured to routinely reject the null hypothesis, assessment of the impact of both Type I and II error should be acknowledged. However, that is rarely the case in published research (see Ferguson & Ketchen, 1999; Magid *et al.,* 1987; and Mone *et al.*, 1996).

Modern behavioral sciences have adopted essentially a hybrid approach to NHST, incorporating elements of both Fisher's and Neyman-Pearson's work (Gigerenzer, 1993). Current interpretation of hypothesis testing processes include Neyman-Pearson's *a priori* specification of the significance level and competing hypotheses, coupled with Fisher's recommended .05 and assertion that nonsignificant statistical test results should evoke no conclusions. In some ways contemporary methods are counter to the original intentions of either work. For instance, Type I and II errors were intended for use in dichotomous decision-making by Neyman-Pearson, but are often used as measures of belief scope by some researchers (Loftus & Masson, 1994). Similarly, findings of significance are often used to imply replicability (Falk & Greenbaum, 1995).

Alpha indeed has a polemic history when original researcher intentions are compared with contemporary usage. The historical development of alpha, or the acceptable level of a Type I error, seems to have included some chance in and of itself, with conceptual misapplication somewhat commonplace, even in textbooks designed to teach basic statistical inference (Cohen, 1994; Dar, Serlin & Omer, 1994). Of particular concern is the use of an alpha of .05 in a nondescript way with little concern for research implications. As Morrison and Henkel (1969) espouse, if α =.05 is indeed "sacred," then researchers are practicing religion, not research, and thus should forget empirical work and develop more rituals. Hence, it seems advisable that researchers should be aware of issues surrounding appropriate α selection, resulting influences of Type I and II error tradeoffs in their own research, as well as researcher influence on acceptable testing parameters for the field as a whole.

Over the years, the appropriateness of the use of α=.05 has been extensively debated in the literature of management, psychology, sociology and statistics. Current approaches to null hypothesis statistical testing indicates researcher choice of alpha may soon be moot given ongoing dialogue among social scientists and statisticians concerning the usefulness and

appropriateness of NHST. While this controversy is nearly as long-lived as the NHST approach itself (see Morrison & Henkel, 1969 and Kirk, 1996 for reviews), renewed rigorous reevaluation recently surfaced. The Task Force on Statistical Significance was formed in 1996 by the American Psychological Association (APA) Board of Scientific Affairs, with the primary charge to study whether NHST should be phased out. This committee of scholars reached into related scholarly communities by seeking the opinions of many, including the American Educational Research Association (AERA), the American Psychological Society (APS), the Society of Mathematical Psychology (SMP), the American Statistical Association (ASA) officials and members of APA Division 5 (*i.e.*, Evaluation, Testing and Measurement). While the committee declined to recommend a ban on NHST in their 1999 report (see Wilkinson *et al.*, 1999), they did recommend a set of proposed guidelines for statistical methods designed to revise the statistical sections of the 1994 *American Psychological Association Publication Manual*. Recommendations included increased emphasis on the reporting of power and sample size issues, as well as effect size and confidence intervals. Furthermore, researchers are encouraged to interpret effects in ways that reflect the level of credibility, generalization and robustness of the findings. This would include comparing current study confidence intervals to those of previous research to help substantiate the stability across studies. Further signifying the relevance of the NHST issue, papers from APS symposia debating the NHST ban issue appeared in a special section of the January 1997 issue of *Psychological Science*.

Confidence Intervals as an Alternative to NHST

While prominent scholars of the field (*e.g.*, Cohen, 1994) have advised no magical alternative to NHST exists, many suggest that confidence intervals may be more appropriate for much behavioral research (*e.g.*, Hunter, 1997; Loftus, 1996; Loftus & Masson, 1994; Schmidt, 1996). Ideas for application vary, with some suggesting confidence intervals provide data helpful in the next logical step, meta-analysis (Schmidt, 1996; Hunter, 1997), while others suggest plotting confidence intervals to explain relationships visually (Loftus & Masson, 1994). Plotting, however, may become burdensome if more than two or three variables are used.

Perhaps researchers should determine and report the point estimate of effect size, as well as the 95% confidence interval around F_1 - F_2 (or other sample statistics). If this confidence interval does not include zero, then the researcher should feel confident in rejecting the hypothesis. Because the 95% confidence interval directly describes a pattern of population parameters, .05 truly represents the probability that a particular observation falls outside the given range (Loftus, 1995). An additional benefit results from the same unit of measurement employed in the data being used in the confidence interval and point estimates, thus results are easier to interpret and trivial effects easier to uncover (Kirk, 1996). Because an inverse relationship exists between confidence interval width and sample size, the larger the sample, the smaller the confidence interval, which is somewhat akin to statistical power (Cohen, 1994).

Use of confidence intervals is not currently a common methodology in much social science/business research, and when employed is often done so in a non-traditional sense. For instance, Lawless, Bergh and Wilsted (1989) used confidence intervals to determine if firms in the same strategic group had identical capabilities, as determined by those that fell within and outside the 95% confidence interval. Results showed that, with certain variables and within certain groups, up to 85% of the firms fell outside the 95% confidence interval. While Lawless and colleagues used this to support their hypothesis that firms in the same strategic group differed on capability dimensions, a more basic interpretation would be that firms were loosely coupled along particular variables. This unique application of confidence intervals allowed analysis of differences within the strategic group.

Confidence intervals also have application in determining differences across two groups (*i.e.*, if confidence intervals do not include zero, groups can be considered to have different means with 95% confidence). However, use of confidence intervals becomes problematic when more than two groups are identified and overall tests of significance are desired in that no known analogous procedure exists for directly comparing means. Thus, while it is simple to compute a confidence interval around the difference between two group means, multiple mean comparisons pose problems. One alternative would have the researcher construct a single-degree-of-freedom contrast and compute a confidence interval around this contrast. As such, a linear contrast is in some sense a generalization of taking a difference between two means (G. Loftus, personal communication, July 31, 1997). Finally, perhaps reporting confidence intervals in pairwise comparisons may be a prudent compromise for researchers employing ANOVA analysis.

Whereas one group of researchers advocates the outright ban on NHST, another suggests its usefulness at times, along with the idea that an outright ban is too severe (*e.g.*, Wilkinson *et al.*, 1999; Estes, 1997; Harris, 1997). A compromise to the null hypothesis significance testing-confidence interval contention may be to use null hypothesis testing without the threshold of significance (*i.e.*, alpha) designation in favor of range null hypotheses, such as the effect size is no larger than "X" (Cohen, 1994). Regardless of approach, it seems reasonable to expect some movement away from current NHST practices. In fact, researchers may be able to more readily uncover particular relationships if appropriate statistical techniques are considered (Wilcox, 1998). A brief look at a strategic management research application shows relevant potential differences in theoretical advances, and hence the streams of research deemed fruitful for further exploration.

Alpha, Null Hypothesis Statistical Testing, Confidence Intervals and the Development of Strategic Management Literature

As scientists, our primary function revolves around development of verified or verifiable systematic theory. As such, a look at theory building fundamentals indicates that a "one-size-fits-all" approach to alpha selection may be problematic. In general, theory building research strives to determine: 1) if a relationship exists, 2) if so, to what degree, and 3) the predicted antecedents and outcomes of this relationship, hence facilitating application. Researchers should initially establish the relationship and then proceed to explore variables that

potentially account for a substantial portion of the variance (*i.e.*, a large effect), while incrementally seeking substantiation of more fine-grained relationships (*i.e.*, smaller effects). Thus, as fundamental relationships are understood and we move more toward calculative approximations, a model that can more accurately predict is desirable and higher standards are reasonable (Fern & Monroe, 1996).

A prime example of how statistical choices influence development, or potentially lack thereof, of research fields can be found in the configuration/strategic group stream of strategic management research. Although configurations/strategic groups have contributed greatly to our understanding of strategic management in general, some (*e.g.*, Barney & Hoskisson, 1990) have questioned its position as an important stream of research. However, criticisms leveled may be more related to precarious statistical methodology than to lack of substantial contributions to the theoretical development of the strategic management field. A closer look at the theoretical development of configurations/strategic groups reveals an example of why business researchers should examine the hypothesis testing versus confidence interval debate, along with related issues, such as ranges of alpha selection and statistical power.

Strategic groups are part of the broader organizational configurations epistemology. Organizational configurations (hereafter referred to simply as configurations) can be defined as clusters of strategic, structural and procedural attributes commonly occurring across organizations (Miller, 1987; Miller & Mintzberg, 1983; Mintzberg, 1990). Analysis of configurations can provide rich descriptions of organizations that reveal their complex, gestalt and systematic nature (Miller & Friesen, 1978). For example, Miles and Snow (1978) identified four distinct profiles (defenders, analyzers, prospectors and reactors), each with a unique combination of structure, decision-making processes, market approaches and performance potential.

The epistemological development of configurational literature suggests that perhaps we should consider segmentation of the configurations-performance literature (see Ferguson & Ketchen, 1999, for a recent review), especially when considering the appropriate alpha to use in null hypothesis statistical testing. Some earlier more exploratory research should conceivably be held to less stringent standards than current research. Perhaps early researchers should have considered an alpha of .10 acceptable, allowing easier establishment of relationships with large effects, generally the first relationships of interest in young research streams. However, current configurations research has developed beyond exploration and initial theory testing in that it regularly explores relationships with small effects, and actively seeks predictions, such as performance. Therefore, it is reasonable to expect more recent research to have lower alphas than earlier research. However, rather than using less stringent alpha benchmarks in earlier research and more stringent ones later, the traditional alpha = .05 benchmark appears to have been used consistently across the research. Results indicate some research supported the configurations-performance relationship (*e.g.* Hawes & Crittenden, 1984; Oster, 1982), while others reported no significant relationship (*e.g.*, Dowling & Ruefli, 1992; Porter, 1979).

The equivocality of results concerning the configurations-performance relationship has lead some to suggest that perhaps researchers should abandon the concept, with research efforts redirected toward other potentially viable determinants of performance (Barney & Hoskisson, 1990). However, when the role of statistical power—a concept directly related to alpha levels—was considered, only approximately eight percent of statistical tests had samples large enough to detect all important relationships, both large and small (Ferguson & Ketchen, 1999). This finding supports the contention that researcher choice of particular alpha levels may indeed be responsible for the equivocality of extant research results concerning the configurations-performance relationship, and indirectly the ultimate potentially ill-advised call by some to abandon this particular stream of research.

The chosen alpha level cannot be interpreted as a measure of support or evidence for a particular hypothesized relationship, only a test of whether the relationship is significant at that particular alpha level (Schervish, 1996). Hence, researchers should consider reporting confidence intervals so that reviewers, editors and other researchers interested in the tested relationship could judge for themselves whether the relationship is worthy of future research. In the above-mentioned configurations-performance relationship case, strategic researchers were discussing the demise of this particular stream of research because it seemingly lacked merit (from an alpha-value point of view), when in fact it is fruitful from a substantive point of view in that it can deepen our understanding of the influences on performance. If researchers would have reported confidence intervals in all published research, the perhaps we would have had a clearer picture of the level of support for the proposed relationship. As shown by the previous example the phenomena under study may indeed be there, but methods may not have allowed it to be teased out (Ferguson & Ketchen, 1999). Hence, a look at the theoretical development of configurations research reveals an example of why business researchers should examine all parameters of statistical testing, including the hypothesis testing versus confidence interval debate, along with related issues, such as ranges of alpha selection and statistical power.

CONCLUSION

While NHST has been a critical tool in most researchers' tool bags for the better part of this century, it has nonetheless been subject to major criticisms. Such criticisms include 1) scientific inference and null hypothesis significance testing address different questions, 2) the null hypothesis is generally always false, so we gain little by "proving" it false and 3) a continuum of uncertainty is assessed dichotomously when a fixed significance level is utilized (Kirk, 1996). In addition, many researchers are misinformed about the interpretation of and benefits from NHST (Hunter, 1997; Schmidt, 1996). Consequently, published research and textbooks are teeming with examples of inaccurately interpreted NHST results (Falk & Greenbaum, 1995; Cohen, 1994; Dar, Serlin, & Omer, 1994). Additionally, although both Fisher (1925, 1926) and Neyman-Pearson (1928) approaches espoused researcher judgment as critical to decision making, NHST binary decision making as now practiced ignores this component, and hence appears not suitable for many complex behavioral research problems (Loftus & Masson, 1994).

Statistical significance as heralded by the all-important researcher finding of $p<.05$ has been a guiding force for many researchers and users of their work for decades. However, a more important contribution than statistical significance may be found in practical significance (*i.e.*, results more useful in the real world) in that researchers should be encouraged to decide if the data support the scientific hypothesis (Kirk, 1996). There is a correspondence between the alpha level in NHST and the probability level (*i.e.*, .05) of a confidence interval, although they entail different logic. The alpha in NHST tests the plausibility of a data set given some null hypothesis, while the probability level directly describes a pattern of population parameters (Loftus, 1995). Perhaps in making choices of various research parameters—whether alpha level, the appropriateness of null hypothesis testing or the application of confidence intervals—researchers might consider the profoundly summarizing statement of Lykken (1968, p. 159-160):

> The value of any research can be determined not from the statistical results, but only by skilled subjective evaluation of the coherence and reasonableness of the theory, the degree of experimental control employed, the sophistication of the measuring techniques, the scientific or practical importance of the phenomena studied, and so on.

This simple, eloquent statement from more than 30 years ago still rings true and offers substantial researcher guidance. While the NHST/CI controversy continues, researchers should at least consider the methodological implications on development of their particular research streams and fields of interest. The logic of the above statement as well as what each researcher could or should do relative to the use of alpha and NHST/CI is also subject to review. If NHST continues to be considered a useful researcher tool, it is possible that debate will fuel a move to include confidence intervals in expanded reporting of results. This could serve as a realistic approach to assessment of any research results, as well as allow more room for researcher and practitioner interpretation of results. Such knowledge may indeed more accurately help describe or reflect upon the nature of the phenomenon under study. Choice of the wrong alpha level may spur underpowered research that incorrectly accepts a false null hypothesis, which can be misleading, contradictory or erroneous. By knowing the background behind the selection of an appropriate alpha, and the resultant influences on the usefulness of hypothesis testing and confidence intervals, perhaps management historians and researchers will be in a better position to influence an appropriate course of action in the NHST/CI debate, and its resultant influence on theory development.

REFERENCES

Barney, J. B., & Hoskisson, R. E. (1990). Strategic groups: Untested assertions and research positions. *Managerial and Decision Economics, 11,* 187-198.

Bernoulli, J. (1713). *Ars Conjectandi* (The Art of Conjecture), as described in Chapter 7 of Bernstein (1996), 116-134.

Bernstein, P. L. (1996). *Against the Gods: The Remarkable Story of Risk.* New York: John Wiley & Sons.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist, 45,* 1304-1312.

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

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62,* 75-82.

DeMoivre, A. (1738). *The Doctrine of Chances* (2nd Ed.), as cited in Bernstein (1996), 126-129.

Dowling, M.J., & Ruefli, T. W. (1992). Technological innovations as a gateway to entry: The case of the telecommunications equipment industry. *Research Policy, 21,* 63-77.

Estes, W. K. (1997). Significance testing in psychological research: Some persisting issues. *Psychological Science, 8(1),* 18-19.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard. *Theory & Psychology, 5(1),* 75-98.

Ferguson, T. D., & Ketchen, D. J., Jr. (1999). Organizational configurations and performance: The role of statistical power in extant research. *Strategic Management Journal, 20(4),* 385-395.

Ferguson, W.L., Barrese, J., & Levy, D.T. (1998). The cost of biased insurer ratings. *Journal of Insurance Issues, 21(2),* 138-150.

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research, 23,* 89-105.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinborough: Oliver and Boyd.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture, 33,* 503-513.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues.* Hillsdale, NJ: Erlbaum, 311-339.

Harris, R. J. (1997). Significance tests have their place. *Psychological Science, 8(1),* 8-11.

Hawes, J. M., & Crittenden, W.F. (1984). A taxonomy of competitive retail strategies. *Strategic Management Journal, 5(3),* 275-287.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8(1),* 3-7.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56(5),* 746-759.

Lawless, M. W., Bergh, D. D., & Wilsted, W. D. (1989). Performance variation among strategic group members: An examination of individual firm capability. *Journal of Management, 15,* 619-661.

Loftus, G. R. (1995). Data analysis as insight: Reply to Morrison and Weaver. *Behavior Research Methods, Instruments & Computers, 25,* 57-59.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5(6),* 161-171.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subjects design. *Psychonomic Bulletin and Review, 1,* 476-490.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151-160.

Magid, A., Mazen, A., Hemmasi, M., & Lewis, M. F. (1987). Assessment of statistical power in contemporary strategy research. *Strategic Management Journal, 8,* 385-397.

Miles, R. E., & Snow, C. C. (1978). *Organizational Strategy, Structure and Process.* New York: McGraw Hill.

Miller, A. (1987). The genesis of configurations. *Academy of Management Review, 12,* 686-701.

Miller, D., & Friesen, P. H. (1978). Archtypes of strategy formulation. *Management Science, 24(9),* 921-933.

Miller, A., & Mintzberg, H. (1983). The case for configuration. In G. Morgan (ed.), *Beyond method: strategies for social research.* Newbury Park, CA: Sage Publications, 57-73.

Mintzberg, H. (1990). Strategic formation: Ten schools of thought. In J. Fredrickson (ed.), *Perspectives on Strategic Management*. Cambridge, MA: Ballinger, 105-235.

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.

Morrison, D. E., & Henkel, R. E. (1969). Significance tests reconsidered. *The American Sociologist, May,* 131-143.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 20,* 175-240, 263-294.

Oster, S. (1982). Intraindustry structure and the ease of strategic change. *Review of Economics and Statistics, 64,* 376-384.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. New York: Harper & Row.

Porter, M. E. (1979). The structure within industries and companies' performance. *Review of Economics and Statistics, 61,* 214-227.

Sauley, K. S., & Bedeian, A. G. (1989). .05: A case of the tail wagging the distribution. *Journal of Management, 15(2),* 335-344.

Schervish, M. J. (1996). P values: What they are and what they are not. *American Statistician, 50(3),* 203-206.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115-129.

Skipper, J. K., Jr., Guenther, A. L., & Nass, G. (1967). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist, 2,* 16-18.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical techniques? *American Psychologist, 53(3),* 300-314.

Wilkinson, L., & The Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54(8),* 594-604.

Winer, B. J. (1962). *Statistical principles in experimental design.* New York: McGraw-Hill.

Winer, B. J. (1971). *Statistical principles in experimental design,* (2nd ed.). New York: McGraw-Hill.

Yule, G. U., & Kendall, M. G. (1950). *An introduction to the theory of statistics*, (14th ed.). London: Griffin.