The Statistical versus Clinical Significance Debate

Sandra M. LeFort

There is general agreement that tests of statistical significance do not provide information about the clinical significance or practical importance of research results. Yet the concept of clinical significance has received little discussion or debate in the nursing literature. The purpose of this paper is to compare and contrast statistical and clinical significance, to provide an overview of the issues surrounding their uses as described in the methodological literature from a variety of disciplines, and to discuss some implications of these concepts for nursing research.

[Keywords: outcomes; methodological research; statistics]

* * *

In the past 20 years, there has been growing recognition that relying on statistical significance testing as the yardstick by which research results are evaluated may have little relevance from a clinical perspective (Jacobson & Truax, 1991). In many health-related and social science disciplines, this recognition has led to a debate about issues concerning the proper interpretation and limitations of tests of statistical significance; the definition of clinical significance; approaches to its assessment; whether measures of association strength, such as effect size and percent of variance accounted for, are the same as clinical significance; the kinds of studies that should address clinical significance; and other issues such as the sensitivity of measurement tools to discern small but clinically meaningful change.

While approaches to statistical significance have been clearly defined, the meaning of clinical significance remains vague. Generally speaking, statistical significance refers to whether an observed difference is more likely to be a real difference rather than a chance occurrence. However, tests of statistical significance do not provide any information about the importance or meaningfulness of research findings (Barlow, 1981). Clinical or substantive significance, on the other hand, relates precisely to the question of importance or potential importance of a finding to the clinical population.

Recent Interest in Clinical Significance

The recent interest in clinical significance appears to be in response to three issues. First, a number of commentaries have discussed the under use of research findings in the clinical setting (Barlow, 1981; Mercer, 1984). In part, this is attributed to the way results are reported particularly in outcome studies (Jacobson, Follette & Revenstorf, 1984). Reports of summary statistics, such as group means and standard deviations alone, do not provide sufficient information for a clinician to determine whether a particular approach might be helpful for specific clients. Secondly, the current climate of accountability in health care requires researchers to demonstrate the efficacy of clinical interventions

Sandra M. LeFort, RN, MN, is a doctoral candidate in the School of Nursing at McGill University. The author gratefully acknowledges the contributions of Laurie Gottlieb, RN, PhD, Associate Professor, McGill School of Nursing, for reviewing an earlier draft of this paper, and Mary Ellen Jeans, RN, PhD, Director, School of Nursing, McGill University for first calling my attention to the importance of clinical significance. The research for this paper was supported in part by the National Health Research and Development Program, Health and Welfare Canada through a National Health Fellowship. Correspondence to School of Nursing, McGill University, Wilson Hall, 3506 University Street, Montreal, Quebec H3A 2A7 Canada.

Accepted for publication February 10, 1992.

(i.e., the extent of change, whether the change makes a real difference to subjects' lives, how long the effect lasts, consumer acceptability, cost-effectiveness and ease of implementation). Thirdly, epidemiology research is shifting from the measurement of disease indicators such as morbidity and mortality to the measurement of health status indicators such as quality of life. This shift from gross indicators of disease to self-report measures of health has meant that measurement tools need to be much more sensitive to small but meaningful changes, a property called "tool responsiveness" (Jaeschke, Singer & Guyatt, 1989).

Recognition of these issues has stimulated a growing body of literature on statistical versus clinical significance in a number of practice disciplines, notably outcome research in psychotherapy (Jacobson & Truax, 1991), behavioral therapy (Hugdahl & Ost, 1981), family therapy (Jacobson, 1985), behavioral medicine (Blanchard & Schwarz, 1988), epidemiology (Jaeschke et al., 1989), rehabilitation (Ottenbacher, Johnson & Hojem, 1988), and education (Carver, 1978). Having acknowledged the limitations of statistical significance testing, these fields are attempting to define and assess clinical significance and clinically significant change. Correspondingly, a number of authors writing in nursing journals have acknowledged that traditional statistical significance testing is not an indicator of clinical importance (Chinn, 1985; Cobb, 1985; Mercer, 1984; Slakter, Wu & Suzuki-Slakter, 1991). However, there has been relatively little discussion about the concept of clinical significance within the nursing literature.

Statistical Significance

Statistical significance is a term indicating that the results obtained in an analysis of sample data are unlikely to be due to chance at some specified level of probability given the truth of the null hypothesis (Jaeger, 1983). By convention and quite arbitrarily, the level of probability of Type One error (alpha or p level) chosen by most researchers is .05 or .01; the determination of statistical significance is achieved by hypothesis testing. Typically, researchers develop two hypotheses, the null and an alternate. The null hypothesis states that no relationship (or no difference) exists between groups on variables of interest and any relationship that may be observed is due to chance or to sampling fluctuations. The alternate hypothesis states that a relationship or difference does exist, is not due to chance and is real. Belief in the validity of the null hypothesis continues unless evidence collected from the study sample is sufficient to make continued belief appear unreasonable (Jaeger, 1983). The logic of hypothesis testing does not allow us to prove the truth of the null hypothesis, let alone the alternate. One can only reject or fail to reject the null. Therefore, scientific knowledge is obtained not by proving a hypothesis true but by finding that it cannot be proven false (Henkel, 1976). The importance of replication studies that build a body of evidence to support a particular hypothesis is clearly paramount for knowledge development in a field.

Tests of significance have become an integral part of the research process in most empirically oriented disciplines and nursing is no exception. The reason for the great popularity of tests of significance is that they complement the scientific method and offer "a deterministic scheme, mechanical and objective, independent of content, that lead to clear-cut yes-no decisions" (Cohen, 1990, p. 1307). Nonetheless, statistical significance testing has limitations and has been widely criticized in the literature (Carver, 1978; Cohen, 1990). Before reviewing the major criticisms, however, it is important to distinguish between the two existing approaches to statistical significance testing.

Approaches to Statistical Inference

Historically, two distinct schools of thought regarding statistical inference have developed: the traditional approach introduced by Fisher in 1925, and the Neyman-Pearson approach first discussed in 1928, and brought to wider attention by Cohen (1977). Both the Fisherian and the Neyman-Pearson approaches follow the same basic logic as previously described and both posit an exact null hypothesis (i.e., the observed relationship or the observed difference between groups equals zero). The point of departure for the two approaches rests on the statement of the alternate hypothesis and issues relating to the power of the test.

In the Fisherian tradition, the alternate hypothesis is an inexact hypothesis that specifies a range of values for a parameter—any value other than zero (Cohen, 1977; 1990). Neither the strength of the observed relationship nor the magnitude of the difference between two means is at issue; the only issue is whether the difference is more likely to be real rather than due to chance variation. Thus, even trivial or minute differences may be statistically significant (Carver, 1978).

By contrast, a researcher using the Neyman-Pearson approach poses a *specific* non-zero value for the alternate hypothesis (Cohen, 1977). The selection of this specific value, termed effect size, is based on the researcher's understanding and knowledge of the field under investigation, supplemented by evidence from literature reviews and pilot studies. Thus, the alternate hypothesis specifies the *exact* size of the effect the research study is designed to detect. Since the publication of Cohen's work on power analysis in 1977, appropriate sample size, p level and power can be determined once an effect size has been estimated. This expanded form of statistical inference, then, answers two questions: (1) Are the results likely to be real or due to chance? (2) Are the results large enough to be considered non-trivial? Thus, unlike Fisherian hypothesis testing, the Neyman-Pearson approach provides additional information on the actual size of the effect observed.

Critique of Statistical Significance Testing

Critiques of statistical significance testing have been aimed almost exclusively at the Fisherian approach (Chow, 1988). This probably reflects the fact that, despite the recommendations in the literature to consider power, effect size and sample size, researchers in many fields including nursing continue to use the Fisherian approach (Polit & Sherman, 1990; Sedlmeier & Gigerenzer, 1989). The major criticisms of Fisherian statistical significance testing can be summarized in the following five areas:

Sample Size and Variability: Results of Fisherian statistical significance testing are highly influenced by sample size and by variability within the sample (Chow, 1988; Garfield, 1978). Studies with large samples are much more likely to lead to statistically significant results compared to identical studies with small samples. Similarly, results from homogeneous samples are more likely to be statistically significant than are results from a sample with greater variability. (See Mercer, 1984, p.49-50 for a discussion of the sample size/variability problem in a nursing research example).

Effect Size: As previously described, the use of Fisherian statistical significance testing leads to the loss of valuable information relating to the size of an effect. Often, readers of research reports erroneously assume that small p values necessarily imply large effects (Smith, 1983). However, small p values indicate only that results are less likely to have occurred by chance given the null hypothesis is true; they do not indicate the *actual* size of the observed effect.

Clinical Significance: There is general agreement that the importance of a finding is not addressed by Fisherian statistical significance testing, and some authors contend that it is not addressed by the Neyman-Pearson approach either (O'Grady, 1982). Being confident that a finding is unlikely to be due to chance, and being confident that it is of a certain magnitude does not necessarily provide information regarding the clinical significance of research outcomes. (See Jacobson, 1985, p.150-151 for an illustrative example).

Incompatibility with Health and Behavioral Sciences: Some authors argue that the accept/reject decision-making approach in statistical significance testing is an inappropriate basis for the development of knowledge in the human sciences (Cohen, 1990). Fisher developed null hypothesis testing to answer questions related to agricultural research that required yes/no decisions. By contrast, questions relating to social, psychological, developmental, and environmental variables that affect health are quite different. The *degree* to which such variables affect health is often more important than whether or not the variable is present or absent. Questions of degree are not answered by traditional Fisherian null hypothesis testing.

General Objections: There are a number of general philosophical objections to the Fisherian procedure itself. For example, authors condemn the mechanistic approach to decision-making that they feel is inherent in the statistical significance testing procedure (Henkel, 1976). They contend that mature scientific judgment and creative theoretical thinking are superseded by the dichotomous nature of this approach. Another criticism is that the choice of p level is not based on any theoretical notions, but is quite arbitrary. As Rosnow and Rosenthal state (no doubt, with tongue in cheek): "we wish to emphasize that dichotomous significance testing has no ontological basis. That is, we want to underscore that, surely, God loves the .06 nearly as much as the .05" (1989, p. 1277). Finally, most researchers would agree that scientific knowledge is cumulative. Yet each test of the null hypothesis is an independent, isolated event that does not take prior evidence of the strength of a hypothesis into account as does Bayesian statistics (Goodman & Royall, 1988).

In defense of Fisher's approach, the reader is referred to Chow (1988) who has refuted many of these criticisms in detail. Of particular importance here is Chow's discussion of the rationale for research and the appropriate use of Fisherian statistical significance testing. Chow categorizes research in two broad classes: theory testing and descriptive. Research that tests the plausibility of explanatory theories (theory-testing research) provides a basis for choosing between rival assertions. Chow maintains that the accept/reject logic of hypothesis testing fits well with research of this kind. Descriptive research, on the other hand, seeks to describe phenomena of interest in detail and is consequently more concerned with issues of magnitude and whether the observed effect is clinically important. Thus, Chow concedes that traditional hypothesis testing is insufficient for descriptive research.

Using Chow's classification, most nursing research studies are descriptive. Thus, the meaningfulness or importance of findings are of utmost concern and should be *explicitly* addressed in research reports.

Clinical Significance

There is little consensus regarding what substantive or clinical significance is, except the universal agreement that it is not merely statistical significance in the Fisherian sense (Jacobson et al., 1984). This disparity is reflected by the multiple terms used by researchers to describe "needing more than statistical significance" to assess the meaningfulness of research findings. Adjectives used to describe this type of significance include: clinical, substantive, scientific, meaningful, biologic, practical, actual, material, social and applied. Still others consider the concept within the broader context of validity. Although these terms may indicate different variations of the construct of clinical significance, for purposes of this discussion it will be assumed that these terms are more or less conceptually equivalent since many writers appear to use them interchangeably.

Approaches to Clinical Significance

Since there exists no universal agreement about the definition of clinical significance, it is not surprising that there are a number of distinct approaches to its assessment. However, all approaches fall under two general orientations: statistically-based or value-based. A number of authors have cogently argued that although the concept of clinical significance is inherently value-based, statistical tests can be used as tools in its assessment (Hayes & Haas, 1988; Hollon & Flick, 1988). A problem arises when users of statistical approaches do not candidly recognize the inherently value-based nature of the concept and hence do not explicitly discuss the judgments of the researcher that, of course, figure in the decision that a finding is important or clinically significant.

Most current methods of assessing clinical significance can be grouped under one or more of the following four approaches:

Statistical Measures of Magnitude of Effect: Measures of magnitude of effect are ways of estimating the size of the overall group effect or of estimating the overall magnitude of the relationship between the independent and the dependent variable. The most widely used estimates of magnitude of effect are: (a) effect size estimates in the form of standardized scores, and (b)

association strength estimates such as measures of explained variance (i.e., the proportion of dependent variable variance accounted for by an independent variable of interest). Unlike tests of significance, these measures are independent of sample size (Maxwell & Delaney, 1990).

Authors from many disciplines, including nursing, espouse the value of interpreting measures of magnitude of effect as an index of clinical significance (Murray & Dosser, 1987; Polit & Sherman, 1990). However, others caution that they may be misleading or inappropriate indicators of the importance of a finding (Hayes & Haas, 1988). O'Grady (1982) points out that many factors—measurement error, reliability of constructs under investigation, study and sampling design, the number of levels of a treatment factor, the number of variables under study, the type of research question, and the heterogeneity of the population sampled—all affect the percent of variance accounted for in any particular study. In addition, any single variable attempting to explain multi-factorial phenomena, such as health, is likely to explain only a small proportion of the variance. But this is not to say that the variable is unimportant.

Another concern is that both measures beg the question of clinical significance. For example, how much variance would a variable need to explain in order to be considered clinically significant? How big does an effect size need to be for it to be considered important? In an attempt to answer these questions, one proposed solution is that improvement be defined *a priori,* based on consistent findings in the research literature (Blanchard & Schwarz, 1988). For example, some authors suggest that meta-analysis of a large collection of studies in a field could be done to detect the minimal effect size or minimal explained variance considered to be important or clinically significant (Murray & Dosser, 1987). However, this method presupposes the existence of sensitive measurement instruments, a large body of conceptually-related studies based on the same or similar outcome variables, and convergent findings. At present, most areas of nursing research cannot met these criteria (Lynn, 1989). Thus, decisions regarding what effect size or what percent of variance explained will be considered important are based on judgments made by the researcher.

Individual Approaches: Authors have suggested that the meaningfulness of an outcome in group-comparison designs be linked to the individual subject, rather than to the overall group effect (Jacobson & Truax, 1991). Hugdahl and Ost (1981) propose that researchers calculate, report (either in table or graph form) and discuss: (a) the proportion of subjects who show improvement, show no change or show deterioration; and, (b) the degree of change within each individual in order to highlight within-group variability. (See Jacobson et al., 1984, p. 347 for a graphic example and Jacobson and Truax, 1991, p. 16 for a tabular example). It is argued that the proportion of improved subjects is a more valid indicator of clinical significance than are statistically significant group level statistics such as means, standard deviations and effect sizes. In addition, this approach has the potential to be of help to clinicians who are primarily interested in individual subject performance; it might also stimulate researchers to ask specific questions regarding who improves, who doesn't, and why.

In terms of evaluating the importance of change over time (such as in pretest/post-test designs or in longitudinal designs), the major problem with the "individual improvement" approach is that it, too, begs the question of clinical significance. How improvement is defined, how degrees of improvement are measured, and how much of an improvement an individual would have to demonstrate to be considered clinically significant is not made explicit by the major proponents of this approach. Clearly, these are issues of values and judgments.

In health care research, part of the problem in defining how much change constitutes clinical significance relates to issues of measurement. Schumm (1982) suggested that there is little understanding about what differences in scores actually mean from a clinical standpoint. For example, on a measure of marital adjustment with possible scores ranging from zero to 151 units, what does a change of 5.5 units actually mean? Much more work needs to be done in this area of psychometrics.

Normative Approaches: There are a number of researchers who propose that clinically significant change be based on a return to normal functioning (Saunders, Howard & Newman, 1988). In a series of studies relating to psychotherapy outcomes, Jacobson and associates (1984, 1985, 1988, 1991) describe an approach to clinically significant change in which every subject is evaluated regarding whether post-treatment scores can be considered within the distribution of the functional or the dysfunctional populations. A discrete cut-off based on normative data available on the variable being measured is used to define functional and non-functional. In addition, to be certain that the change in function is statistically reliable and not due to measurement error, an index of reliable change (RC index) is also calculated. Any subject moving toward the functional distribution, who is above the defined cut-off point and who shows reliable change is said to have experienced clinically significant change. The developers claim that this two-pronged criterion provides an objective, bias-free, psychometrically sound method for the evaluation of clinically significant change that has the added advantage of applicability to a wide range of clinical problems.

The normalization approach is not without its problems. First, the criteria proposed by Jacobson and colleagues are only as good as the outcome measures available in the field. Secondly, normative data on the functional population is required but adequate norms do not exist for many widely used measurement instruments in nursing. A more fundamental issue is the definition of clinically significant change as being tied to absolute normalization. Saunders et al. (1988) suggest that there are at least three different conceptions of normality: (a) normality as reasonable functioning (e.g., absence of disease or improvement in symptoms); (b) normality as optimal functioning (e.g., self-actualization); and, (c) normality as average functioning (e.g., bell-shaped distribution). Those writing about clinically significant change have opted for the "normality as average" concept. However, this is problematic when the average is not the goal of intervention, or when study populations, such as those with chronic illness, may not have the potential to reach average functioning on a given variable but are, nonetheless, capable of meaningful change. In these cases, what are the appropriate referent groups to determine clinically significant change? It may be the case that an understanding of

normality that is culturally and gender-sensitive and that highlights self-actualization may be a better fit with the values of nursing.

Social Validation: The last major technique of evaluating clinical significance is to ask those who may be in the best position to judge: the clients themselves, significant others who are in the client's natural environment, and clinicians who have experience in the area. This approach is based on the premise that for change to be clinically significant, it must make an obvious qualitative difference in people's lives, and should lead to noticeable improvement in their everyday functioning (Kazdin, 1977) and, I would add, sense of well-being. Thus, the opinions and judgments about the change and the difference it has made in their lives are solicited from the subject and/or others. Disciplines such as psychology have tended to use standardized questionnaires such as global ratings of function to elicit these subjective judgments. However, the use of qualitative methodologies that use in-depth interviewing may prove especially useful in gaining insight into how clients perceive their unique situation and why nursing interventions may or may not have helped.

The social validation approach is the only method to acknowledge unreservedly that clinical significance is ultimately a matter of values, and therefore, what is meaningful or important will depend, in part, on who is asked. Hollon and Flick (1988) contend that there are three constituencies interested in health outcomes:

1. The clients/families themselves who are interested in their everyday functioning and well-being;
2. The clinician/researcher who may have theoretically ascribed or discipline ascribed approaches to what is considered meaningful; and,
3. Society at large, usually interested in maintaining norms, in sustaining current health policy and belief systems, and in other issues, such as cost-effectiveness.

Baer (1988) cautions researchers and clinicians to be very clear about whose definition of clinical significance is being used. Clearly, nursing's focus on caring orients the discipline to choosing the first definition relating to clients and families.

The social validation approach to the assessment of clinical significance seems intuitively appropriate to nursing research. The method emphasizes the importance of context and of the naturalistic setting, both of which are integral to nursing. In addition, nurse researchers have identified the need for multiple methods and multiple data sources to access both the insider and the outsider view, an approach consistent with social validation (Murphy, 1986). The major difficulty in using the approach may relate to the subtlety and complexity of many of the concepts central to nursing such as health. The social validation approach, as with all the approaches to clinical significance discussed in this paper, assumes that clients are pathological or have a problem. Clinically significant change during an illness event may be more readily identifiable than for healthy transitions throughout the life cycle, for example.

Recommendations

A social validation orientation does not dismiss the statistical approaches described. Statistical approaches used to assess clinical significance are not unlike diagnostic tests that serve to confirm what a good clinician already suspects from the history. They are tools to be used intelligently, interpreted in light of the researcher's knowledge and experience to provide further evidence for a position. Thus for nursing, the judgments, opinions and behavior of subjects in addition to observations of key individuals should be the primary evidence confirming clinical significance; results from statistical approaches should provide supportive evidence.

The following are suggestions regarding how statistical approaches to clinical significance might be incorporated into nursing research reports:

1. The provision of descriptive data about study subjects that highlights individual change or important relationships about individual characteristics is a useful, simple and manageable addition to the results section of research reports particularly when the number of study subjects is small.
2. Measures of magnitude of effect can be important estimators of the magnitude of the relationship between variables; however, they need to be interpreted in the context of other data and in light of their limitations. In addition, the researcher must state and justify at what point results are considered clinically significant given the nature of the research question and given nursing's perspective. As Cohen (1977) has pointed out: one discipline's trivial effect size is another discipline's large effect size.
3. Before normative approaches to clinical significance are used in nursing research, more normative data that is gender and culture-specific needs to be collected on variables of interest to nursing.
4. Issues relating to the "clinical" limitations of the measurement tools used in nursing studies need to be addressed in the methodology and discussion sections of research reports.

A final comment relates to research design. Assessing the clinical significance of some nursing interventions may require innovative longitudinal designs with multiple data collection points in order to assess the long-term importance of a change.

References

Baer, D.M. (1988). If you know why you're changing a behavior, you'll know when you've changed it enough. **Behavioral Assessment, 10**, 219-223.

Barlow, D.H. (1981). On the relation of clinical research to clinical practice: Current issues, new directions. **Journal of Consulting and Clinical Psychology, 49**, 147-155.

Blanchard, E.B. & Schwarz, S.P. (1988). Clinically significant changes in behavioral medicine. **Behavioral Assessment, 10**, 171-188.

Carver, R.P. (1978). The case against statistical significance testing. **Harvard Educational Review, 48**, 378-399.

Chinn, P.L. (1985). Debunking myths in nursing theory and research. **Image, 17**, 45-49.

Chow, S.L. (1988). Significance test or effect size? **Psychological Bulletin, 103**, 105-110.

Cobb, E.B. (1985). Planning research studies: An alternative to power analysis. **Nursing Research, 34**, 386-388.

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

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

Garfield, S.L. (1978). Research problems in clinical diagnosis. **Journal of Consulting and Clinical Psychology, 46**, 596-607.

Goodman, S.N. & Royall, R. (1988). Evidence and scientific research. **American Journal of Public Health, 78**, 1568-1574.

Hayes, S.C. & Haas, J.R. (1988). A reevaluation of the concept of clinical significance: Goals, methods, and methodology. **Behavioral Assessment, 10**, 189-196.

Henkel, R.E. (1976). **Tests of significance.** Sage University Paper series on quantitative applications in the social sciences, 07-004. Beverly Hills and London: Sage.

Hollon, S.D. & Flick, S.N. (1988). On the meaning and methods of clinical significance. **Behavioral Assessment, 10**, 197-206.

Hugdahl, K. & Ost, L. (1981). On the difference between statistical and clinical significance. **Behavioral Assessment, 3**, 289-295.

Jacobson, N.S. (1985). Family therapy outcome research: Potential pitfalls and prospects. **Journal of Marital and Family Therapy, 11**, 149-158.

Jacobson, N.S., Follette, W.C. & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. **Behavior Therapy, 15**, 336-352.

Jacobson, N.S. & Revenstorf, D. (1988). Statistics for assessing the clinical significance of psychotherapy techniques: Issues, problems, and new developments. **Behavioral Assessment, 10**, 133-145.

Jacobson, N.S. & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. **Journal of Consulting and Clinical Psychology, 59**, 12-19.

Jaeger, R.M. (1983). **Statistics: A spectator sport.** Newbury Park, CA: Sage.

Jaeschke, R., Singer, J. & Guyatt, G.H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. **Controlled Clinical Trials, 10**, 407-415.

Kazdin, A.E. (1977). Assessing the clinical or applied importance of behavior change through social validation. **Behavior Modification, 1**, 427-451.

Lynn, M.R. (1989). Meta-analysis: Appropriate tool for the integration of nursing research? **Nursing Research, 38**, 302-305.

Maxwell, S.E. & Delaney, H.D. (1990). **Designing experiments and analyzing data: A model comparison perspective.** Belmont: CA: Wadsworth.

Mercer, R.T. (1984). Nursing research: The bridge to excellence in practice. **Image, 16**, 47-51.

Murphy, S. (1986). Family study and nursing research. **Image, 18**, 170-174.

Murray, L.W. & Dosser, D.A. (1987). How significant is a significant difference? Problems with the measurement of magnitude of effect. **Journal of Counseling Psychology, 34**, 68-72.

O'Grady, K.E. (1982). Measures of explained variance: Cautions and limitations. **Psychological Bulletin, 92**, 766-777.

Ottenbacher, K.J., Johnson, M.B. & Hojem, M. (1988). The significance of clinical change and clinical change of significance: Issues and methods. **Occupational Therapy, 42**, 156-163.

Polit, D.F. & Sherman, R.E. (1990). Statistical power in nursing research. **Nursing Research, 39**, 365-369.

Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. **American Psychologist, 44**, 1276-1284.

Saunders, S.M., Howard, K.I. & Newman, F.L. (1988). Evaluating the clinical significance of treatment effects: Norms and normality. **Behavioral Assessment, 10**, 207-218.

Schumm, W.R. (1982). Integrating theory, measurement and data analysis in family studies survey research. **Journal of Marriage and the Family, 44**, 983-998.

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

Slakter, M.J., Wu, Y.B. & Suzuki-Slakter, N.S. (1991). *, **, and ***: Statistical nonsense at the .00000 level. **Nursing Research, 40**, 248-249.

Smith, K. (1983). Tests of significance: Some frequent misunderstandings. **American Journal of Orthopsychiatry, 53**, 315-321.