Methodological Concerns in the Testing of Nursing Interventions

Carol A Deets, R.N., Ed.D.
Associate Professor
Chairperson of Nursing Synthesis
School of Nursing
Indiana University
Indianapolis, Indiana

THREE ISSUES are problematic in the development of rigorous nursing intervention studies; they are problems of methodology, of definition and of theory.

Both the definitional and theory issues are extremely important to the development of nursing interventions. The definitional problems result from a lack of agreement on the meaning of specific concepts and the need to define operationally many of the concepts currently in use, e.g., nursing, professional, clinical. A lack of understanding or valuing of theory seems to be a central cause of the theory problems. Both areas are extremely important and deserve individual consideration.

METHODOLOGICAL PROBLEMS

Certainly definitional and theoretical problems are important, but they are also long-term problems that often need consensus within the profession to resolve. On the other hand, many of the methodo-

0161-9268/80/0022-0001$2.00

logical problems considered in this section are ones that can be dealt with immediately. In fact, if researchers were to resolve some of these methodological problems, thus strengthening the quality of their research of nursing interventions, it may be that their studies would provide the necessary data for resolving the definitional and theoretical problems more rapidly than would occur without this effort.

The attempt to control the variance in a study is the concern for each of the methodological problems discussed in this section. Control of variance in clinical research can be accomplished through the design of the study and use of statistical techniques. The researcher's ability to identify and control extraneous variables will greatly affect the quality of the study's design. The use of complex, sophisticated statistics will result in a more rigorous control of variance, but it is often difficult to interpret the results of these techniques in a meaningful manner. This problem is further aggravated by the reality that few nurses are prepared to use appropriately such statistical procedures. Often the best research design is a simple, well-controlled one.

Campbell and Stanley discussed different types of quasi-experimental designs and classes of extraneous variables that threaten the internal and external validity of quasi-experimental designs.[1] Internal validity of a design refers to the ability of that design to appropriately answer the research question. On the other hand, external validity refers to the ability of the research design to allow generalization of the findings from that study. The researcher needs to consider the theoretical framework and its implications plus previous research findings and the Campbell and Stanley classes of extraneous variables when identifying potential threats to the design's internal and external validity. Since Campbell and Stanley's work is relatively well known in nursing, the classes of extraneous variables will be used as an organizer for this section.

Major advances have been made in controlling some of Campbell and Stanley's variables. For others such as "history," "maturation" and "mortality," the researcher's ability to use common sense and to develop rigorous research designs are still the best means for reducing these threats. It is also possible to control maturation using several statistical techniques, primarily time series analysis (cf. Box & Jenkins,[2] Anderson,[3] Glass,[4] Bloomfield[5]).

Testing

Campbell and Stanley indicated that a major problem with quasi-experimental designs was the use of pretest and posttest examinations. It was their contention that a sensitizing effect would occur as the result of taking the pretest. Sensitizing could occur because the test taker was cued to the importance of specific content by the pretest items or by the mere fact that the test taker retained information learned at pretesting to posttesting. Certainly it is possible that some biasing effect or cueing could occur when the posttest is exactly the same as the pretest and administered on the same day as the pretest. When time elapses between the pretests and posttests the retention factor is not so operative (cf. Lana,[6] Bracht &

Glass,[7] Welch & Walberg[8]). If objectives for the learning sequence are used there would be as much cueing of important points by the reading of the objectives as would occur by responding to the pretest items. It is possible to reduce the possibility of retaining information from pretesting to posttesting by allowing sufficient time between the two tests or by creating parallel pretests and posttests. Parallel tests are ones that measure the same phenomenon using equally difficult items that have different wording to measure the same objective. In addition, the means and standard deviations for the two tests must be equal.

Through the use of such techniques as these the art of developing and administrating pretest and posttest has improved considerably in the last 15 years. These improvements have greatly reduced the need for concern originally expressed by Campbell and Stanley. The value of being able to determine what an individual knows or what his or her skills are before and after an intervention is an excellent means for determining the success of the intervention. The ability to document improvement is of greater value than any detriment that could occur from using a pretest when appropriate safeguards are taken.

The art of developing and administrating pretest and posttest has improved considerably in the last 15 years, greatly reducing the need for concern originally expressed by Campbell and Stanley.

Throughout their work, Campbell and Stanley referred to a phenomenon called change or gain score. This score is created by subtracting an individual's pretest score from the posttest score. Included in this new score is the measurement error from both tests that results in a different reliability for the change scores than for either the pretest or posttest measures. It should be noted that the potential is for either an artificial increase or decrease in the reliability, and one never knows which actually occurred. Campbell and Stanley suggested using the change or gain score as the dependent variable in many of their designs. However, most authors agree that using either a correlated *t* test or the analysis of variance with repeated measures is the more appropriate technique for analyzing pretest and posttest scores.[9-13]

Instrumentation

A closer reading of Campbell and Stanley's examples in this area indicates that these examples were all concerned with some type of reliability. For example, they cautioned that just because equipment was used, consistency in measurement could not be assumed. This aspect of their concern for reliability is even more important today when nurse researchers are using sophisticated bioinstrumentation that they may not be able to adequately operate (cf. Abbey,[14] Huether,[15] Jacobs,[16] Weeks[17]).

Most of the instrumentation examples in Campbell and Stanley dealt with some aspect of observer agreement. Since observation is one means of data collection

frequently employed by the nurse researcher, this type of reliability is most important. The usual approach is for the researcher to compute the percent of agreement in ratings between observers. However, the percent of agreement may reflect agreement in measuring error rather than the phenomenon that was to have been measured. Most authors believe that agreement among observers in not so critical as consistency in the measurement of the phenomenon (cf. Cronbach, et al.,[18] McGraw, Wardrop & Bunda,[19] Herbert & Attridge,[20] Rowley,[21] Frick & Semmel[22]).

An alternative approach to the percent of agreement technique is to compute interrater or interjudge reliability. This approach deals with amount of variation among the raters or judges and determines whether the difference in the amount of variance is sufficiently great to have occurred by chance. Considerable accomplishments have occurred in the last few years that allow the researcher to estimate the reliability of the observed measures more accurately. This area is far more complex than simply determining the percent of agreement among judges or raters, and an in-depth analysis of possible techniques to use when determining the reliability of data collected through observations should be undertaken whenever that technique is to be employed.

All the previous comments concerning reliability are as relevant—perhaps even more so—when applied to paper and pencil tools. It is beyond the scope of this article to discuss the pros and cons of obtaining reliability coefficients for internal consistency, stability and equivalence (cf. Ebel,[23] Mehrens & Lehman[24]). Nonetheless, these should be concerns of the researcher when using this type of tool. Recently the nursing literature has reflected an increased interest in the quality of tools (cf. Fitzpatrick & Donoran,[25] Kearney & Fleischer,[26] Ward & Fetler,[27] Ward & Lindemann[28]). These researchers and consumers of nurse research must learn that statistically significant differences for data collected by an unreliable tool may be due to error and not the treatment.

Statistical Regression

In most distributions of scores there are those scores that are at either extreme of the continuum, i.e., beyond three standard deviations. The phenomenon statistical regression is concerned with these extreme scores. If an individual who has scored at one extreme of a distribution is retested there is a high probability that his second score will be considerably closer to the mean than his first score. In other words, in an experiment it would be possible for an individual's score to be changed as a result of the regression effect and not of the treatment effect.

Statistical regression is influenced by the reliability of the instrument used to measure the subjects. As Campbell and Stanley indicated, the lower the reliability of the measure the greater the problem of statistical regression. Probably the best means for controlling this variable is to be alert for instances when it could occur and then take preventive action as prescribed by Campbell and Stanley.

Differential Selection

A basic assumption of inferential statistics is that the sample being included in

the experiment was randomly selected and, ideally, randomly assigned to the treatment or control groups (or both). In the real world of nursing research this is seldom a realistic expectation. The alternative is a convenience sample. Even when the sample is selected by including every second or tenth person, it still remains a convenience sample and not a random sample. By definition, a convenience sample is a differential selection process. In other words, some specific variable differentiates those subjects who are or are not included in the sample.

The purpose of randomly selecting the sample and then randomly assigning subjects to groups is to ensure equivalence of groups and to permit the use of inferential statistics. When a convenience sample is used and equivalence of groups cannot be assumed, variables that are critical descriptors of the population can be statistically tested to determine if the treatment group differs from the population or the control group on these variables. Finding no significant differences among the groups may not place the researcher in so strong a position as having randomly selected and/or assigned the subjects to the groups, but the groups are statistically equivalent. Of course, the researcher cannot generalize to a representative population, but the internal validity of the design is no longer threatened by this concern and the use of inferential statistics is appropriate (cf. Edgington,[29] Glass & Stanley,[30] Slakter,[31] Wooldridge, Leonard & Skipper,[32] Polit & Hungler[33]).

One means for resolving the problem of generalizing to a representative population when forced to use convenient samples is the use of replications. Although seldom used in nursing, this approach can result in the demonstration of the consistency of findings across many samples, thus indicating the generalizability of the findings.[30,32,34] One hopes the value system of faculty in graduate programs will shift from emphasizing the importance of *new* information to emphasizing the importance of *reliable* information so that frequency of replications through master's studies will increase.

Reactive Arrangements

Often subjects who discover they are participating in an experiment will react to that fact. They will feel that because they have been selected they are special. This reaction is what Campbell and Stanley referred to as reactive arrangements. This response by subjects is also known as the Hawthorne or Halo effect. No matter how reliable and valid the tool used, it cannot combat this particular reaction.

The need to secure informed consent from each subject has increased the need for considering this threat. The concern in controlling reactive responses is that the subjects would respond to the attention provided by the study, thus creating an effect on the dependent variable in relation to the independent variable. One means for controlling this factor has been not to tell the subjects that they were participating in a study. While this approach is still possible, the committee for protecting human subjects would need to be convinced of the necessity for deception. In reality few studies are conducted that do not obtain informed consent from each participant prior to the study.

Data are not available that provide insight into this problem. The practice and

ethical aspects of informed consent are frequently discussed but seldom from the viewpoint of informed consent functioning as an intervening variable. Studies in this area should be conducted in the near future.

Multiple Treatment

A multiple treatment effect occurs whenever a proceeding treatment may affect a subject's later treatment. For example, if a client learns about his condition through a slide tape presentation and then is exposed to a learning sequence conducted by a nurse, it is impossible to determine which of the two treatments resulted in the client's learning. As nursing becomes more concerned with the effectiveness of the treatment and the retention factor related to that treatment, the problem of multiple treatments will become more complex. It is insufficient for the client to know enough about his condition to respond to simple questions; it is essential that he be able to retain this information and use it in his home setting. A legitimate multiple treatment question should be, "Would one interaction with a client every week over a period of six weeks be superior to the concentrated interaction with a client over a shorter period of time?" To answer this question would require complex, sophisticated statistical techniques. The use of repeated measures or split-plot designs provides the researcher with statistical techniques to assist in identifying variance due to specific factors;[34-36] however neither technique resolves the problem. The researcher must consider carefully the importance of multiple treatments and justify their use keeping in mind that it may not be possible to identify which treatment produced the desired effect.

Interaction of Variables

While controlling the different classes of variables posited by Campbell and Stanley is difficult, perhaps the most important problem is the reality that these extraneous variables also have the potential for interacting with an experiment's treatment effect. On the basis of recent studies concerning the sensitizing effect of the pretest, it has been found that the probability of the pretest interacting with the treatment is reduced, but not eliminated. A rigorous researcher will seriously consider each extraneous variable and its control. In addition, the researcher must evaluate the possibility of that variable's interacting with the treatment and decide how this could be controlled.

Designs

The designs included in Campbell and Stanley were used to illustrate the importance of considering extraneous variables. These designs are still being taught in nursing research classes as possible designs for use in nursing. Advances have been made in the area of research design such as aptitude/trait-treatment interactions (cf. Berliner & Cahen,[37] Borich, Godbout & Wunderlick,[38] Snow,[39] Cronbach & Snow[40]) and should be considered in nursing research classes.

Researchers need to create designs that answer their research questions rather than searching for designs and altering their questions to fit the designs. It was not Campbell and Stanley's intent to create a

Researchers need to create designs that answer their research questions rather than searching for designs and altering their questions to fit the designs.

design cookbook that novice researchers could use in place of rational, careful consideration of their theoretical framework, possible extraneous variables and knowledge of design. Unfortunately, the selection of the design before the question has been completely articulated has occurred all too often.

Conspicuously missing in Campbell and Stanley's work is any reference to crossed versus nested variables or random versus fixed variables. The ability to incorporate nested variables in a design allows nurse researchers to more accurately represent the clinical reality. For instance, a researcher finds that to obtain a sufficient sample several hospitals must be included in the study. In addition to the design needed to answer the question, the researcher needs to control for the fact that variance is added to the design as a result of these hospitals. In addition it is likely that the subjects will be selected from different floors or units. The researcher now needs to control for the variable units. The variable units are nested in the variable hospitals, i.e., unit A in one hospital is not the same as unit A in the other hospital. (For more specific information see Glass & Stanley,[30] Winer,[34] Huynh & Feldt,[36] Stoloff,[41] Poynor[42].)

The problem of whether a variable is random or fixed influences the "expected mean square" for that variable. When one calculates the "expected mean square" it is to determine which term is to be used in the *F* ratio's denominator. This means that the denominator in the ratio for a random variable might be quite different from the usual denominator (error term) for the fixed factor. In fact, it is possible that this particular variable would be tested using as its denominator one of the interaction terms. Since the smaller the denominator the higher the probability of significant differences and because usually an interactive effect is less than the error effect, the use of the interaction term in the denominator increases the probability of significance.[31]

Other Considerations

Even if all of these points are considered by the researcher there is no guarantee that the clinical research study will be rigorous. There are additional problems that must be considered. For instance, nurses are using more and more *criterion-referenced* tools. A criterion-referenced tool is one in which the expected level of expertise by the test taker has been established in advance by the test maker. In other words, if an individual has sufficiently mastered the material presented he should miss only a preset number of items in the tool. Often the criterion is 5% or 10% of the items in that tool. What effect do these types of tools have on the statistical tests used? What influence does the arbitrary cutoff score have in interpreting data, especially if the cutoff point is too high or too low? How does one calculate the reliability of a criterion-referenced tool? Possible answers to these questions are just beginning to

appear in the literature (cf. Meskauskas,[43] Subkoviak & Baker,[44] Hambleton, et al.[45])

The sensitivity of the instrument used to measure the variable is another important consideration. Assume a theory indicates that the galvanic skin response (GSR), blood pressure (BP), pulse and other physical indicators measure anxiety levels (cf. Wang,[46] Forbes,[47] Edelberg,[48] Prokasky & Raskin,[49] Seligman[50]). Most of the techniques readily available to nurses for measuring BP, pulse and other physiological indicators are indirect measures and indicate only gross changes. The researcher is in the position of not being able to demonstrate whether or not the treatment really functioned because the change could be so slight the measurement tools were insensitive to the changes.

The power of a statistical test is seldom reported or discussed in the nursing literature. Yet this information is critical in assessing the value of the findings. It is possible to obtain statistically significant differences yet account for only 10% to 12% of the variance between the dependent and independent variables. There is little value of creating or using an intervention on the strength of these findings. The power of a test can be increased, and anyone interested in more complicated statistics should be conversant with the problems of power (cf. Kennedy,[51] Halderson & Glasnapp,[52] Cohen,[53,54] Meyer,[55] Brewer[56]).

Another consideration that must be addressed is the value of a control group. It is true that in many situations one is needed. It is also possible that extraneous variables influence the control group to the extent that it no longer fulfills its purpose. There are some research questions that may not need a control group. When the question deals with the meeting of educational objectives, a control group may or may not be needed. Some authors contend that it is safe to assume that without some type of educational intervention the control group would remain unchanged. The question of using a control group must also be considered from the cost of the effort compared with the value of the data. There seems to be no consistent answer to this question (cf. Poynor,[42] Feshback[57]).

Finally, the issue of whether applied (usually perceived as atheoretical) or basic (theoretical) research will produce the type of information necessary to improve the quality of nursing care must be considered. Erroneously, many nurses have assumed that because research is conducted in the clinical or field setting it is atheoretical. The use of the applied research approach does not mean that a concept is taken out of its theoretical context. Findings from rigorous applied research can be used to develop, test and refine a theory.

In reality there is probably not enough data to determine if basic or applied research will be the most productive approach. Currently the emphasis in nursing is on the conduct of applied research; however, evidence from the Comroe and Dripps study challenges this emphasis. Comroe and Dripps identified scientific advances as reported in 529 articles—61.7% of the articles reported results of basic research; 21.2% represented a wide variety

of research types; 15.3% were concerned with developing technical apparatus or procedures; and 1.8% were review or synthesis efforts.[58] Certainly this study does not "prove" anything; however, it does provide support for the contention that both applied and basic research needs to be conducted in nursing.

Ironically, once researchers decide to conduct either applied or basic research they encounter the same problem: the simplistic versions of theories reported in the nursing literature. Why such versions are offered is not clear, but often the power of a theory is greatly reduced by summarizing it in broad generalizations. Evidently these generalizations were developed in an attempt to make the theory usable in all types of situations. In reality, the use of that theory in nursing is all but impossible because it has lost its power to produce understanding or to predict. Certainly this has been true of some of the writing concerning systems theory, role theory and others. Few nurses are even aware that there are several different role theories rather than one global theory, much less that one might be more appropriate for conceptualizing nursing interventions than are others.

PREPARING NURSES FOR CLINICAL RESEARCH

There is no doubt that the nurse researcher who wishes to conduct rigorous research in a clinical setting needs at least an acquaintance with information relating to measurement, theory, design and statistics. Since the extraneous variables testing, instrumentation and statistical regression are influenced by reliability of the instruments used in obtaining data, nurse researchers should know the techniques needed to create reliable instruments and know how to compute and to interpret reliability coefficients.

Nurse researchers should know the techniques needed to create reliable instruments and know how to compute and to interpret reliability coefficients.

There are two approaches to preparing nurse researchers to conduct clinical research. A step-by-step procedure can be developed by which the study can be conducted without the researcher's having an in-depth understanding of the rationale for each step. Even with this step-by-step approach the researcher still needs to be able to interact intelligently with experts in the areas of measurement, theory, design and statistics. The other approach is to create learning experiences that allow the potential researcher to develop expertise in these skills. This approach is not well accepted, because many nurses believe that since their interests are clinical in nature they do not really need to know much about research. However, information in this article illustrates that a considerable knowledge base is required if the nurse is to be able to make judgments about the adequacy of clinical research.

To conduct rigorous research in a clinical setting, the nurse researcher needs to have acquired skills in theory, measurement, statistics and design. If research is really the basis for sound

10

clinical practice, nurses must reevaluate their current positions regarding the value of basic versus applied research, the role of theory in relation to clinical practice and the importance of mastering the necessary research skills to conduct adequately conceived research in the clinical setting.

REFERENCES

1. Campbell, D. T. and Stanley, J. C. *Experimental and Quasi-Experimental Designs Research* (Chicago: Rand McNally 1963).
2. Box, G. E. P., and Jenkins, G. M. *Time Series Analysis: Forecasting and Control* (San Francisco: Holden-Day 1970).
3. Anderson, T. W. *The Statistical Analysis of Time Series* (New York: John Wiley & Sons 1971).
4. Glass, G. V. "Estimating the Effects of Intervention Into a Non-Stationary Time Series." *AERJ* 9 (1972) p. 9, 463-477.
5. Bloomfield, P. *Fourier Analysis of Time Series: An Introduction* (New York: John Wiley & Sons 1976).
6. Lana, R. E. "Inhibitory Effects of a Pretest on Opinion Change." *Educ Psychol Meas* 26 (1966) p. 139-150.
7. Bracht, G. H., and Glass, G. V. "The External Validity of Experiments." *AERJ* 5 (1968) p. 437-474.
8. Welch, W. W. and Walberg, H. J. "Pretest and Sensitization Effects in Curriculm Evaluation." *AERJ* 7 (1970) p. 605-614.
9. Harris, C. W., ed. *Problems in Measuring Change* (Madison: University of Wisconsin 1963).
10. Tucker, L. R., Domarin, F. and Messick, S. "A Base-Free Measure of Change." *Psychometrika* 3 (1966) p. 457-473.
11. Cronbach, L. J. and Furby, L. "How We Should Measure 'Change'—Or Should We?" *Psychol Bull* 74 (1970) p. 68-80.
12. O'Connor, E. F. "Extending Classical Test Theory to the Measurement of Change." *Rev Educ Res* 42 (1972) p. 73-97.
13. Richards, J. M. "A Simulation Study of the Use of Change Measures to Compare Educational Programs." *AERJ* 12 (1975) p. 299-311.
14. Abbey, F. C., "Bioinstrumentation: Twentieth Century Slave." *Nurs Clin N Am* 13:4 (1978) p. 631-640.
15. Huether, S. E. "Biomedical Instrumentation and the Nurse Practitioner." *Nurs Clin N Am* 13:4 (1978) p. 561-571.
16. Jacobs, M. K. "Sources of Measurement Error in Noninvasive Electronic Instrumentation." *Nurs Clin N Am* 13:4 (1978) p. 573-587.
17. Weeks, H. "Bioinstrumentation in the Care of the Neonate." *Nurs Clin N Am* 13:4 (1978) p. 597-609.
18. Cronbach, L. J. et al. *The Dependability of Behavioral Measures: Generalizability of Scores and Profiles* (New York: John Wiley & Sons 1972).
19. McGraw, B., Wardrop, J. L. and Bunda, M. A. "Classroom Observation Schemes—Where Are the Errors?" *AERJ* 9 (1972) p. 13-27.
20. Herbert, J. D. and Attridge, C. "A Guide For Developers and Users of Observational Systems and Manuals." *AERJ* 12 (1975) p. 1-20.
21. Rowley, G. L. "The Reliability of Observational Measures." *AERJ* 13 (1976) p. 51-59.
22. Frick, T. and Semmel, M. I. "Observer Agreement and Reliabilities of Classroom Observational Measures." *Rev Educ Res* 48 (1978) p. 157-184.
23. Ebel, R. L. *Essentials of Educational Measurement* (Englewood Cliffs, N.J.: Prentice-Hall 1972).
24. Mehrens, W. A. and Lehman, I. J. "Measurement and Evaluation in Education and Psychology." (New York: Holt, Rinehart and Winston 1973).
25. Fitzpatrick, J. J. and Donoran, M. J. "A Follow-Up Study of the Reliability and Validity of the Motor Activity Rating Scale." *Nurs Res* 28:3 (1979) p. 179-181.
26. Kearney, B. Y. and Fleischer, B. J. "Development of an Instrument to Measure Exercise of Selfcare Agency." *Res Nurs Health* 2:1 (1979) p. 25-34.
27. Ward, M. J. and Fetler, M. *Instruments for Use in Nursing Education Research* (Boulder, Colo.: Western Interstate Commission for Higher Education 1979).
28. Ward, M. J. and Lindemann, C. A. *Instruments for Measuring Nursing Practice and Other Health Care Variables* Vol. I and II. DHEW Pub. No. HRA 78-53 and 78-54 (Washington, D.C. 1979).
29. Edgington, E. S. "Statistical Reference and Non-Random Samples" in Heermann, E.F. and Braskamp, L.A. eds. *Readings in Statistics for the Behavioral Sciences* (Englewood Cliffs, N. J.: Prentice-Hall 1970).
30. Glass, G. V. and Stanley, J. C. *Statistical Method in Education and Psychology* (Englewood Cliffs, N. J.: Prentice-Hall 1970).
31. Slakter, M. J. *Statistical Reference for Educational Researchers* (Reading, Mass.: Addison-Wesley 1972).

32. Wooldridge, P. J., Leonard, R. C. and Skipper, J. K. *Method of Clinical Experimentation to Improve Patient Care* (St. Louis: The C.V. Mosby Co. 1978).
33. Polit, D. F. and Hungler, B. P. *Nursing Research: Principles and Methods* (New York: J. B. Lippincott Co. 1978).
34. Winer, B. J. *Statistical Principles in Experimental Design* (New York: McGraw-Hill 1971).
35. Huynh, H. and Feldt, L. S. "Conditions Under Which Mean Square Ratios in Repeated Measurements Designs Have Exact F-Distributions." *J Am Stat Assoc* 65 (1970) p. 1582-1589.
36. Huynh, H. and Feldt, L. S. "Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split Plot Designs." *J Educ Stat* (1976) p. 69-82.
37. Berliner, D. C. and Cahen, L. S. "Trait-Treatment Interaction and Learning" in Kerlinger, F. N., ed. *Review of Research in Education* (Itasca, Ill.: F. E. Peacock Publishers 1973).
38. Borich, G., Godbout, R. and Wunderlick, K. *The Analysis of Aptitude-Treatment Interactions: Computer Programs and Calculations* (Chicago: International Educational Services 1976).
39. Snow, R. E. "Research on Aptitude for Learning: A Progress Report" in Shulman, L., ed. *Review of Research in Education* (Itasca, Ill.: F. E. Peacock 1976) p. 50-105.
40. Cronbach, L. J. and Snow, R. E. *Aptitudes and Instructional Methods: A Handbook for Research and Interactions* (New York: Irvington 1977).
41. Stoloff, P. H. "Correcting for Heterogeneity of Covariance for Repeated Measures Designs of Variance." *Educ Psychol Meas* 30 (1970) p. 909-924.
42. Poynor, H. "Selecting Units of Analysis" in Borich, G. D., ed. *Evaluating Educational Programs and Products* (Englewood, Calif.: Educational Technology Publication 1974).
43. Meskauskas, J. A. "Evaluation Models for Criterion and Referenced Testing: Views Regarding Mastery and Standard Setting. *Rev Educ Res* 46 (1976) p. 133-158.
44. Subkoviak, M. J. and Baker, F. B. "Test Theory" in Shulman, L. S., ed. *Review of Research in Education* (Itasca, Ill.: F. E. Peacock 1977).
45. Hambleton, R. K. et al. "Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments." *Rev Educ Res* 48 (1978) p. 1-43.
46. Wang, G. H. "The Galvanic Skin Reflex: A Review of Old and Recent Works from a Physiological Point of View." *Am J Phys Med* 37 (1958) p. 35-37.
47. Forbes, T. W. "Problems in Measurement of Electrodermal Phenomena—Choice of Method and Phenomena—Potential, Impedance, Resistance." *Psychophysiology* 1 (1964) p. 26-30.
48. Edelberg, R. "The Electrodermal System" in Greenfield, N. S. and Sternback, R. A., eds. *Handbook of Psychology* New York: Holt, Rinehart and Winston 1972).
49. Prokasky, W. F. and Raskin, D. C., eds. *Electrodermal Activity in Psychological Research* (New York: Academic Press 1973).
50. Seligman, L. "Skin Potential as an Indicator of Emotion." *J Couns Psych* 22 (1975) p. 489-493.
51. Kennedy, J. J. "The eta Coefficient in ANOVA Designs." *Educ Psychol Meas* 30 (1970) p. 885-889.
52. Halderson, J. S. and Glasnapp, D. R. "Generalized Rules for Calculating the Magnitude of an Effect in Factoral and Repeated Measures ANOVA Designs." *AERJ* 9 (1972) p. 301-316.
53. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (New York: Academic Press 1969).
54. Cohen, J. "Statistical Power Analysis and Research Results." *AERJ* 10 (1973) p. 225-229.
55. Meyer, D. "Statistical Tests and Surveys of Power: A Critique." *AERJ* 11 (1974) p. 179-188.
56. Brewer, J. K. "On the Power of Statistical Tests in AERJ." *AERJ* 9 (1972) p. 391-401.
57. Feshback, S. "The Illusory Control Group in Field Research: Comments on the ESAA Evaluation." *J Educ Stat* 3 (1978) p. 87-93.
58. Comroe, J. H. and Dripps, R. D. "Scientific Basis for the Support of Biomedical Science." *Science* 1076 (1976) p. 105-111.