Selected Critical Measurement and Statistical Issues in Health Education Evaluation and Research

Mohammad R. Torabi, Ph.D., M.P.H., C.H.E.S.[1]; Kele Ding, M.S., MD.[2]

[1] Chancellor's Professor and Assistant Department Chairperson, Indiana University

[2] Doctoral Candidate, Department of Applied Health Science, Indiana University,

Corresponding Author: Mohammad R. Torabi, Indiana University, Department of Applied Health Science, HPER 116, Bloomington, IN 47405; 812.855.3627 (phone), 812.855.3936 (fax), TORABI@INDIANA.EDU

Introduction

The backbone of any well established, undisputed, and credible applied discipline is established through sound research evaluational studies. In comparison with other applied disciplines like medicine, law, and education, health education as a discipline has had a relatively short history. Still there are those who question the credibility of health education as a discipline and challenge health education as a profession. These disputed and image problems may be partially due to the fact that some of our research and evaluational studies lack vigor with some methodological flaws. These criticisms are not unusual for a very young discipline. As a matter of fact, health education has remained a very dynamic and responsive profession. It has made great progress on all fronts. In some instances, our discipline and its leadership have surpassed other comparable disciplines. Examples of these triumphs can be cited in the establishment of Electronic Journals like this Journal, in credentialing, in tremendous improvement of the quality of publications in existing journals, and of publications from funded research projects sponsored by the NCI, CDC, and so forth. Despite this progress, there are critical measurement and statistical issues that need to be addressed. This article uses an extensive review of literature to address selected statistical and measurement issues and problems related to educational research in general and health education research in specific. These selected topics vary from self-report data to sample size and they are briefly discussed with sub-headings below.

Self-report versus True Data

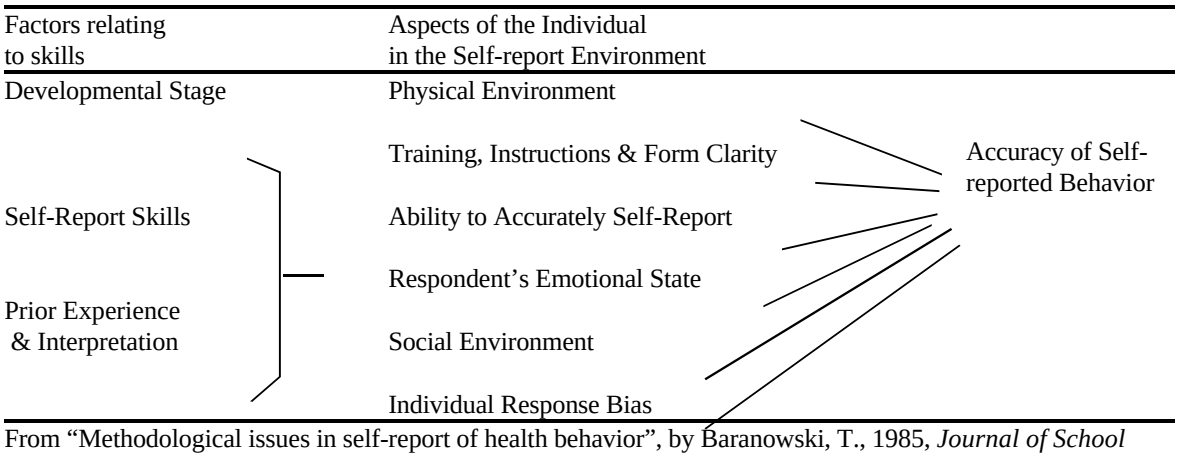
The ultimate purpose of a health education program is to help individuals to make intelligent decisions and to behave accordingly with regard to their own and community health and well being (Torabi, 1995). Self-report measures of attitude and behavior are among the most widely used in health education research and program evaluation. Self-report, as indicated by Baranowski (1985), has been used for the measurement of a wide variety of variables such as personality, assessment, family interaction assessment, and test of knowledge, ability, and attitudes. There are several advantages of self-report measurement:

1. The individual is in the best position to observe, describe, and report upon his own behavior (Brown, 1970). It is quite easy to understand that the only observer who can follow one's everyday behavior and know one's thoughts, reasons for taking certain action, and attitude toward certain object is the person himself;
2. Self-report measures can be easily and relatively quickly obtained (Baranowski, 1985); and
3. Self-report forms seem to measure exactly what the investigator wants and the type of data obtained by these methods usually are unavailable from other methods (Baranowski, 1985).

However, there are disadvantages with this measurement method. In quantitative study, reliability and validity of self-reported data is a key question (Baranowski, Dworkin, Cieslik, & Hooks, 1984). May and Foxcroft (1995) found bias in self-reports in qualitative research is also a key problem. Many factors affect the accuracy of the data collected via self-report. Baranowski (1985) summarized those common factors influencing the accuracy of self-report of behavior as physical environment, training, instructions and form clarity, ability to accurately self-report, respondent's emotional state, social environment, and individual response bias (Figure 1). Among these, response bias probably is the toughest problem for investigators to prevent.

Response bias affects both the reliability and validity of the measurement. Reliability estimation is affected because consistency of response may be increased if all questions are answered in a socially desirable direction yielding a high test-retest and internal consistency coefficient. This coefficient may be interpreted falsely as a favorable reliability estimate (Baranowski, 1985). Validity is affected because a construct other than the construct of interest is used for response (Cook and Campbell, 1979). Health is a social desirable ideal. According to Phillips (1971) and Edwards (1957), the term social desirability is used to describe a biasing factor that results from the respondent's attempt to demonstrate behavior that is socially desirable or preferred. In regards to health, it is desirable to appear healthy, or it is desirable to be disabled. In Papenfuss and Beier's study (1984), behavior changes were found among 100% of subjects in the experimental group and 16% of the control group after intervention. One of the explanations for this result was that the subjects may have anticipated that the project staff wanted such changes, and reported the changes either to make the staff feel good or to avoid

Figure 1
Factors Influencing the Accuracy of Self-Report of Behavior



From "Methodological issues in self-report of health behavior", by Baranowski, T., 1985, *Journal of School Health*, 55(5), p. 180.

some anticipated approbation for not having made such changes. Another sample of response bias of self-report is provided by Catania, Gibson, Chitwood, and Coates (1990). They found it is very reasonable to suspect inconsistencies between people's actual sexual behavior and their self-reports in an interview. Privacy, embarrassment, and fear of reprisals are but a few of the reasons people do not provide true data.

As Torabi suggested, "recognizing the limitations of measuring the actual health behavior, health evaluators should be innovative in collecting more data related to actual behavior while protecting human subjects" (Torabi, 1995, p. 3). Approaches to improve self-report measurement have been reported in literature. Barabiowski (1985, p. 181-182) proposed 8 steps to promote more accurate self-report of health behaviors as follows.

1. Select measures that clearly reflect program outcomes
2. Select measures that have been designed to anticipate response problems and that have been validated
3. Conduct a pilot study with the population
4. Anticipate and correct major sources of unreliability
5. Employ quality control procedures to detect other sources of error
6. Employ multiple methods
7. Use multiple measures
8. Use experimental and control groups with random assignment to control for biases in self-report.

Using social desirability test is also suggested as a method to detect social desirability by many authors (Miller, 1990; Mueller, 1985; Crowne & Marlow, 1960).

Statistics versus Practical Significance

Since Fisher, Neyman and Pearson set up the tone of null hypothesis testing, 70 years have passed (Kirk, 1996). The test has been an integral part of the research enterprise in which behavioral and educational researchers engage. A null hypothesis that there is no relationship (or no difference) between two variables is tested in order to find whether or not test data retains or rejects the null hypothesis. Such a test is sometimes referred to as a test of significance or significance test (Hays, 1963; Snedecor & Cochran, 1967). When a null hypothesis is rejected, the relationships (or the differences) of sample results are said to be significant because such results signify rejection of the null hypothesis (Snedecor & Cochran, 1967). Apparently, significance test tells whether or not a difference or relationship exists between two variables. However, when the question is about how magnitude the difference or the relationship is, a p value seems helpless. In Shaver's words (1996, p. 302), "Statistics significance

indicates neither the magnitude nor the importance of a result." That brings out the issue of practical significance.

Practical significance refers to the extent of the difference or the relationship between two tested variables (Torabi, 1995), and to the scientific or practical importance of conditions that exist in populations (Daniel, 1977). Levin (1996, p. 378-379) once used a problem in his article in explaining statistical significance and practical significance as follows.

"Investigator A conducts a two-group study with n=2,000 subjects per condition, performs a two-sample t test, and reports that for the hypothesis tested p=. 048. Investigator B conducts a similar study with n=2 subject per condition, performs the same statistical test, and also finds that p=. 048. Two comprehension-demanding questions are: (a) Which investigator's findings are more credible (i.e., likely to be non-chance)? and (b) Which investigator's findings are more impressive (e.g., worth getting excited about or investing in)?"

Here, the first question addresses the notion of statistical significance, whereas the second addresses practical significance.

As Torabi(1986, p. 232) indicated, "Investigators in applied disciplines sometimes consider statistical significance as the ultimate answer to their research problems. A finding of statistical significance may or may not have any practical meaning unless the results are examined further for practical significance." This indication matches many authors' opinion (e.g., Daniel, 1977; Gill, McNamara and Skinkle, 1980; Levin, 1996; Carver, 1996; Shaver, 1996). While comments regarding the null hypothesis tests are widely spread from no use at all, minimizing the use, to use as usual, testing and reporting practical significance is commonly suggested. Confidence interval, effect size, power test, replication, the Holm and the Shaffer procedures, and the conventional and bootstrap significance tests are among the recommended techniques being used to generate more practical information other than ps (Torabi, 1986; Carver, 1996; Daniel, 1977; Gill, McNamara and Skinkle, 1980; Serlin, 1996; Loftus, 1991).

According to Levin (1996), the American Psychological Association's (APA) publication manual editors "unanimously endorsed the position that information about á levels, p values, effect sizes, and power routinely be included in reports of empirical research" in the manual's next edition (p. 381). This might implies a future trend of the use of statistical significance tests in health education research.

Qualitative versus Quantitative Approaches

Quantitative approaches prevail in Health Education research. Quantitative method provided health education a substantial base for practice and for legitimacy and contributed to standards of accountability (Green and Lewis, 1986). Nevertheless, the predominating use of quantitative approaches has received increasing critics such as inadequate and unsuitable in social programming (Cronbach, 1975; Gebhardt, 1980; House, 1980). More researchers have been advocating the acceptance and integration of qualitative methods in health education, and many advocate a combined use of both approaches (e.g. Cook & Campbell, 1979; Patton, 1987; Basch, 1987; Steckler, McLeroy, Goodmab, Bird, & McCormic, 1992; Israel et al, 1995; Torabi, 1995).

The differences between qualitative and qualitative paradigms have been discussed thoroughly in literature. Quantitative research uses a rationalistic paradigm. It assumes that reality exists "out there" for anyone to see or experience through the sense. Three features of quantitative research are summarized as reductionism (that parts can be separated from the whole for study), repeatability (what has been discovered by one should be repeatable by another), and refutation (what is asserted should be conformable or refutable). To the contrary, qualitative research uses a naturalistic paradigm. It assumes that "reality does not exist out there for everyone to see and experience in the same way, but the world is both found (as objective reality) and made (that is socially constructed by each individual)" (Bhola, 1990, p. 29). The methods used are different from quantitative study. The main points are (1) the evaluator/researcher is himself or herself part of the phenomenon under study; (2) the design is emergent; and (3) the instruments are always unstructured and generate qualitative data (Bhola, 1990). Why qualitative research? As indicated by many authors, qualitative approaches have their advantages comparing to quantitative approaches. Patton (1987) points out that qualitative methods to evaluation can provide (1) detailed descriptions of program implementation, (2) analysis of major program process, (3) description of different types of participants and participation, (4) descriptions of how the program affects participants, (5) analysis of observed changes (or lack thereof) for program implementation and evaluation, and (6) analysis of program strengths and weaknesses. Qualitative data can provide insight into the internal validity of a quantitative design (Skeckler, 1989). When combined with quantitative data, qualitative data can provide much more meaningful results and lead to practical recommendations (Torabi, 1995). This

combined use of both quantitative and qualitative approaches increases the validity of the conclusions and the range of information collected (Reichardt & Cook, 1979; Yin, 1984). Furthermore, "qualitative methods provide contextual understanding of health behavior and program results (Steckler, McLeroy, Goodmab, Bird, & McCormic, 1992, p. 2). In summary, as indicated by Steckler et al (1992), both approaches have weaknesses, which are compensated for by the strengths of the other.

It is noticeable that some social scientists and philosophers disagree about the combined use of both quantitative and qualitative methods (e.g. Rosenberg, 1988; Lincoln, 1985; Noblitt & Hare, 1988). The argument is that research pursues either prediction or intelligibility; it cannot be both (Steckler et al, 1992). And the two paradigms are so different that any reconciliation between them is bound to destroy the epistemological foundations of each (Rosenberg, 1988). However, practitioners in health education research do not stop their practice of integrating both qualitative and quantitative approaches. As Steckler et al (1992) stated, both approaches have their own advantages and can be considered as separate methodologies contributing to social science in general and health education in particular.

Literature has showed that a variety of data collection techniques of qualitative study has been practiced in health education research. Frequently used and recommended methods include open-ended questions on questionnaire survey, ethnographic interviews, open-ended and in-depth interviews, ethnographic field notes, focus groups and group interviews, participant observation, and so forth (Steckler et al, 1992; Torabi, 1995; Israel et al, 1995; Vries, Weijts, Dijkstra, & Kok, 1992).

Steckler et al (1992) suggested four models for integrating quantitative and qualitative methods in health education research and program evaluation: (1) qualitative methods are used to help develop quantitative measures and instruments; (2) qualitative methods are used to help explain quantitative findings; (3) quantitative methods are used to embellish a primarily qualitative study; and (4) qualitative and quantitative methods are used equally and parallel.

Test Length and Test Reliability

Test length determines test reliability that is the longer the test, the higher the reliability (Grounlund, 1981; Ebel, 1980; Noll & Scannell, 1972; Brown, 1970). In fact, one can often make a fairly accurate guess as to the reliability of a test just by knowing its length (Gulliksen, 1950). Spearman-Brown Prophecy Formula is used for proving it. The formula is:

$$r_{nn} = nr \Big/ 1 + (n-1)r$$

Where r_{nn} = estimated reliability when the length of the test is increased n times, and r = the reliability of the test in question (Noll & Scannell, 1972). Obviously, when numerator n increases, r_{nn} increases too. However, this formula does not include other conditions that also affect a test's reliability.

And only when all other conditions are equal, increasing the number of questions can result in a higher reliability coefficient. Ebel (1980) introduced other conditions including item homogeneous, item discriminating power, item difficulty level, sample homogenous, and test speed.

It is worth noting that the correlate of test length and test reliability is usually examined by using a theoretical formula such as Spearman-Brown Prophecy Formula. When testing this relation with empirical data, the result could be different. In Torabi's (1988) study, a 54-item alcohol attitude scale was administered to a group of 700 students. The Cronbach Alpha reached 0.96 and the Split-half 0.91 for the data collected. By selecting a different number of subjects and a different number of items, the Cronbach Alpha was found to increase as the number of items increase until the number of items reaches 18. Adding more items did not increase Cronbach Alpha reliability coefficient significantly. This finding might imply a nonlinear relationship between the number of test items and the magnitude of test consistency. Nevertheless, Wainer (1986, p. 171) said, "when a relative short test shows unusually high reliability it should be cause for concern rather than unbridled jubilation." In his study, Wainer (1986) indicated, if reliability is too high or too low for a relative short test (28 items in his sample), test scores must be scrutinized. Otherwise, summary statistics can be seriously misleading.

Reliability versus Validity

Measurement is a central part of health education research and a measurement instrument must produce valid and reliable results (Torabi 1995). "Validity and reliability are two benchmark criteria for assessing the quality of all measurement devices and procedures" (Muller, 1985, p. 57). Often reliability is heavily relied versus validity as a single important psychometric evidence for an instrument. However, validity evidence is also important through various means because an instrument can produce a highly reliable result and not necessarily measure what it is supposed to measure (Torabi, 1995).

Validity, as described in most research methodology books (e.g., Horrocks, 1964; Muller, 1985; Cozby, 1985;

Baumgartner and Strong, 1994) is the extent to which a test measures what it is supposed to measure. But authors such as Rulon (1964) thought this is an unsatisfactory and not very useful concept since thus the validity may be high for one use and low for another. Usually four types of validity are introduced by most of research methodology books. They are content validity, concurrent validity, predictive validity, and construct validity (e.g., Horrocks, 1964; Muller, 1985; Cozby, 1985; Baumgartner and Strong, 1994). Face validity is also emphasized by some authors (Horrocks, 1964; Muller, 1985). Face validity makes items "look like" what it is supposed to measure."The instrument is judged after it is constructed; it is a check on completed product"(Green & Lewis, 1986, p. 106).

Although most of the textbooks did not specify the validity issue in health behavior measurement, Horrocks (1964) suggested five criteria should be taken into consideration when developing a valid measurement instrument. They are the true outcomes of the construct in question, observable, measurable in some quantitative fashion readily definable, and agreed upon by the individuals concerned. However, as indicated by Horrocks (1964) perfect validity is an ideal and it is rarely approached. Evidence of validity have to be evaluated in terms of the appropriateness of the criteria, the degree of agreement, the extent to which the test will help to achieve the present situation demands, and so on. For example, Catania et all (1990) suggested involving validating self-reports with a biological mark as a strategy. Udry and Morris (1967) tested urine for the presence of sperm and correlated these findings with self-reported incidence of coitus. However, low validity is also acceptable in some cases because it still "provides better conclusions than those we would have without it" (Horrocks, 1964, p. 63).

Reliability is "the accuracy or precision of a measuring instrument" (Kerlinger, 1964, p. 443), and means "stability, predictability, dependability, consistency" (Kerlinger, 1979, p. 132). As Cozby (1985) noted, any measure can be thought of as made up of two components: (1) a true score, which is the real score on the variable, and (2) measurement error. A reliable measure contains little measurement error.

Generally, five types of reliability are usually discussed as internal consistency, test-retest, alternate forms reliability, split-half and odd-even reliability, and inter/intra-rater reliability (Torabi, 1994; Mueller, 1985; Cozby, 1985; Kerlinger, 1979; Horrocks, 1964;). Although internal consistency is the most important indicator of a reliable instrument, it might have limitations when used in behavioral measurement. As Horrocks indicated, "Internal consistency is of major importance when the measure is dealing with a single or generalized trait or aspect of behavior; but if the measure is dealing with many diverse traits or aspects of behavior internal consistency may be low for such tests other questions of reliability are more important" (Horrocks, 1964, p. 64).

Wording Issues in Measurement

Wording is important to survey instrument construct and behavior measurement. The purpose of the words chosen is to communicate explicit meaning as efficiently as possible (Ebel, 1980). However, problems are found in practice when negative and positive wording are used together or when similar worded items were used in measuring attitude and behavior.

Measuring both attitude and behavior by one survey questionnaire is common in health education research. Studies showed item wording could cause problem in this practice. Feldman and Lynch (1988) used similarly worded items to measure attitudes and behaviors. The results suggested that a response to a question is likely to be retrieved as a basis for a subsequent response if they are presented consecutively in one questionnaire. In Beland, Maheux, and Lambert's (1991) experimental study, 21 attitudinal items and 21 similarly worded behavioral items were used together as experimental group and separate for control groups. The finding showed that the correlation between attitude and behavior is higher in experimental group than that in the control group. It is explained that a questionnaire wherein attitude items are followed by similar worded behavioral items, respondents answered both types of questions in a similar way. The recommendation, therefore, is if similarly worded items were used in one questionnaire testing both attitudes and behaviors two questionnaires are required.

Negative and positive wording is also a concern in health education research. The use of equal negative and positive worded items on personality, attitude, and other rating scale instruments is the rule in test construct used and recommended by many researchers and text books (e.g., shiffman & Jarvik, 1976; Tiffany & Drobes, 1991; Mueller, 1985; Anastasi, 1988; Mehrens & Lehmann, 1991). This construction of a questionnaire is said to result in a more psychometrically sound instrument because positively and negatively worded items are measuring the same construct, and including both of them makes the questionnaire more valid. However, reports from Benson & Hocevar (1985), Schriesheim & Hill (1981), Taylor & bowers (1972), Marsh (1986), and Sweeney, Pillitteri, & Kozlowski (1996) told different stories.

Benson & Hocevar (1985) developed three parallel scales that consisted of all positively worded items, all negatively worded items, and a mixture of the two. They found that scales defined by positive and negative items differed significantly in terms of scale means, scale variances, and scale reliabilities; subjects had difficulty responding appropriately to the negatively worded items. When both positive and negative items were included in the same form, responses were more affected by the positive or negative phrasing than by the item content. They concluded that the inclusion of negative items adversely affects the validity of responses by their subjects. Marsh (1986) explained that in order to respond appropriately to negative items, respondents may have to invoke a double negative logic that requires a higher level of verbal reasoning than that required by positive items. In his sample, the item "I am not smart at mathematics" requires a response of "false" to indicate that "I am smart at mathematics." If this logic is not appropriately employed, respondents may give an answer that has a meaning exactly opposite to that of their intended response. This is late called as "negative item bias". Here, "negative item bias" is defined as occurring when a child responds inappropriately by saying "true" to a negative statement when his or her responses to positive items have consistently indicated that the opposite response would be more appropriate, or vise versa (Marsh, 1986). More explicitly, it is defined as occurring when mean score of negative worded items tends to be lower or higher than positive worded items (Sweeney, Pillitteri & Kozlowski, 1996).

Subjects in Benson and Hocevar's (1985) and Marsh's (1986) studies were preschool adolescents. The negative item bias is therefore further explained as a cognitive-developmental phenomenon. Younger children and children with poorer reading skills are thought less able to respond appropriately to negative items. However, studies conducted by Sweeney, Pillitteri and Kozlowski (1996), by Schrieshiem and Hill (1981), and by Taylor and Bowers (1972) also encountered negative item bias with their adult subjects.

Sweeney, Pillitteri and Kozlowski (1996) adopted Tiffany and Drobes'(1991) Questionnaire of Smoking Urges. They reversed the original 32 items to make negatively worded items. Thus a final 64-item questionnaire was developed and applied in an adult sample. The survey results showed that mean scores for negatively worded items tended to be higher than mean scores for positively worded items indicating that subjects generally agreed more/disagreed less with statements worded negatively than with statements worded positively. Furthermore, some statements, particularly those double negatives, proved to be especially troublesome for respondents. They conclude that negatively worded items were less valid than positive worded items and practice of balancing scales should be discontinued.

Parametric and Nonparametric Statistics

Researchers in health education often confront a question related to their nominal or ordinal data as to whether they use parametric or nonparametric method in data analysis. Yet, in most cases parametric methods were used. Using nonparametric tests instead of parametric tests has two general considerations. One consideration is about underling assumptions of parametric tests and the other is the problem of whether or not the measurement scale is suitable for application of parametric procedures. It is noticed that nonparametric methods are particularly appropriate in psychology, education, and behavioral science because of the type of data (Siegel & Castellan, 1988).

Generally, data for using parametric statistics are assumed to be normally distributed with a homogeneity of variances and linearity (Kuzma, 1992). The data scale is interval. Nonparametric methods are sometimes referred to as distribution-free methods because the observations can not be normally distributed (Kuzma, 1992; Weimer, 1993; Conover, 1980), or because the sampling distribution does not depend on the specific distribution of the population from which the sample was drawn (Gibbons, 1993). The data scale is usually ordinal or ranking order scale. In Siegel & Castellan's (1988) words, the data are "with scores which are not exact in any numerical sense, but which in effect are simply ranks (p.XV)". These data are usually collected in health educational research.

In health education, researchers frequently want to describe the correlation, association, or agreement between two or more related population groups and test whether there is a significant difference. But even if data from two groups of people are collected by taking random samples from the same population, they could be different to some extent. Statistical tests would enable one to find whether or not such difference occurs by chance.

Literally, controversy about the use of nonparametric methods exists relative to being less efficient and not sensitive to detecting real difference (Kuzma, 1992; Weimer, 1993; Conover, 1980; Bradley, 1977). As Jenkins and Fuqua (1984) pointed out this would bring a high probability for leading the researcher to commit a Type II error. However, Dixon (1954) and Hodges and Lehman (1956) found in certain cases that Wilcoxon tests

with a sample of 100 has the same power as the test based on 95 observations. The Kruskal-Wallis one-way analysis of variance has been shown to have an power of 0.955 when compared to the classic F test (Bradley, 1968). In their study, Ittenbach, Chayer, Bruininks, Thurlow, & Beirne-Smith (1993) used four approaches - parametric multivariate analysis of variance, nonparametric multivariate analysis of variance, multiple nonparametric analysis of variance, and multinominal logistic regression in analyzing a set of data regarding. They found the patterns of significance, indices of substantive significance, and measures of statistical power were virtually identical in three of the four principle procedures. The nonparametric multivariate analysis of variance procedure retained a modest advantage over the other techniques used. From these cases, it is seen that nonparametric approaches can be as powerful as parametric methods in detecting a difference. As Hunter and May (1993) stated, although the parametric tests are more powerful than nonparametric tests, meeting their assumptions make the test appealing. Thus, as Bradley (1978) indicated, when some of the assumptions of a parametric test are not met, nonparametric tests could be more powerful than comparable parametric tests.

Another selection of using parametric or nonparametric approach is based on data scale. Stevens (1946, 1968) outlined four categories into which variables are assigned. They are: nominal, ordinal, interval, and ratio. Generally, statistics textbooks suggest that parametric procedures are used for interval and ratio scale and nonparametric techniques are used for nominal and ordinal data (Siegel, 1956; Kuzma, 1992; Weimer, 1993; Gibbons, 1971; Lehmann, 1975; Siegel & Castellan, 1988). However, this criteria in selecting parametric and nonparametric tests is criticized as a religious prescription (Harwell, 1988). The rationale for the controversy focused on relations of the nature of the data and the mechanical act of validly using a statistical test (Gardner, 1975; Savage, 1957; Harwell, 1988). As Harwell (1988) indicated, "the sole statistical criterion is the fit between the model and the data; if the fit is good, the test can be performed validly" (p. 37).

How does the researcher make a choice between parametric or nonparametric techniques based on the above discussion when there is no widely acceptable prescription? Some authors (e.g., Jenkins & Fuqua, 1984; Hunter & May, 1993; Harwell, 1988) suggested

1. When the hypothesis can be tested by either method, the most statistically powerful method should be used;
2. When the sample size is small, nonparametric test should be computed, or the normality assumption should be verified. If the data fit parametric assumptions, parametric methods can be used; if not, nonparametric methods should be preferred; and
3. Methods selection also depends on the nature of data and on a substantive basis.

Testing or Not Testing Null Hypothesis

Null hypotheses tests are taught in nearly every university setting and written in numerous textbooks. However, critics of the use has been under attack for over 30 years (Shaver, 1996). A strong statement can be found in Carver's article (1978) which says "even if properly used in the scientific method, educational research would still be better off without statistical significance testing" (p. 398). Other authors also expressed their opposing opinions against statistically significant tests (Skinner, 1956; Bakan, 1967; Meehl, 1967; Morison & Henkel, 1970; Seeman, 1973). Several points can be made which argue that testing null hypotheses has its drawbacks.

In Carver's (1996) opinion, statistical significance is a function of effect size and sampling error. A t-test actually tests the ratio of effect size and sampling error. Therefore, "it is much better to report effect size and sampling error and forget about their ratio and ratio's associated p value "(Carver, 1996, p. 290).

Random sample is "the building blocks for hypothesis testing" (Glass & Hopkins, 1984, p. 202) and is the essential assumption for the hypothesis test. If a sample or samples are not random, null hypothesis test won't yield a meaningful probability statement (Shaver, 1996).

The probability of a false null hypothesis in a statistically significant test states the occurrence in the long run, with repeated random sampling or random assignment. As Shaver (1996) pointed out, "it provides no basis for a conclusion about the probability that a particular result is attributable to chance" (p. 300). Replication is essential at this point. However, as Shaver (1996) indicated "statistical significance not only provide no information about the probability that replications of a study would yield the same result, but is of little relevance in judging whether actual replications yield similar results" (p. 304).

A result of statistically significant findings does not mean the probability that the null hypothesis is true or false. It provides information about the likelihood of a result given that the null hypothesis is true (Shaver, 1996; Carver, 1978; Cohen, 1990). To reject the null hypothesis based on one statistically significant result is a conclusion that is too absolute. On the other hand, a test

of statistical significance does not provide information on the probability that an alternative hypothesis is true or false.

Again, statistical significance tests do not provide information about the magnitude of a difference or association being tested because the effect size is a function of sample size. Sample size plays a role in determining the statistics test to be significant or not. It is believed by many authors (e.g. Kaiser, 1976; Thompson, 1987; Meehl, 1978; Hays, 1981) that if sample size is big enough, any study can be made to show significant results by statistical significance tests.

Thompson (1993) suggested three alternatives to supplement statistical significance tests. They are: (1) Evaluating result importance by consulting effect sizes; (2) Evaluating results in a sample size context; and (3) Interpreting results based on likelihood of replication. Other suggestions include reporting confidence intervals (Serlin, 1996; Cohen, 1990), interpreting statistical significance tests results with respect to the data (carver, 1996), using range versus point null hypothesis (Serlin, 1996), and so forth. So far, statistical significance tests are still widely used and accepted by the majority of researchers. The focus is on how to interpret its results and provide more background information.

Large Sample Size or Small Sample Size

"How many samples should be included in a study" is a question asked by researchers early in the design of the study. More frequently, the question also concerns the number of subjects needed in order to come up with significant results (Torabi, 1990; Austin, 1983). That sample size affects statistical analyses is a common agreement (e.g. Hay, 1953; Stolurow & Frinke, 1966; Aleamoni, 1973; Torabi, 1990). Even a "stupidest man" knows that the larger one's sample of observation the more confidence one can have in being close to the truth about the phenomena observed (Sedlmeier & Gigerenzer, 1997).

Yet, as several authors indicated (Torabi, 1990; Tuckman, 1978) the primary concern in selecting study samples is not necessarily sample size but representative of the population within acceptable error limits. A representative sample keeps the same characteristics of the population with a margin of error. It is clear that the smaller the sample the less likely one is to obtain a true picture of what is studied. But if the sample is not representative of the target population, a large sample can produce a very misleading result (Torabi, 1990). In other words, if a sample is not randomly selected, no matter how large, it has problems, and generalization from the sample is not possible. If the sample is randomly selected, a large sample is preferable to a small one because it reduces sample error. Therefore, sample size must be determined in relationship to the error that will be tolerated.

Different formula have been developed to calculate desired sample size (Blalock, 1979; Swisher & McClure, 1984; Busha & Harter, 1980; Hopkins & Glass, 1984; Kuzma, 1984). Factors involved vary including desired confidence level, acceptable degree of sampling error, estimated variability in the population, the power of the statistics tests, type of tailed tests used (one-tailed versus two-tailed), and the design of the study (Blalock, 1979; Swisher & McClure, 1984; Busha & Harter, 1980; Hopkins & Glass, 1984; Kuzma, 1984; Torabi, 1990; Austin, 1983). Torabi (1990) suggested a simple formula in determining sample size in normal distribution as follows.

$$\overline{X} = \frac{\overline{X} - m}{s/\sqrt{n}} \quad \text{or} \quad n = \left(\frac{Z\,s}{\overline{X} - m} \right)$$

Here, X-μ is the precision level, the difference between sample mean and population mean, ó is the standard deviation, and Z is determined by the level of confidence. When given X-μ=5, ó=10, Z=1.645(90% confidence level), the sample size is 10.82 or 11 if rounding to the nearest number. The size is small due to a low confidence level, small variability and low level of precision. If a researcher is interested in higher precision such as X-μ=1 and a high confidence level such as 99%, the sample size is much higher (n=663.06 or 664).

Response rate is also a consideration in the determination of sample size discussed by a number of authors (Kalton, 1983; Brown, 1986). Suppose the above sample size will have a response rate of 75 percent. Then the selected sample must be 664/0.75= instead of 664. In this case, the desired sample size is obtained. But nonresponse bias will remain a problem. Nearly all surveys report something less than 100 percept response rates (Brown, 1986). Obviously, any response rate less than 100 percent increases the error limit and subsequently reduces the generalizability of the results. Brown (1986), Babbie and Huitt (1979), and Kish (1965) suggest the following two evidences for researchers to provide to reduce the error limit concerns. (1) The nonrespondents are systematically different from the respondents; and (2) The respondents are systematically different from the sample or from the total population.

Another simple way to find a needy sample size is the use of sample size tables. Available tables can be found in the articles by Krejcie and Morgan(1970) and in NEA Research Division (1960). However, selecting a

sample size also depends on a number of non-statistical factors such as time to conduct the survey, cost of collecting the data, and type of survey method used (Brown, 1986). Some researchers worked on how to reduce sample size while still obtaining valid data. For example, it is possible to use more reliable measure to reduce sample size requirement (Leon, Marzuk, & Portera, 1995), to eliminate outliers to produce a result unaffected by sample size (Van-Selst & Jolicoeur, 1994), to increase effect size by decreasing within-group variance without increasing sample size (Kraemer, 1991), and so forth. Finally, if a survey had a small sample size and negative results, the effect size of the treatment and the power of the statistical test should be reported (Baer and Ahern 1993).

It is noticeable that sample size for qualitative research is becoming a concern in health education research. A common misconception is that the number of samples in qualitative research is unimportant (Sandelowski, 1995). Studies have found some qualitative research studies with too small sample size to support their claims (e.g. Lincoln & Guba, 1985; Strauss & Corbin, 1990). In general, sample size for qualitative research depends on the qualitative methods used in the study (Sandelowski, 1995). More (1994) recommended using 6 participants for phenomenological studies, 30 to 50 interviewers for ethnographies and grounded theory studies, and 100 to 200 units for qualitative ethnological studies. A principle provided by Sandelowski (1995) states that "an adequate sample in qualitative research is one that permits the deep, case-oriented analysis that is hallmark of all qualitative inquiry, and that results in a new and richly textured understanding of experience "(p. 183).

Epilogue

Our profession has been advanced in a significant way due in part by researchers who have conducted scientific studies and evaluational investigations related to health education and health promotion. The younger investigators should follow the path of our established researchers in the field who have utilized sound methodology in conducting and completing their research projects. Knowing the fact that there is no perfect research, we all need to improve upon our research, measurement, and statistical skills.In addition we need to continuously retool ourselves with the new technology increasingly becoming available to our profession.

In selecting a statistical or measurement technique over other options, researchers need to weigh advantages and disadvantages and make an objective decision. Obviously, there is no short cut in conducting scientific research. However, due to certain budgetary or human subject restrictions, one may choose a convenient approach over an ideal one. These decisions will ultimately impact the findings and extent of any inference one can make from the findings. For instance, selecting convenience sampling method over a stratified sampling technique will impact the degree of generalizability of the finding. Consequently, in deciding which techniques to utilize, one needs to be fully informed of the method and issues involved. This paper attempted to highlight some of those variable and issues that may influence the decisions made by younger researchers during the course of conducting their research projects.

References

Aleamoni, L. M. (1973). Effects of size of sample on eigenvalues, observed commonalties, and factor loading. Journal of Applied Psychology, 58, 266-269

Anastasi, A. (1988). Psychological testing (6th ed.). New York, NY: Macmillan

Austin, H. W. (1983). Sample size: How much is enough? Quality and Quantity, 17, 239-245

Babbie, E. R., & Huitt, R. E. (1979). Practicing social research (2nd ed.). Belmont, CA: Wadsworth

Baer, L., & Ahern, D. K. (1993). Statistical problems with small sample size. American Journal of Psychiatry, 150(2), 356

Bakan, D. (1967). One method: Toward a reconstruction of psychological investigation. San Francisco, CA: Jossey-Bass

Baranowski, T, Dworkin, R., Cieslik, C. J., & Hooks, P. (1984). Reliability and validity of children's self-report of aerobic activity: Family health project. Research Exercise Report, 55(4), 309-317

Baranowski, T. (1985). Methodological issues in self-report of health behavior. Journal of School Health, 55(5), 179-182

Basch, E. (1987). Focus group interview: An underutilized research technique for improving theory and practice in health education. Health Education Quarterly, 14, 441-448

Baumgartner, T. A., & Strong, C. H. (1994). Conducting and reading research in health and human performance. Dubuque, IA: Wm. C. Brown Communications, Inc.

Beland, F., Maheux, B., & Lambert, J. (1991). Measurement of attitudes and behaviors in public health surveys. American Journal of Public Health, 81(1), 103-105

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. Journal of Educational Measurement. 22, 231-240

Bhola, H. S. (1990). Evaluating "literacy for development" projects, programs and campaigns. Hamburg, German: UNESCO Institute for Education and German Foundation for International Development

Blalock, H. M. (1979). Social Statistics. New York, NY: McGraw-Hill

Bradley, J. V. (1968). Distribution free statistical tests. Englewood cliffs, NJ: Prentice-Hall

Bradley, J. V. (1977). Bizarre distribution shapes. American Statistician, 31, 147-150

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 114-152

Brown, F. G. (1970). Principles of educational and psychological testing. New York: Holt, Rinehart and Winston

Brown, K. W. (1986). Action research in business education. Business Education Forum, February, 31-34

Busha, C. H., & Harter, S. P. (1980). Research methods in liberarianship: Techniques and interpretation. San Diego, CA: Academic

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

Carver, R. P. (1996). The case against statistical significance testing, revisited. Journal of Experimental Education, 61(4), 287-292

Catania, J. A., Gibson, D. R., Chitwood, D. D., & Coates, T. J. (1990). Methodological problems in AIDS behavior research: Influences on measurement error and participation bias in studies of sexual behavior. Psychology Bulletin, 108(3), 339-362

Cohen, S. A. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312

Conover, W. J. (1980). Practical nonparametric statistics. New York, NY: Wiley

Cook, T. D., & Campbell, D. T. (Eds.). (1979). Quasi-experimentation, design and analysis issues for field settings. Chicago: Rand McNally

Cozby, P. C. (1985). Methods in behavioral research. Palo Alto, CA: Mayfield Publishing Company

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. American Psychologist, 30, 116-127

Crowne, D. P., & Marlow, D. (1960). A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology, 24, 349-354

Daniel, W. W. (1977). Statistical significance versus practical significance. Science Education, 61(3), 423-427

Dixon, W. J. (1954). Power undernormality of several nonparametric tests. Annals of Mathematical Statistics, 25, 610-614

Ebel, R. L. (1980). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall Inc.

Edwards, A. L. (1957). The social desirability variable in personality assessment and research. New York: Dryden Press

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitudes, intention, and behaviors. Journal of Applied Psychology, 73, 421-435

Gardner, P. L. (1975). Scales and statistics. Review of Educational Research, 45, 43-57

Gebhardt, M. E. (1980). Health education evaluation: An alternate research paradigm. Evaluation and the Health Professions, 3(2), 205-210

Gibbons, J. D. (1971). Nonparametric statistical inference. New York, NY: McGraw-Hill

Gibbons, J. D. (1993). Nonparametric statistics: An introduction. New Bury Park, CA: Sage Publication, Inc.

Gill, D. H., McNamara, J. F., & Skinkle, J. D. (1980). The practical significance of research reported in the journal of industrial teacher education. Journal of Industrial Teacher Education, 17(2), 5-19

Glass, G. V., & Hopkins, K. D. (1984). Statistical methods in education and psychology (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall

Green, L. W., & Lewis, F. M. (1986). Measurement and evaluation in health education and health promotion. Palo Alto, CA: Mayfield Publishing Company

Grounlund, N. E. (1981). Measurement and evaluation in teaching. New York: Macmillan Publishing Co., Inc.

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley

Harwell, M. R. (1988). Choosing between parametric and nonparametric tests. Journal of Counseling and Development, 67, 35-38

Hay, E. (1953). A note on small samples. Journal of Applied Psychology, 37, 445.

Hays, W. L. (1963). Statistics for psychologists. New York: Holt, Rinehart and Winston

Hays, W. L. (1981). Statistics (3rd ed.). New York, NY: Holt, Rinehart and Winston

Hodges, J. L., & Lehman, E. L. (1956). The efficacy of some nonparametric competitors of the t test. Annals of Mathematical Statistics, 27, 324-335

Hopkins, K. D., & Glass, G. V. (1984). Statistical methods in education and psychology. Englewood Cllifs, NJ: Prentice-Hall

Horrocks, J. E. (1964). Assessment of behavior. Columbus, OH: Charles E. Merrill Books, Inc.

House, E. R. (1980). Evaluating with validity. Beverly Hills, CA: Sage

Hunter, M. A., & May, R. B. (1993). Some myths concerning parametric and nonparametric tests. Canadian Psychology, 34(4), 384-389

Israel, B. A., Cummings, K. M., Dignan, M. B., Heaney, C. A., Perales, D. P., Simons-Morton, B. G., & Zimmerman, M. A. (1995). Evaluation of health education programs: Current assessment and future directions. Health Education Quarterly, 22(3), 364-389

Ittenbach, R. F., Chayer, D. E., Bruininks, R. H., Thurlow, M. L., & Beirne-Smith, M. (1993). Adjustment of young adults with mental retardation in community settings: Comparison of parametric and nonparametric statistical techniques. American Journal on Mental Retardation, 97(6), 607-615

Jenkins, S. J., & Fuqua, D. R. (1984). Evaluating criteria for selection of nonparametric statistics. Perceptual and Motor Skills, 58, 979-984

Kaiser,H. F. (1976). Review of factor analysis as a statistical method. Educational and Psychological Measurement, 36, 586-589

Kalton, G. (1983). Introduction to survey sampling. Newbury Park, NJ: SAGE publications

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

Kish, L. (1965). Survey sampling. New York, NY: John Wiley

Kerlinger, F. N. (1964). Foundations of behavioral research; educational and psychological inquiry. New York, NY: Holt, Rinehart and Winston

Kerlinger, F. N. (1979). Behavioral research: A conceptual approach. New York: Holt, Rinehart and Winston.

Kraemer, H. C. (1991). To increase power in randomized clinical trials without increasing sample size. Psychopharmacology Bulletin, 27(3), 217-224

Krejcie, R. V., & Morgan, D. (1970). Determining sample size for research activities. Educational and Psychological Measurement, 30, 607-610

Kuzma, J. W. (1984). Basic statistics for the health sciences. Palo Alto, CA: Mayfield Publishing Co.

Kuzma, J. W. (1992). Basic statistics for the health sciences. Mountain View, CA: Mayfield Publishing company

Lehmann, E. L. (1975). Nonparametrics: Statistical methods based on ranks. San Francisco, CA: Holden-Day

Leon, A. C., Marzuk, P. M., & Portera, L. (1995). More reliable outcome measures can reduce sample size requirements. Archives of General Psychiatry, 52(10), 867-871

Levin, J. R. (1996). Statistical significance testing from three perspectives. Journal of Experimental Education, 61(4), 378-382

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Beverly Hills, CA: Sage

Lincoln, Y. S. (1985). Naturalistic Inquiry. Newbury Park, CA: Sage

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. Contemporary Psychology, 36, 102-105

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. Developmental Psychology, 22(1), 37-49

May, C., & Foxcroft, D. (1995). Minimizing bias in self-repots of health beliefs and behaviors. Health Education Research Theory and Practice, 10(1), 107-112

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. Philosophy of Science, 34, 103-115

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834

Mehrens, W. A., & Lehmann, I. J. (1991). Measurement and evaluation in education and psychology (4th ed.). New York, NY: Holt, Rinehart and Winston, Inc.

Miller, E. H. (1990). The positive health index: Development and psychometric evaluation. Chicago: University of Illinois at Chicago

Miller, T., & Cleary, T.A. (1993). Direction of wording effects in balanced scales. Educational and Psychological Measurement, 53(1), 51-60.

More, J. M. (1994. Designing funded qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), Handbook of qualitative research (pp. 220-235). Thousand Oaks, CA: Sage

Mueller, D. J. (1985). Measuring social attitudes: A handbook for researches and practitioners. Bloomington, IN: Indiana University

Morison, D. E., & Henkel, R. E. (Eds.). (1970). The significance test controversy: A reader. Chicago, IL: Aldine

NEA Research Division. (1960). Small sample techniques. New Research Bulletin, 99-104

Noblitt, O. W., & Hare, R. D. (1988). Meta-Echnography: Synthesizing qualitative studies. Newbury Park, CA: Sage

Noll, V. H., & Scannell, D. P. (1972). Introduction to educational measurement. Boston, MA: Houghton Mifflin Co.

Papenfuss, R. & Beier, B. J. (1984). Developing, implementing, and evaluation a wellness education program. Journal of School Health, 54(7), 360-362

Patton, M. Q. (1980). Qualitative evaluation methods. Beverly Hills, CA: Sage

Patton, M. Q. (1987). How to use qualitative and quantitative methods in evaluation research. Newbury Park, CA: Sage

Phillips, D. L. (1971). Knowledge from what? Chicago, IL: Rand McNally

Reichardt, C. S., & Cook, T. D. (1979). Beyond qualitative versus quantitative methods. In T. D. Cook & C. S. Reichardt (Eds.), qualitative and quantitative methods in evaluation research. Beverly Hills, CA: Sage

Rosenberg, A. (1988). Philosophy of social science. Boulder, CO: Westview Press Inc.

Rulon, P. J. (1964). Validity of educational tests. Test Service Notebook No. 3. Yonkers: world Book Co.

Sandelowski, M. (1995). Sample size in qualitative research. Research in Nursing and Health, 18, 179-183

Savage, I. R. (1957). Nonparametric statistics. Journal of the American Statistical Association, 52, 331-344

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. Educational and Psychological Measurement, 41, 1101-1114

Sedlmeier, R. & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. Journal of Behavioral Decision Making, 10(1), 33-51

Seeman, J. (1973). On supervising student research. American Psychologist, 28, 900-906

Serlin, R. C. (1996). Confidence intervals and the scientific method: A case for Holm on the range. Journal of Experimental Education, 61(4), 350-360

Shaver, J. P. (1996). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61(4), 293-316

Siegel, S. & Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences. New York, NY: McGraw-Hill Book Company

Siegel, S. (1956). Nonparametric statistics for the behavioral science. New York, NY: McGraw-Hill

Shiffman, S. M. & Jarvik, M. E. (1976). Smoking withdrawal symptoms in two weeks of abstinence. Psychopharmacology, 50, 35-39

Skinner, B. F. (1956). A case history in scientific method. American Psychologist, 11, 221-223

Snedecor, G. W. & Cochran, W. G. (Eds, 1967). Statistical methods .Ames, IW: the Iowa State University Press.

Steckler, A. (1989). The use of qualitative evaluation methods to test internal validity. Evaluation and the Health Professions, 12, 115-133

Steckler, A., McLeroy, K. R., Goodmab, R. M., Bird,, S. T., & McCormic, L. (1992). Toward integrating qualitative and quantitative methods: An introduction. Health Education Quarterly, 19(1), 1-8

Stevens, S. S. (1946). On the theory of scales of measurement. Science, 103, 677-680

Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. Science, 161, 849-856

Stolurow, L. & Frinke, G. (1966). A study of sample size in making decisions about instructional materials. Educational and Psychological Measurement, 26, 643-649

Strauss, A. & Corbin, J. (1990). Basics of qualitative research: Grounded theory procedures and techniques. Newbury Park, CA: Sage

Sweeney, C. T., Pillitteri, J. L., & Kozlowski, L. T. (1996). Measuring drug urges by questionnaire: Do not balance scales. Addictive Behavior, 21(2), 199-204

Swisher, R., & McClure, C. R. (1984). Research for decision making. Chicago, IL: ALA

Taylor, J. C., & Bowers, D. G. (1972). Survey of organizations: A machine-scored standardized questionnaire instrument. Ann Arbor, Michigan: Institute for Social Research, University of Michigan

Thompson, B. (1987). The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice. Paper presented at the annual meeting of the American Education Research Association, Washington, DC

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61(4), 361-377

Tiffany, S. T., & Drobes, D. J. (1991). The development and initial validation of a questionnaire on smoking urges. British Journal of Addiction, 86, 1467-1476

Torabi, M. R. (1986). How to estimate practical significance in health education research. Journal of School Health, 56(6), 232-234

Torabi, M. R. (1988). Factors affecting reliability coefficients of health attitude scales. Journal of school health, 58(5), 186-189

Torabi, M. R. (1990). The question of sample size. Health Values, 14(5), 53-56

Torabi, M. R. (1995). Critical issues in health education program evaluation: Implications for the U.S. Health objectives for the year 2000. HPER Dimension. Bloomington, IN: Indiana University

Tuckman, B. W. (1978). Conducting educational research (2nd ed.). New York: NY: Harcourt Brace Jovanovich

Udry, J., & Morris. N. (1967). A method for validation of reported sexual data. Journal of Marriage and the Family, 442-446

Van-Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. Quarterly Journal of Experimental Psychology Human Experimental Psychology, 47A(3), 631-650

Vries, H. D., Weijts, W., Dijkstra, M., & Kok, G. (1992). The utilization of qualitative and quantitative data for health education program planning implementation and evaluation: A spiral approach. Health Education Quarterly, 19(1), 101-115

Wainer, H. (1986). Can a test be too reliable? Journal of Education Measurement, 23(2), 171-173

Weimer, R. C. (1993). Statistics. Dubuque, IA: Wm. C. Brown Publishers

Yin, R. K. (1984). Case study research: Design and methods. Beverly Hills, CA: Sage