DOCUMENT RESUME

ED 359 201 TM 019 326

AUTHOR Dwyer, Evelyn E.

TITLE Attitude Scale Construction: A Review of the Literature.

PUB DATE [93]

NOTE 48p.

PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Attitude Measures; *Definitions; Estimation (Mathematics); Literature Reviews; Mathematics Anxiety; Mathematics Tests; *Rating Scales; *Teacher Attitudes; *Test Construction; Test Reliability; Test Validity

ABSTRACT

The literature was examined to find studies specifically concerned with the development of instruments designed to measure attitudes. Close attention was paid to procedures used to determine reliability and validity in the development and subsequent use of such instruments. An effort was made to select the most relevant literature from the vast number of studies undertaken to determine attitudes. The review is organized into the following categories: (1) definitions and components of attitudes; (2) the measurement of attitude; (3) techniques for attitude scale construction; (4) test construction statistics; and (5) mathematics related attitude scales. Attitude is often considered to have cognitive, affective, and behavioral components. Information about attitudes is usually gathered through observation and through self-report measures. Advantages and problems of each approach are discussed. An overview is presented of four basic techniques of attitude scale construction: Thurstone scales, Likert scales, Guttman scales, and semantic differential scales. Methods for estimating reliability and validity are reviewed. Attitude scales related to mathematics include those for mathematics anxiety, attitudes toward mathematics, and teacher attitudes. Contains 132 references. (SLD)

ATTITUDE SCALE CONSTRUCTION:

A REVIEW OF LITERATURE

By Evelyn E. Dwyer, Ph.D.

Assistant Professor, Mathematics Dept.
Walters State Community College
500 South Davy Crockett Pkwy.
Morristown, Tennessee 37813-6899

ATTITUDE SCALE CONSTRUCTION:

A REVIEW OF LITERATURE

The literature was examined to find studies specifically concerned with the development of instruments designed to measure attitudes. Close attention was paid to procedures used to determine reliability and validity in the development and subsequent use of such instruments. An effort was made to select the most relevant literature from among the vast number of studies undertaken to determine attitudes toward a great many objects. This review is organized into the following categories: (1) Definitions and components of attitude, (2) The measurement of attitude, (3) Techniques for attitude scale construction, (4) Test construction statistics and, (5) Related attitude scales.

Definitions and Components of Attitude

Definitions

Among the more commonly accepted definitions of attitude are the following:

> An attitude is a mental and neural state of readiness, organized through experience, exerting a directive or dynamic influence upon the individual's response to all objects and situations with which it is related.
> (Allport, 1935, p. 810)

> An attitude can be defined as an enduring organization of motivational, emotional, perceptual, and cognitive processes with respect to some aspects of the individual's world.
> (Krech and Crutchfield, 1948, p. 152)

> An individual's social attitude is a syndrome of response consistency with regard to social objects.
> (Campbell, 1950, p. 31)

> An attitude is an idea charged with emotion which predisposes a class of actions to a particular class of social situations.
> (Triandis, 1971, p. 2)

The definition of attitude proposed by Triandis suggests that attitude has three components: (a) a cognitive component (the idea), (b) an affective component (the emotions), and (c) a behavioral component (the action). Further, the definition of attitude proposed by Krech and Crutchfield (1948) described above also implies three similar components of attitude. A discussion concerning components of attitude is presented below:

Components

The cognitive component of attitude was described by Triandis (1971) as the ideas or beliefs that subjects have about an attitudinal object, the object, in this context, being the focal point of attention. The affective component was described as the emotions or feelings about the attitudinal object, while the behavioral component was described as predisposition to action with regard to the same object. Triandis found the behavioral component measurable though direct observation of overt actions or through analysis of verbal statements concerning intended behavior. Triandis further indicated that, although the three components are closely related, the components can appear to be inconsistent with one another based on overall analysis of attitude scale responses from individuals.

Other researchers refered to three similar subcomponents of attitude and recommended attitude measurement approaches reflecting those subcomponents. In this light, Hassan & Shrigley (1984) categorized attitude scale components as (1) egocentered, (2) social-centered and, (3) action-centered. The three item types suggested by Hassan and Shrigley appear similar to the affective, cognitive, and behavioral components of attitude described by Triandis (1971). Likewise, Chein (1948) and Harding (1954) discussed attitudes in terms of three components. In a similar approach, Greenwald (1968,) also described the three components of attitude as "affects, cognitions, and action tendencies" (p. 363).

Fishbein & Ajzen (1975), suggested a classification of four components of attitude rather than the more commonly used three. While maintaining affect (feeling), and cognition (belief), Fishbein & Ajzen divided the behavioral component into two parts: the actual behavior (observed overt acts) and the conation (behavioral intentions). Further, these researchers concluded that if attitude must be measured as a single dimension and reported in a single score, it is most accurately measured through the affective part of the attitude concept. The last contention of Fishbein and Ajzen is consistent with the apparent widespread agreement among researchers that, although affect cannot capture the full complexity of the attitude concept, it is the most essential, consistent, stable and reliable measure of attitude.

The Measurement of Attitudes

Information about attitude can be gathered in two basic ways: through observing subjects and/or by asking subjects what they believe. In this light, Anderson (1981) stated that information is gathered about attitude or any affective characteristic though observational methods and/or through self-report methods. The purpose of this section is to present information about observational and self-report methods of attitude assessment and to highlight advantages and problems inherent in each.

Observational Methods

Using observational methods for obtaining information about attitude is based on the assumption that it is possible to infer attitude from the observation of overt behavior or physiological reactions (Fox, 1969; Anderson, 1981). Three major problems are reportedly inherent in observational research methodology:

1. The problem of inaccurately inferring affective characteristics from overt behavior.
2. The problem of determining which behaviors to observe and how to accurately record those behaviors.
3. The problem of misinterpreting the behavior noted by the observer.

Anderson (1981) proposed potential solutions to the three problems inherent in observational methods of obtaining information about attitude. For the first problem related to making inferences, Anderson suggested that correct inferences are more likely to be made if multiple observations are made of the same behavior in a variety of settings or over time in the same setting. With regard to problem number two, observing relevant behaviors, Anderson suggested that appropriate inferences can be made if the affective characteristics are clearly defined at the outset and care is taken to observe only those clearly defined behaviors in an appropriate context. With regard to the third problem, that of misinterpreting behaviors, Anderson suggested using more than one carefully trained observer in the same setting to minimize misinterpretation.

Purkey, Cage, and Graves (1973) assessed affective characteristics of 357 students at two elementary schools. The researchers designed a measure they called the Florida Key. Teachers of the 357 subjects were asked to evaluate their students based on observations of the students behaviors. In the Florida Key, 18 behaviors were designated for evaluation and subsequent analysis. While the researchers reported only a modest relationship between the Florida Key and a self-report measure of affective characteristics, the study presents an interesting research design pairing observational research with quantitative research methodology. Further descriptions and presentations of data concerning data collected through observing subjects are presented by Cook and Sellitz (1964), Lemon (1973), and Crano and Brewer (1973).

The measurement of attitude through observation of physiological reactions was studied by, Porier & Lott (1967), Westie & DeFleur (1959), Woodmansee (1970), and Mueller (1970). Such techniques are based on the assumption that there is a close relationship between physiological responses and affective states. Researchers noted that autonomonic responses might function as valid indicators of strong attitude but might be insensitive to less extreme attitudinal reactions. Further, researchers generally have noted that the ability to determine the directionality of response through analysis of physiological reactions is extremely limited. The two main types of physiological responses discussed in the literature are the Galvanic Skin Response (GSR), a calculation of the amount of electrical conductance of the skin, and pupillography, a measure of change in reaction of the pupil in the eye to various attitudinal stimuli.

Self-Report Methods

Self-report methods of attitude assessment are usually a series of questions, adjectives, or statements about an attitudinal object. Respondants are asked to read and react to each question, adjective, or statement about an attitudinal object in terms of agreement or disagreement. Responses are then scored in terms of positiveness toward the attitudinal object. In some instances, responses are summed to attain a total score.

According to Anderson (1981), the major difficulty associated with self-report methods of attitude assessment is that subjects may provide misinformation to the researcher. Anderson contends that misinformation is sometimes supplied to the researcher when individuals respond to a question, statement, or adjective in a way they think will be socially acceptable to the researcher or when they respond in an acquiescent manner. Acquiescence, in this instance, refers to the tendency of an individual to agree with a question, statement or adjective when they are actually unsure of their response. Thurstone & Chave (1929, pg. 10) considered the issue of misinformation and offered the following advice to researchers:

> All that we can do with an attitude scale is to measure the attitude expressed with the full realization that the subject may be consciously hiding his true attitude or that the social pressure of the situation made him really believe what he expresses...All we can do is minimize as far as possible the conditions that prevent our subjects from telling the truth, or else to adjust our interpretation accordingly.

Selected Techniques of Attitude Scale Construction

The four major types of attitude scales described in the literature were: Thurstone scales (Thurstone and Chave, 1929); Likert scales (Likert, 1932); Guttman scales (Guttman, 1944); and semantic differential scales (Osgood, Suci, and Tannenbaum, 1957). An overview of each of the four attitude measurement scale types is presented below:

Thurstone Technique

Thurstone & Chave (1929), developed the method of equal-appearing intervals to measure attitudes. According to Thurstone, the essential characteristic of the method of equal-appearing intervals is the series "...of evenly graduated opinions so arranged that equal steps or intervals on the scale seem to most people to represent equally noticeable shifts in attitude" (pg. 554). Edwards (1957) reported on the usefulness of the method of equal-appearing intervals especially when a large number of statements must be scaled. Edwards further described the method of equal-appearing intervals as much preferable to the earlier more laborious paired-comparison technique of attitude scale construction also introduced by Thurstone in 1927.

Procedures.

Using the method of equal-appearing intervals developed by Thurstone and Chave (1929), opinions about an attitudinal object can be collected from designated samples and from related academic literature. The collected opinion statements about the object of focus can then be edited. The editing process is undertaken to select statements covering the widest possible range from the most intensely negative to the most intensely positive attitudes toward the object. The selected items are each printed on a separate slip of paper and subjects (sometimes called "judges" in the literature) are given a copy of each item.

The subjects are asked to sort the items into 11 piles representing an evenly graduated series of attitudes from extremely negative (pile 1) through extremely positive (pile 11) toward the attitudinal object. After sorting, data are tabulated to show how each subject placed every one of the statements. Figure 1 shows the method used by Thurstone and Chave to summarize the sorting of items by subjects. The first column gives the item number. The second and third column contain, respectively, the scale value and the Q value (see "Scale and Q Values" below). The remaining columns, progressing from left to right, give the cumulative frequency of times the specified item was placed in each pile by subjects.

Item	Scale	Q	A	B	C	D	E	F	G	H	I	J	K
#	Value		1	2	3	4	5	6	7	8	9	10	11
1	9.9	2.4	.00	.00	.00	.00	.00	.08	.17	.23	.33	.52	1.00

Figure 1. Hypothetical example of how an item is sorted using the equal-appearing intervals technique.

Scale and Q Values.

A scale value for each item was determined graphically by Thurstone & Chave (1929). Considering each item separately, the cumulative proportion of responses to an item (Y axis) was plotted against the corresponding sorted scale values for the same item (X axis). An overall scale value for the item was then determined by locating the item's median scale value.

After, Thurstone & Chave located the scale value and the upper and lower quartile response scores of each item, an overall Q value was then determined for that item. The Q value was calculated by subtracting the lower quartile score from the upper quartile score. The Q value was considered to be a measure of ambiguity and also a measure of dispersion of judgments for an item. If the dispersion of judgements for a statement is high in comparison with other statements, the statement would be considered ambiguous and eliminated from consideration on the final scale. The Q value has also been referred to in the literature as the semi-interquartile range (Guilford, 1965). Guilford defined the semi-interquartile range as one-half the range of the middle 50 per cent of judgements about an item: $Q = (C_{75} - C_{25})/2$.

The Final Thurstone Scale.

After considering and comparing the scale and Q value of every item, and after giving logical consideration to the content of every item, Thurstone & Chave (1929) selected items for inclusion on their final attitude scale. The statements selected approximated as closely as possible a uniformly graduated series of scale values. The scale was then presented to subjects who were asked to place a check mark beside each statement with which they agreed.

Thurstone & Chave (1928) described two methods of scoring the equal appearing interval attitude scale. The first method involved summing the scale-values of all items with which a subject agreed and then obtaining the arithmetic mean of those scale values. The second method of scoring a subject's set of responses consisted of assigning a numerical rank to each of the items on the scale. The rank values of all items with which a subject agreed were summed and an arithmetic average determined.

Likert Scales

Likert scales are an extremely popular method for measuring attitude (Oppenheim, 1966; Crano & Brewer, 1973; and Anderson, 1981). The researchers cited above the Likert method of scale construction is less laborious than the Thurstone technique. Further, the researchers suggested that it is the most efficient and effective method of developing highly reliable scales.

The Likert Scale was developed by Rensis Likert (1932). Likert's primary concern for such a scale was that it measure a unidimensional construct, that is, that all items measure the same thing. Edwards (1957) and Sellitz (1959) referred to the Likert scaling technique as the method of summated ratings because the total score for each subject is obtained by summing the subject's response to each item. The summated score, therefore, represents the degree of favorable or unfavorable attitude toward the object under consideration. Components and strategies for developing Likert scales are presented below:

Procedures.

Items should be clearly favorable or unfavorable with regard to the attitudinal object. Likert (1932) determined it desirable to prepare and select more statements than are likely ever to be used, since many of the items would be found unsatisfactory for the intended purpose of an instrument. Years later, Lemon (1973) suggested using approximately the same number of positive and negatively stated items in a Likert scale. However, other researchers, including Hassan & Shrigley (1984), favored using more negative than positive statements because negatively stated items "are less prone to withstand the rigor of Likert's item analysis" (pg. 660).

After preliminary items on the Likert scale have been written, several judges are asked to classify each item as positive, negative, or neutral with regard to the attitudinal object. Items not classified by the majority of judges as either positive or negative with regard to the attitudinal object are eliminated from consideration for use in the final scale.

A decision must be made relative to the number of response alternatives for each statement. Likert originally used a five response format: 1. strongly disagree, 2. disagree, 3. undecided, 4. agree, and 5. strongly agree. However, modifications in the number of response alternatives are acceptable. A number of response categories ranging from two to seven are described by Anderson (1981) with the even numbered categories yielding a forced choice i.e. no neutral response is possible. Anderson further suggested increasing the number of response categories as a means to strengthen reliability.

The self-report instrument is then administered to a sample of the audience for whom the instrument is intended. Data are analyzed to estimate the validity and reliability of the scale. A revised final scale is then constructed based on conclusions drawn from the data.

Scoring.

The respondant is asked to react to each item in terms of several degrees of agreement or disagreement; for example, (1) strongly agree, (2) agree, (3) undecided, (4) disagree, and (5) strongly disagree. The response alternatives are weighted so the most favorable response carries the highest weight. For example, if a statement is favorable regarding the attitudinal object, "strongly agree" carries the highest weight. On the other hand, if the statement is unfavorable toward the object, then "strongly disagree" carries the highest weight. Consequently, when scoring, the tallies on negative items would be reversed.

Likert's original method of weighted scoring (Edwards, 1957) was based on Likert's conclusion that a normal distribution often results when the five point response system is used. Likert determined the proportion of subjects falling into each of the five response categories for a favorable statement and then calculated the corresponding normal deviate weights i.e. Z score for each item. The overall score was obtained by summing the weights for all items. As mentioned above, the weights were reversed for unfavorable statements so that the strongly disagree category had the highest positive weight for those negative items.

Likert (1932) also devised a less complex method for assigning weights to the five response categories by eliminating the need for Z score transformation. In the simpler system, for favorable items, Likert assigned the "strongly agree" response a weight of 4, the "agree" response a weight of 3, the "undecided" response a weight of 2, the "disagree" response a weight of 1, and the "strongly disagree" response a weight of 0. For unfavorable items, the scoring was reversed. For each respondant, a total score was then obtained by summing all scores for all items.

Item Selection Criteria.

The criterion of internal consistency is commonly used as a method of selecting items for inclusion on a final Likert scale (Likert, 1932; Ferguson, 1981; Crano & Brewer, 1973; Anderson 1981). The criterion of internal consistency is applied by correlating item scores with total scores. Any item with a nonsignificant item to total correlation is eliminated from consideration for use in the final scale. Researchers agree that high correlations between scores on a particular item and total test scores suggest the item represents the attitude under study.

According to Hassan & Shrigley (1984) and Edwards (1957), another test of the validity of a particular item is the discriminating quality of the item. A positively written item is valid only if those individuals with a generally positive attitude toward the attitudinal object agree or strongly agree with the item and if those with a generally negatiave attitude disagree or strongly disagree with the item. The researchers cited above suggested establishing positive and negative criterion groups composed of subjects having the highest and lowest 27% of scores within the overall group being considered. Student t scores would then be calculated comparing the mean score for each criterion group. A significant difference in the mean scores of the two criterion groups would suggest that the item has discriminating quality.

Guttman Scales

Guttman (1944) and Guttman and Suchman (1947) developed what Edwards (1957) suggested is more a procedure for evaluating a set of statements about an attitudinal object rather than an actual attitude scale. Nevertheless, the procedure has become known throughout the literature as the Guttman Scale. A description of the Guttman procedure follows.

Procedures.

In constructing a Guttman scale, according to Crano and Brewer (1973), statements appearing to have the following characteristics are written or selected:

1. Statements have commmon content
2. Statements are ordered along a continuum from least positive to most positive
3. Agreement with a given statement implies agreement with every other less positive statement.

Given an instrument with statements about an attitudinal object meeting the criteria described above, subjects are then instructed to check each statement with which they agree. When a subject agrees with an attitude statement, the subject receives a score of 1 for the item. However, if the individual disagrees with the attitude statement, the subject receives a score of 0 for the item. The subject's total score is the sum of all his/her item scores on the scale. The overall score suggests the subject's degree of favorability toward the attitudinal object. Data are then submitted to Guttman scale analysis.

Scalogram Analysis.

Guttman scale analysis involves the computation of the coefficient of reproducibility (CR) (Guttman, 1944, 1947). The calculation of the coefficient of reproducibility is illustrated through the following hypothetical example: A scale with five statements is administered to a group of subjects. The five statements were written along a continuum from least positive (statement number one) through most positive (statement number five). If the Guttman assumption is met, several potential response patterns are acceptable. All acceptable response patterns illustrate that a person who agrees with statement number five (the most positive), must also agree with statements one through four. Through analysis of all response patterns, the number of errors due to inappropriate responses can be calculated. Figure 2 illustrates an acceptable response pattern for a five item Guttman scale.

Acceptable		**Statement**	**Numbers**		
Pattern	**1**	**2**	**3**	**4**	**5**
A	D	D	D	D	D
B	A	D	D	D	D
C	A	A	D	D	D
D	A	A	A	D	D
E	A	A	A	A	D
F	**A**	**A**	**A**	**A**	**A**

Figure 2. Acceptable response patterns for a five item Guttman scale.

In calculating the coefficient of reproducibility (CR), the total number of errors (deviations from acceptable patterns) is counted for all subjects. A percentage of error is then computed by dividing the total number of errors made by all subjects by the total number of responses. The total number of responses refers to the number of subjects multiplied by the number of statements. The CR is then obtained by subtracting the error rate from 10C percent. For example, suppose 25 subjects were administered a five item Guttman scale. The total number of potential responses is 25

* 5 = 125. If the total number of errors made by all the subjects is 15, then the error rate is 15/125 = .12 or 12 %. The CR calcuation: 100% - 12% = 88%. Guttman (1944) suggested that the error rate should be no larger than 10% for the set of statements to be considered an acceptable scale.

The Cornell technique (Guttman, 1947) and the Goodenough technique (Goodenough, 1944) are two prominent methods of scalogram analysis. Both the Cornell and Goodenough scalogram methods calculate the percent of accuracy the data obtained from responses to a Guttman attitude scale can be reproduced from the total scores. For example, if the coefficient of reproducibility for a scale is .88, this means that 88% of the subjects' responses could be predicted from knowledge of total test scores alone.

Scalogram analysis can also be generalized to more than two categories of response. For example, three categories of response can be used such as: agree, undecided, and disagree with weights of 2, 1, and 0 assigned, respectively. A more comprehensive description of these procedures is found in Edwards (1957).

Semantic Differential Scale

The semantic differential technique was introduced by Osgood, Suci, and Tannenbaum (1957) for measuring attitude. This technique is adjective based and measures reactions of subjects to pairs of bi-polar adjectives with meanings as nearly opposite as possible (Osgood, 1952). Examples might include: good-bad, happy-sad, etc. The semantic differential

(SD) measures directionality of a reaction and also intensity of reaction (Osgood & Suci, 1955). Heise (1967) reported that ratings on SD scales tend to be correlated around three basic dimensions of response accounting for most of the covariance in ratings: evaluation, potency, and activity (EPA). SD scales generally contain adjectives from all three dimensions. Examples of EPA types might include: Evaluation - good/bad, Potency - hard/soft, Activity - fast/slow. Lists of evaluative adjective pairs are included in a text by Osgood, Suci, and Tannenbaum (1957).

Procedures.

In constructing a semantic differential scale, the name of the attitudinal object is placed at the top of the scale. Then, five to ten emotion laden adjective pairs are chosen and a response sheet is constructed. The bi-polar adjective pairs are placed at different ends of a numerical continuum of seven equal segments. Figure 3 illustrates an example of a semantic differential response sheet.

High Risk High School Students

Good ___: ___: ___: ___: ___: ___: ___ Bad

Dishonest ___: ___: ___: ___: ___: ___: ___ Honest

Weak ___: ___: ___: ___: ___: ___: ___ Stong

Fast ___: ___: ___: ___: ___: ___: ___ Slow

Figure 3. Example of a Semantic Differential Scale

After adjectives are selected and a response scale is constructed, it is administered to a sample. Subjects are instructed to place a check mark along the continuum at the point best describing how they feel about the object presented at the top of the sheet. A check mark near either end of the continuum indicates strong positive or negative feelings, while a center check mark indicates neutral feelings. Positive integer values of one through seven are assigned to each response option with the most favorable attitude toward an object given a weight of seven. The total score on the scale is the sum of the subject's response to each item.

Analysis of Data.

Analysis of data obtained through administration of the semantic differential scale is similar analysis of data obtained from a Likert scale. Correlations between each adjective pair and the total scale score can be computed. Adjective pairs not correlating significantly with the total scale score are eliminated.

A comprehensive description of various statistical procedures available for processing data obtained from administration of a semantic differential scale is contained in a review of related research by Heise (1967). Further, reviews, methodological studies, and validity studies related to the semantic differential technique are found in Snider and Osgood (1969).

Methods for Estimating Reliability and Validity

Methods for estimating reliability and validity of tests will be discussed in this section. In part one, the concept of reliability and an overview of computational procedures related to calculating reliability will be reviewed. In part two, an overview of the types of validity and statistical procedures for calculating validity coefficients will be presented.

Reliability

The reliability of a test indicates the trustworthiness of scores obtained. The reliability of a test is an expression of both the stability and consistency of test scores (Cureton, 1958; Thorndike, 1966; Dick & Haggerty, 1971). Concerning stability, resesearchers determine whether the score obtained for a subject (S1) would be the same if S1 were tested again at a later date. The reliability coefficient then indicates whether the two test scores for S1 are stable indicators of S1's performance. Researchers also consider whether the reliability coefficient estimates the accuracy of S1's true score.

A reliability coefficient is represented by a numerical value between 0 and 1 reflecting the stability of the instrument. To compute reliability coefficients, four basic methods are generally used (Ferguson, 1981):

1. Test-retest method - the same test is administered twice to the same group of subjects with administrations separated by an interval of time

2. Parallel-forms method - an alternative test form is administered to the same group after a period of time.

3. Split-half method - A test is divided into two parts and two scores are obtained. The paired observations are correlated.

4. Internal-consistency methods - based on the average correlation among items and the number of items on a test

In all four of the basic methods mentioned above for approximating reliability, the calculation of correlation coefficients between paired observations is required. Many varities of correlation have been developed for use with different types of variables and for data with special characteristics. An overview and discussion of all correlation coefficients is beyond the scope of this dissertation. However, a few of the more widly used measures of correlation will be briefly presented.

Product-Moment Correlations.

The test-retest and alternate forms methods of estimating reliability are determined based on correlating two sets of test scores. Alternate formulas, derived from standard score form, exist for computing product-moment correlations between test scores. The most widely used product-moment correlation coefficient is the Pearson correlation coefficient (Ferguson, 1981). One form of the Pearson Product Moment, denoted by r, follows:

$$r = \frac{(\Sigma XY) - n\bar{X}\bar{Y}}{(n-1)\ s_x s_y}$$

where n is the number of cases, $\bar{X}$, $\bar{Y}$ are the means, and s_x and s_y are the standard deviations of the two variables.

The split-half method of reliability estimation requires the use of an additional formula (Guilford, 1965; Ferguson, 1981). As mentioned above, in the split-half method the same test is divided into two parts and the scores are correlated. The result is a correlation between scores on tests having half as many items as the original instrument. For example, on a 20 item test, 10 of the items would be correlated with the 10 other items with each set of correlated items having similar content. In effect, correlation would occur between paired scores based on scores from two 10 item tests. However, the reliability for the total 20 item test is needed. Therefore, the use of the Spearman Brown (SB) formula approximates the reliability for the total test. One form of the Spearman Brown formula (Ferguson, 1981) is shown below:

$$r_{tt} = \frac{nr_{11}}{1 + (n-1)r_{11}}$$

Where, n is the ratio of the number of items on the desired test to the number of items on the original test and r is the already obtained reliability for the partial test. The Spearman-Brown formula can also be utilized to estimate reliabilities obtained by the test-retest and alternate forms methods.

Kuder-Richardson.

An internal-consistency measure commonly used to estimate reliability was derived by Kuder and Richardson (1937). The two assumptions underlying use of Kuder-Richardson formulas are: (1) the items are dicotomously scored, that is, items are scored 1 for a correct response and scored 0 for an incorrect resonse; and, (2) the items are unidimensional since they measure the same characteristic.

There are many ways a test can be split in order to compute half-test scores. For each possible split, a different reliability coefficient can be obtained. The Kuder-Richardson formula averages all the possible split half reliability coefficients of a particular test. The basic Kuder-Richardson formula (Guilford, 1954; Ferguson, 1981), referred to as formula 20 or KR-20, is shown below:

$$r_{tt} = \frac{k}{(k-1)} * \frac{(\sigma^2_o - p_i q_i)}{(\sigma^2_o)}$$

where, k is the number of items in the test; p = the proportion of students responding correctly to item i; q = 1 - p, the proportion of students responding incorrectly to item i; σ^2_o = test variance, and $p_i q_i$= sum of p times q for all items.

When individual item statistics are not available, an alternative Kuder-Richardson formula can be used to give a conservative estimate of test reliability (Kuder & Richardson, 1937; Guilford, 1954; Ferguson, 1981). It is reasonable to assume that all test items have approximately the same level of difficulty; therefore, the term pq in the KR-20 formula can be replaced by kpq in the alternative Kuder-Richardson formula, where k is the number of test items.

A special case of the Kuder-Richarson formula, is Chronbach's coefficient alpha (α) (Cronbach, 1951). Coefficient alpha is the basic formula for determining the reliability of test scores based on internal consistency for items not dichotomously scores (Nunnally, 1967). According to Cronbach (1951), the coefficient alpha (α) is the mean of all possible split-half coefficients which can result from different splittings of a test and can be used as an index of inter-item homogeneity..

Validity

Test validity is an indication of how well a test measures what it was designed to measure (Garrett, 1937, 1947; Mehrens & Lehmann, 1980). Validity is always stated in reference to a given group, a given area, or a given circumstance. A test can be valid for one group but inappropriate for another. Validity involves gathering and evaluating information for determining how well a test measures what its authors purport it measures. Other definitions and discussions of validity can be found in works by Lindquist (1942), Guilford (1946), Cureton (1951), and Anastasi (1976).

Types of Validity

Although there are many procedures for determining validity, all aspects of validity are interrelated. Types of validity usually considered when instruments are developed for measuring psychological traits are: 1. content, 2. concurrent, 3. construct, and 4. predictive (Wainer & Braun, 1988). Some of the other types of validity mentioned in the literature are: 1. face, 2. curricular, and 3. differential. The specific approaches for determining validity listed above will be described in the section that follows.

Content Validity

The following definition of content validity was offered by the American Psychological Association (1966, p. 12):

> The test user wishes to determine how an individual performs at present in a universe of situations that the test situation is claimed to represent.

If test items are to have content validity, items should be representative of the characteristic being measured. For example, if teacher attitude toward low achievers in mathematics at the middle school level is being measured, items should be written based on middle school teachers' comments about low achievers in mathematics, on other scales measuring the same characteristic, or on relevant items found in the literature.. In this way appropriateness of test content can be determined.

Predictive and Concurrent Validity

In describing predictive validity the American Psychological Association (1966, pg. 12) stated:

> The test user wishes to forecast an individual's future or to estimate an individual's present standing on some variable of particular significance that is different from the test.

When tests correlate highly with subsequent performance, the tests are said to have predictive validity. Validation of this type sometimes takes a long period of time. For example, the ACT mathematics scores of high school juniors might have predictive validity for grade point average in college freshman mathematics classes. There is no way to determine whether it does other than to wait and see how the subjects perform in college.

Concurrent validity, sometimes termed "immediate predictive validity," correlates a test in the process of being developed with scores obtained from previously established measures. For example, in establishing concurrent validity for an instrument designed to measure mathematics anxiety of pre-service elementary school teachers, a researcher might choose to correlate scores obtained on this measure with scores obtained from the same individuals taking the previously established Mathematics Anxiety Rating Scale (MARS) (Suinn, 1972). By obtaining a significant positive correlative between scores obtained on the two measures, researchers could infer that the anxiety scale written for pre-service teachers does indeed appear to measure mathematics anxiety.

Construct Validity

In defining construct validity, the American Psychological Association (1966, pg. 12) stated:

> The test user wishes to infer the degree to which the individual possesses some hypothetical trait or quality (construct) presumed to be reflected in the test performance.

Construct validity involves formulating a theory of relationships and cannot generally be expressed in terms of one coefficient. Cronbach and Meehl (1959) contend that the following types of evidence, among others, must be taken into consideration when attempting to achieve construct validity: content validity, interitem correlations, intertest correlations, studies of stability over time and after experimental intervention.

Face Validity

This type of validity merely answers the question, "Does the test appear to measure what it purports to measure"? For example, the Math Anxiety Rating Scale (MARS) (Suinn, 1972) appears from the name of the instrument and a perusal of items therein to measure what it was designed to measure, mathematics anxiety.

Curricular Validity

Cronbach (1960) introduced the term "curricular validity." This type of validity required determining if tests are representative of instructional content and reflect goals of instruction. For example, the mathematics teacher who is concerned with students' achievement of specific objectives would make certain that his/her test measures those same objectives.

Differential Validity

Anastasi (1986) defined differential validity as the difference between two correlation coefficients when one measure is correlated with two different measures This procedure is undertaken to determine what a test measures and what it does not measure. For example, as a classification test, an honors level high school calculus achievement test might be administered to all students in the honors calculus class. The results of the classification test could then be correlated with two separate criteria: (a) a test of creative ability and (b) a test of mechanical ability. If the classification test correlates .11 with the creative ability test and .92 with the mechanical ability test, then the differential validity of the classification test would be .92 - .11 = .81.

Computational Procedures

In the preceding section entitled "Reliability," several methods were given for approximating the reliability of a test. Whether using statistical methods applicable to reliability established through the use of alternate forms, test-retest, split-half, or internal-consistency reliability, the correlation coefficient given was obtained through correlating a test in some manner with itself. Correlations can also approximate validity coefficients. When statistical procedures correlate a test (x) and some other external criterion (y), such as another test, then they become calculations of validity coefficients. Statistical procedures for calculating validity coefficients and considerations concerning the choice of statistical procedures are found in works by Ferguson (1981), Guilford (1965), Wainer & Braun (1988), Edwards (1972), Nunnally (1967), Guilford (1954) and, Mehrens and Ebel (1967).

Another procedure, factor-analysis, has been suggested by researchers as a useful indicator of the construct validity of scales (Oppenheim, 1966; Hassan & Shrigley, 1984; Gorsuch, 1974; and Mulaik, 1972). Through the use of factor analysis, researchers can test how well statistical clusterings of items match the intended construct groupings. The clusters of items that appear as a result of factor analysis can be examined to determine if they represent the component or subcomponents of the attitude under study.

Innovations

The Mantel-Haenszel procedure was proposed as a "practical and powerful way to detect test items that function differently in two groups" (Holland, 1985, pg. 129). This statistical application can be used to shed light concerning the effect of experiential background relative to subject reaction to test items. Similarily, other researchers have conducted studies relative to what has become known in the literature as differential item functioning (Thissen, Steinberg, & Wainer, 1988). Methodologies described by the researchers cited above are designed to investigate methods of locating test items likely to be responded to differently based on the characteristics of groups setting them apart from others.

Meta-analysis is another statistical innovation in validity assessment. In relationship to validity, meta-analysis is concerned with quantitative methods for combining evidence from different studies. Wainer and Braun (1988) presented information from a variety of sources concerning the calculation and merits of meta-analysis, including the empirical Baysian approach.

Attitude Scales Related to Mathematics

Analysis of the literature suggests a vast array of studies undertaken to determine attitudes among a variety of samples concerning countless areas of interest. On the other hand, the comprehensive review of the literature has not produced evidence of any substantial study in the realm of measuring teacher attitudes toward low achievers in mathematics, the focus of this study. Therefore, the attitudinal instruments presented in this section relate to the measurement of affective attributes related to mathematics.

Mathematics Anxiety

The Fennema-Sherman Mathematics Attitude Scales (1976) consist of a group of five instruments: (1) Mathematics Anxiety Scale, (2) Attitude Toward Success in Mathematics Scale, (3) Effectance Motivation in Mathematics Scale, (4) Usefulness of Mathematics Scale and, (5) Confidence in Learning Mathematics Scale. The Fennema-Sherman scales are designed for administration to high school students. Item responses for the five tests are obtained on a four point Likert scale. Each test consists of 12 items, half of which are positively worded while the other half are negatively worded. Split-half reliability for the five tests were given by the researchers with coefficients ranging from .87 to .93. The Fennema-Sherman studies were innovative in the suggestion that a psychological trait such as mathematics anxiety, might be a multi-dimensional construct. Investigators found relatively low intercorrelations among test scores obtained through administration of the five instruments mentioned above. The researchers, therefore, concluded that each scale measured a different construct.

In a factor-analytic study of mathematics anxiety, conducted by Ling (1982), the five Fennema-Sherman scales (1976) were administered to 500 college freshman in mathematics courses. In addition to the five Fennema-Sherman scales, subjects were also administered the Short-Form Dogmatism Scale (Troldahl and Powell, 1965), The Adjective Check List (Gough, 1952), and the Test Anxiety Inventory (Spielberger, 1978). The study was designed to investigate mathematics anxiety and the possibility that it might be a multi-dimensional construct related to a variety of personality characteristics. However, after analysis of data, Ling (1982) concluded that mathematics anxiety is a unidimensional construct strongly related to attitude toward mathematics in general but not related to other personality characteristics represented by the instruments administered in the study.

Richardson and Suinn (1972) developed the Mathematics Anxiety Rating Scale (MARS). The scale consists of 98 items describing situations producing varying levels of anxiety to numbers. In the original study, 397 secondary level students responded to the items in the scale. An internal-consistency measure yielded a coefficient alpha of .97, while a test-retest procedure yielded a reliability coefficient of .85. In additional studies, a numerical ability measure was compared with the MARS, producing correlation coefficients suggesting that high levels of mathematics anxiety appear to interfere with achievement in mathematics.

Sandman (1974) developed the Mathematics Attitude Inventory (MAI) designed to measure several constructs related to mathematics: 1. Anxiety Toward Mathematics, 2. Value of Mathematics in Society, 3. Self-Concept in Mathematics, 4. Enjoyment of Mathematics, 5. Motivation in Mathematics, and 6. Perception of the Mathematics Teacher. The total scale contains 48 Likert items with each of the above mentioned subscales represented by eight items. Factor analysis of data obtained from 2,547 eighth and eleventh grade students provided support for the validity of the subscale constructs represented in the total scale.

Attitudes Toward Mathematics

Aiken and Dreger (1961) developed the Math Attitude Scale and the Revised Math Attitude Scale (1974). There are 20 items on the scale written in a Likert format with 10 of the items stated positively and 10 stated negatively. In the original study (Aiken & Dreger, 1961), application of the test-retest procedure yielded a reliability coefficient of .94. In the Aiken and Dreger studies, the Math Attitude Scale was correlated with instruments designed to measure achievement in mathematics, experience with mathematics, and other personality variables. Researchers concluded that attitude toward mathematics appears to be related to achievement and ability in mathematics but not to temperament or other personality variables represented by instruments in the study.

The Dutton Scale (Dutton, 1954) was originally designed as a Thurstone type scale measuring feelings toward arithmetic. The scale was comprised of twenty-two statements with scale values ranging from 1.0 to 10.5 divided equally between favorable and unfavorable statements. In 1954, the test was administered to 289 education majors yielding a test-retest reliability coefficient of .94. The scale was revised by Dutton (1962) and its length reduced to 15 items. With a sample of 127 education majors, the test-retest reliability coefficient on the revised measure was .94. Dutton and Blum (1968) changed the scale again, this time to a Likert format having twenty-five items. The sample in the later study consisted of 346 middle school pupils from four socioeconomic groups. The Dutton-Likert scale yielded a split-half reliability coefficient of .84.

Gladstone, Deal, and Drevdahl (1960) developed a 12 item, modified Likert-type scale for use in studying the effects of remedial mathematics courses on attitude. The items were designed to measure attitudes toward mathematics as compared to attitudes toward other school subjects. No reliability estimates were found for the scale. However, some evidence of predictive validity of the scale items related to subjects' dispositions toward mathematics were found.

Aiken (1974) constructed scales designed to measure enjoyment of mathematics (E Scale) and the value of mathematics (V Scale). The scales were combined into a 40 item Likert-type scale and administered to 190 college freshmean. The internal-consistency reliability, coefficient alpha, for the instrument was found to be .95 for the E Scale and .85 for the V Scale. The correlation coefficient between the E and V scores was .64.

A Mathematics Attitude Inventory was constructed in two forms by Ellinston (1962) using Thurstone's method of equal appearing intervals. The two equivalent forms of the inventory, containing 25 items each, were administered to 755 students in 31 junior and senior high school mathematics classes. The scores were correlated yielding a coefficient of .77 . Teachers were asked to rate the attitude of those same students toward mathematics on a scale of one to nine with nine being the most highest positive score. Data were also obtained relative to current grade in mathematics, overall grade point average, mental ability score, composite achievement and mathematics achievement scores, and percentiles. Teacher rating of student attitude toward mathematics and student scores on the Attitude Inventory correlated moderately (r=.48). However, Inventory scores were significantly correlated with composite achievement test percentile ranks (r=.64). Although other correlation coefficients were obtained, the reported relationships appeared minimal.

Teacher Attitudes

Bowling (1977) developed an instrument containing three subscales designed to measure attitudes of prospective teachers toward mathematics. Aiken's E and V Scales (measuring enjoyment and value of mathematics) were utilized along with a new N scale measuring prospective teachers attitudes toward the nature of mathematics. Bowling randomly organized 48 items from the three scales, Aiken's E and V scales and the N scale, and administered the resulting scale to 126 pre-service teachers. A revised 33 item scale was then administered to 328 prospective and inservice teachers. Coefficient alpha reliabilities ranged from .90 to .95 for the E scale portion, .70 for the V scale, and .85 for the N scale.

McCallon and Brown (1971) developed a semantic differential scale designed to measure attitudes of education majors toward mathematics. The researchers developed 15 items containing bi-polar adjectives placed at both ends of a continuum. The scores of 68 subjects were then correlated with the scores obtained from administration of the Aiken-Dreger Math Attitude Scale and a correlation coefficient of .90 was found .

Childress (1976) conducted studies investigating the relationship between collete students attitudes toward mathematics and student ratings of teachers and courses in mathematics. A questionaire containing 90 items was administered to 204 students enrolled in pre-calculus classes. Subscores were obtained from the 90 items measuring: (1) enjoyment of mathematics, (2) value of mathematics, (3) attitude toward mathematics, (4) teacher ratings, (5) course ratings, and (6) a combination of course and teacher ratings. Findings led Childress to conclude that general attitude toward mathematics was significantly related to course and instructor ratings.

Using the Dutton Scale (1968), Phillips (1973) conducted studies relative to the effect of teacher attitude toward arithmetic on student attitude and achievement in mathematics. In the Phillips study, 306 seventh grade students and 59 teachers were tested. Analysis of data indicated that teacher attitude was significantly related to student attitude but not to student achievement. The study also provided evidence suggesting that the effect of teacher attitude on student attitude and achievement is cumulative. Students appeared to achieve higher in arithemetic if they had a sequence of three teachers with favorable attitudes toward mathematics.

REFERENCES

Aiken, L.R., Jr., (1972). Research on attitudes toward mathematics. Arithmetic Teacher, 19, 229-234.

Aiken, L.R., Jr. (1974). Two scales of attitude toward mathematics. Journal for Research in Mathematics Education, 5, 67-71.

Aiken, L.R., Jr. (1976). Update on attitudes and other affective variables in learning mathematics. Review of Educational Research, 46, 293-311.

Aiken, L.R. & Dreger, R.M. (1961).The effect of attitudes on performance in learning mathematics. Journal of Educational Psychology, 52, 19-24.

Aiken, L.R., Jr. & Dreger, R.M. (1963). Personality correlates of attitude toward mathematics. Journal of Educational Research, 56, 576-580.

Allport, G.W. (1935). Attitudes. In M. Fishbein (Ed.)(1967), Readings in attitude theory and measurement (pp. 1-13). New York: John Wiley & Sons, Inc.

American Psychological Association (1966). Standards for educational and psychological tests and manuals. Washington, D.C.: author.

Anastasi, A. (1976). Psychological testing (4th ed.). New York: The Macmillan Company.

Anastasi, A. (1986). Evolving concepts of test validation. Annual Reviews of Psychology, 37, 1-15.

Anderson, L.W. (1981). Assessing affective characteristics in the schools. Boston: Allyn and Bacon, Inc.

Bowling, J.M. (1977). Three scales of attitude toward mathematics. Dissertation Abstracts International. 37: 4927A-4928A.

Brophy, J. E. (1979). Teacher behavior and its effects. Journal of Educational Psychology, 71, 733-750.

Brophy, J. (1986). Teaching and learning mathematics: where research should be going. Journal for Research in Mathematics Education, 17, 323-346.

Brophy, J.E. & Good, T.L. (1970). Teacher's communication of differential expectations for children's classroom performance: Some behavioral data. Journal of Educational Psychology, 61, 365-374.

Brown, C.A., Carpenter, T.P., Kouba, V.L., Lindquist, M.M., Silver, E.A., & Swafford, J.O. (1988). Secondary school results for the fourth NAEP mathematics assessment: Algebra, geometry, mathematical methods, and attitudes. Mathematics Teacher, 81, 337-347.

Buchanan, N.K. (1987). Factors contributing to mathematical problem-solving performance: and exploratory study. Educational Studies in Mathematics, 18, 399-415.

Campbell, D.T. (1950) The indirect assessment of social attitudes. In M. Fishbein (Ed.)(1967), Readings in Attitude Theory and Measurement (pp.163-179). New York: John Wiley & Sons, Inc.

Carpenter, T.P., Cobitt, M.K., Kepner, H.S.,Jr., Lindquist, M.M., & Reyes, R.R. (1980). Students' affective responses to mathematics: Secondary school results from national assessment. Mathematics Teacher, 73, 531-539.

Chein, I. (1950). Behavior theory and the behavior of attitudes: Some critical comments. In M. Fishbein (Ed.), Readings in Attitude theory and measurement (pp. 51-57). New York: John Wiley & Sons, Inc.

Childress, D.R. (1976). A study of the relationships between students' attitudes toward mathematics and their ratings of mathematics courses and mathematics instructors. Dissertaion Abstracts International. 36: 4340A.

Confrey, J. (1986). A critique of teacher effectiveness research in mathematics education. Journal for Research in Mathematics Education, 17, 347-360.

Cook, S.W. & Sellitz, C. (1964). A multiple indicator approach to attitude measurement. Psychological Bulletin, 62, 36-55.

Crano, W.D. & Brewer, M.B. (1973). Principles of research in social psychology. New York: McGraw-Hill Book Company.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Cronbach, L.J. (1959). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Cronbach, L.J. (1960). Essentials of psychological testing. New York: Harper & Row, Publishers, Incorporated.

Cureton, E.E. (1951). Validity. In E.F. Lindquist (Ed.), Educational measurement (pp. 621-694). Washington, D.C.: American Council on Education.

Cureton, E.E. (1965). Reliability and validity: Basic assumptions and experimental designs. Educational and Psychological Measurement, 25, 327-346.

Dick, W. & Hagerty, N. (1971). Topics in measurement. New York: McGraw-Hill Book Company.

Dutton, W.H. (1954). Measuring attitudes toward arithmetic. Elementary School Journal, 55, 24-31.

Dutton, W.H. (1962). Attitudes of prospective elementary school teachers toward arithmetic. Arithmetic Teacher, 9, 418-424.

Dutton, W.J. & Blum, M.P. (1968). The measurement of attitudes toard arithmetic with a Likert-type test. Elementary School Journal, 68, 259-264.

Eccles, J.S. (1983). Expectancies, values, and academic behaviors. In J.T. Spence (Ed.), Achievement and Achievement Motivation (pp. 75-146). San Francisco: W.H. Freeman.

Eccles, J.S., Midgley, C., & Adler, T.F. (1984). Grade related changes in the school environment: Effects on achievement motivation. In J.G. Nicholls (Ed.), The Development of Achievement Motivation (pp. 283-331). Greenwich, CT: JAI Press.

Edwards, A.L. (1957). Techniques of attitude scale construction. New York: Appleton-Century-Crofts, Inc.

Edwards, A.L. (1972). Experimental design in psychological research (4th ed.). New York: Holt, Rinehart and Winston, Inc.

Ellingson, J.B. (1962) Evaluation of attitudes of high school students toward mathematics. Dissertation Abstracts. 23: 1604.

Feldlaufer, H., Midgley, C., & Eccles, J.S. (1988). Student, teacher, and observer perceptions of the classroom environment before and after the transition to junior high school. Journal of Early Adolescence, 8, 133-156.

Fennema, E., & Sherman, J.A. (1976). Fennema-Sherman mathematics attitudes scales: Instruments designed to measure attitudes toward the learning of mathematics by males and females, Catalog of Selected Documents in Psychology, 6, 31.

Ferguson, G.A. (1981). Statistical analysis in psychology and education (5th ed.). New York: McGraw-Hill Book Company.

Fishbein, M. & Ajzen, I. (1975). Belief, attitude, intention and behavior: An introduction to theory and research. Reading, Massachusetts: Addison-Wesley Pub. Co.

Foster, G., Algozzine, B., Ysseldyke, J. (1980). Classroom teacher and teacher-in-training susceptibility to stereotypical bias. The Personnel and Guidance Journal, 59, 27-30.

Fox, D.J. (1969). The research process in education. New York: Holt, Rinehart and Winston, Inc.

Frary, R. & Ling J. (1983). A factor analytic study of mathematics anxiety. Educational and Psychological Measurement, 43, 985-993.

Garrett, H.E. (1937). Statistics in psychology and education. New York: Longmans, Green.

Garrett, H.E. (1947). Statistics in psychology and education. New York: Longmans, Green.

Gladstone, R., Deal, R., & Drevdahl, J.E. (1960). An exploratory study of remedial math. In M.E. Shaw & J.M. Wright (Eds.)(1967), Scales for the measurement of attitudes. New York: McGraw-Hill Book Company.

Good, T.L. (1970). Which pupils do teachers call on? Elementary School Journal, 70, 190-198.

Good, T.L. (1981). Teacher expectations and student perceptions: A decade of research. Educational Leadership, 38, 415-422.

Good, T.L. & Brophy, J.E. (1987). Looking in classrooms. New York: Harper & Row, Publishers.

Goodenough, W.H. (1944). A technique of scale analysis. Educational and Psychological Measurement, 4, 179-190.

Gorsuch, R.L. (1974). Factor analysis. Philadelphia: W.B. Saunders Company.

Gough, H.G. The Adjective Check List. Palo Alto, CA: Consulting Psychologists Press.

Green, W. (1991). Maintaining vitality. Mathematics Teacher, 81, 32.

Greenwald, A.G. On defining attitudes and attitude theory. In A.G. Greenwald (Ed.), Psychological foundations of attitudes (pp 361-390). New York: Academic Press.

Guilford, J.P. (1946). New standards for test evaluation. Educational & Psychological Measurement, 6, 427-438.

Guilford, J.P. (1954). Psychometric Methods (2nd ed.). New York: McGraw-Hill Book Company, Inc.

Guilford, J.P. (1959). Personality. New York: McGraw-Hill Book Company, Inc.

Guilford, J.P. (1965). Fundamental statistics in psychology and education (4th ed.). New York: McGraw-Hill Book Company.

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

Guttman, L. (1944). A basis for scaling qualitative data. Sociological Review, 9, 139-150.

Guttman, L. (1947). The cornell technique for scale and intensity analysis. Educational and Psychologycal Measurement, 7, 247-280.

Guttman, L. & Suchman, E.A. (1947). Intensity and a zero point for attitude analysis. In M. Fishbein (Ed.) (1967), Readings in attitude theory and measurement (pp.267-276). New York: John Wiley & Sons, Inc.

Haladyna, T., Shaughnessy, J. & Shaughnessy, J.M. (1983). A causal analysis of attitude toward mathematics. Journal for Research in Mathematics Education, 14, 19-28.

Harman, H.H. (1967). Modern factor analysis. Chicago: University of Chicago Press.

Hart, L.E. (1989). Classroom processes, sex of student, and confidence in learning mathematics. Journal for Research in Mathematics Education, 20, 242-260.

Hassan, A.M. & Shrigley, R.L. (1984). Designing a likert scale to measure chemistry attitudes. School Science and Mathematics, 84, 659-669.

Heise, D.R. (1970). The semantic differential and attitude research. In G.F. Summers (Ed.), Attitude measurement (pp. 235-253). Chicago: Rand McNally & Company.

Holland, P.W. (1985). On the study of differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), Test Validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Hotelling, H. (1935). The most predictable criterion. Journal of Educational Psychology, 26, 139-142.

Hume, D. (1979). A scale to measure attitude toward one's mathematics teacher. (Doctoral dissertation, University of Tennessee, 1979). Dissertation Abstracts International, 40, 3171A.

Kaiser, H.F. (1958). The varimax crierion for analytic rotation in factor analysis. Psychometrika, 23, 187-200.

Kirsch, I.S., and Jungeblut, A. (1986). Literacy profiles of america's young adults. Princeton, N.J.: Educational Testing Service.

Krech, D., Crutchfield, R.S., & Ballachey, E.L. (1962). Individual in society. New York: McGraw-Hill Book Company.

Kuder, G.F. & Richardson, M.W. (1937). The theory and estimation of test reliability. Psychometrika, 2, 151-160.

Kulm, G. (1980). Research on mathematics attitude. Journal for Research in Mathematics Education, 11, 356-381.

Lemon, N. (1973). Attitudes and their measurement. New York: John Wiley & Sons.

Likert, R. (1932). A technique for the measurement of attitudes. In G.F. Summers (Ed.)(1970). Attitude measurement, (pp. 149-158). Chicago, IL: Rand McNally & Company.

Lindquist, E.F. (1942). A first course in statistics. New York: Houghton Mifflin.

Lindquist, E.F. (Ed.). (1951). Educational Measurement. Washington, D.C.: American Council on Education.

Ling, J.L. (1983). A factor-analytic study of mathematics anxiety. Dissertation Abstracts International, 43, 2266A. (University Microfilms No. 82-26,901)

McCallon, E.L. & Brown, J.D. Semantic differential instrument for measuring attitude toward mathematics. Journal for Experimental Education, 39, 69-72.

McKnight, C.C., Crosswhite, F.J., Dossey, J.A., Kifer, J.O. Swafford, J.O., Travers, K.J., and Cooney, T.J. (1987) The underachieving curriculum: Assessing U.S. school mathematics from an international perspective Champaign,IL: Stipes Publishing Co.

McNeil, J. (1971) Toward accountable teachers: Their appraisal and improvement. New York: Holt, Rinehart and Winston.

Meek, A. (1989). On creating ganas: a conversation with Jaime Escalante. Educational Leadership, 46, 46-47.

Mehrens, W.A. & Ebel, R.L. (Eds.) (1967). Principles of educational and psychological measurement. Chicago: Rand McNally & Company.

Mehrens, W.A. & Lehmann, I.J. (1980). Standardized tests in education (3rd ed). New York: Holt, Rinehart and Winston.

Midgley, C., Feldlaufer, H., & Eccles J.S. (1989). Change in teacher efficacy and student self-and task-related beliefs in ma+hematics during the transition to junior high school. Journal of Educational Psychology, 81, 247-258.

Mueller, D.J. (1970). Physiological techniques of attitude measurement. In G.F. Summers (Ed.)(1970), Attitude measurement (pp. 534-552). Chicago, IL: Rand McNally & Company.

Mulaik, S.A. (1972). The foundations of factor analysis. New York: McGraw-Hill Book Company.

National Commission on Excellence in Education (1983). A nation at risk. Washing+on, DC: Superintendent of Documents.

National Council of Teachers of Mathematics. (1991). Professional Standards for Teaching Mathematics. Reston, VA: author.

National Council of Teachers of Mathematics. (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: author.

Norusis, M.J. (Ed.) (1988). SPSS/PC+ V2.0 base manual for the IBM PC/XT/AT and PS/2. Chicago: SPSS Inc.

Nunnally, J.C. (1967). McGraw-Hill series in psychology: psychometric theory. New York: McGraw-Hill Book Company.

Nunnally, Jr., J.C. (1970). Introduction to psychological measurement. New York: McGraw-Hill Book Company.

Oppenheim, A.N. (1966). Questionnaire design and attitude measurement. New York: Basic Books, Inc.

Osgood, C.E. (1952). The nature and measurement of meaning. In J.G. Snider & C.E. Osgood (Eds.) (1969), Semantic Differential Technique (pp. 3-41). Chicago: Aldine Publishing Company.

Osgood, C.E. & Suci, G.J. (1955). In J.G. Snider & C.E. Osgood (Eds.) (1969), Semantic Differential Technique (pp. 42-55). Chicago: Aldine Publishing Company.

Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. (1957). The measurement of meaning. Urbana, IL: University of Illinois Press.

Pambookian, H. (1976). Discrepancy between instructor and student evaluation of instruction: Effect on instruction. Instructional Science, 5, 63-75.

Pederson, K., Bleyer, D.R., & Elmore, P.B. (1985). Attitudes and career interests in junior high school mathematics students: Implications for the classroom. Arithmetic Teacher, 33, 45-48.

Phillips, R.B. (1973). Teacher attitude as related to student attitude and achieverment in elementary school mathematics. School Science and Mathematics, 73, 501-507.

Porier, G.W. & Lott, A.J. (1967). Galvanic skin responses and prejudice. In G.F. Summers (Ed.)(1970), Attitude measurement (pp. 489-496). Chicago: Rand McNally & Company.

Purkey, W.W., Cage, B, & Graves, M. The Florida Key: A scale to infer learner self-concept. Educational and Psychological Measurement, 33, 979-984.

Quilter, D., & Harper, E. (1988). Why we didn't like mathematics, and why we can't do it. Educational Research, 30, 121-134.

Richardson, F.C. & Suinn, R.M. (1972) The mathematics anxiety rating scale: Psychometric data. Journal of Counseling Psychology, 19, 551-554.

Rosenthal, R. & Jacobson, L. (1968). Pygmalion in the classroom: teacher expectation and pupils' intellectual development. New York: Holt, Rinehart & Winston.

Rosenthal, R., & Rubin, D. (1971). Appendix C: Pygmalion reaffirmed. In J. Elashoff & R. Snow (Eds.), Pygmalion reconsidered. Belmont, CA: Wadsworth Publishing.

Rosenthal, R. (1973). The pygmalion effect lives. Psychology Today, 7, 56-63.

Rosenthal, R. (1974). On the social psychology of the self-fulfilling prophecy: Further evidence for pygmalion effects and their mediating mechanisms. New York: MSS Modular Publications.

Schoenfeld A.H. (1989). Explorations of students' mathematical beliefs and behaviors. Journal for Research in Mathematics Education, 20, 338-355.

Schunk, D. (1985). Self-efficacy and classroom learning. Psychology in the Schools, 22, 208-223.

Selitiz, C., Jahoda, M., Deutsch, M. & Cook, S.W. (1959). Attitude scaling. In M. Jahoda & N. Warren (Eds.) (1966), Attitudes (pp. 305-324). Baltimore: Penguin Books Inc.

Snider, J.G. & Osgood, C.E. (Eds.)(1969). Semantic differential technique. Chicago: Aldine Publishing Company.

Spielberger, C.D. (1978). Test anxiety inventory. University of South Florida.

Steeg, J.L., (1983). Behavioral and attitudinal changes of teachers toward low achieving students as a result of the TESA program. (Doctoral dissertation, United States International University, 1982). Dissertation Abstracts International, 43, 425A.

Steen, L.A., (1989). Teaching mathematics for tomorrow's world. Educational Leadership, 47, 18-22.

Stevenson, H.W. (1987). America's math problems. Educational Leadership, 45, 4-10.

Suinn, R.M. (1972). Mathematics anxiety rating scale. Fort Collins, CO: Rocky Mountain Behavioral Science Institute, Inc.

Suinn, R. M., Edie, C.A., Nicoletti, J. & Spinelli, P.R. (1972). The MARS, a measure of mathematics anxiety: Psychometric data. Journal of Clinical Psychology, 28, 373-375.

Thissen, D.,Wainer, H., & Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Thorndike, R.L. (1966). Reliability. In E.F. Lindquist (Ed.), Educational measurement (pp. 560-619). Menasha, WI: George Banta Publishing Company.

Thorndike, R. (1968). Review of pygmalion in the classroom. American Educational Research Journal, 5, 708-711.

Thurstone, L.L. & Chave E.J. (1928). Attitudes can be measured. American Journal of Sociology, 33, 529-554.

Thurstone, L.L. & Chave E.J. (1929). The measurement of attitude. Chicago: The University of Chicago Press.

Triandis, H.D. (1971). Attitude and attitude change. New York: John Wiley & Sons, Inc.

Troldahl, V.C., & Powell, F.A. (1965). A short-form dogamatism scale for use in field studies. Social Forces. 44, 221-224.

Wainer, H. & Braun H.I. (Eds.) (1988). Test validity. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Weiner, B. (1971). Achievement motivation and attribution theory. Morristown, NJ: General Learning Press.

Westie, F.R. & Defleur, M.L. (1959). Autonomic responses and their relationship to race attitudes. In G.F. Summers (Ed.) (1970), Attitude measurement (pp. 497-506). Chicago: Rand McNally & Company.

Woodmansee, J.J. (1970). The pupil response as a measure of social attitudes. In G.F. Summers (Ed.), Attitude measurement (pp. 514-533). Chicago: Rand McNally & Company.