

QUANTIFYING THE UNQUANTIFIABLE: DEVELOPING QUALITY MEASURES FOR AFFECTIVE OBJECTIVES

Rehana Masrur

Department of Secondary Teacher Education
Allama Iqbal Open University, Pakistan

Masrur Alam Khan

Department of Communication
International Islamic University, Malaysia

Artikel ini bertujuan untuk membimbing pengkaji membangunkan kemahiran asas dalam mereka dan menggunakan instrumen untuk mengukur objektif afektif (sikap). Isi kandungan artikel ini berasaskan pengalaman beberapa pakar penilaian dan pengukuran dalam bidang pendidikan dan sains sosial. Mengukur sikap sesebuah sampel bukan sesuatu tugas yang mudah. Konsep sikap, seperti konsep abstrak yang lain, adalah suatu konstruk. Penulis membuat kesimpulan tentang langkah yang perlu diambil untuk membina skala sikap dengan cara yang mudah. Skala Likert sering digunakan untuk mengukur sikap kerana skala ini ringkas dan uni-dimensional. Artikel ini membimbing pengkaji yang ingin membangunkan instrumen yang boleh dipercayai untuk kajian berkualiti dalam bidang penyelidikan pendidikan.

Every component of research is logically embedded in a systematic process. The separation of one component from another is not only impossible but is also very illogical. Every section is a base for the next section. Consequently, it is very pertinent to say that research results can be no better than the measures used to obtain them. The findings and conclusions of a research study will be of little value if drawn from the data collected by unsound, unreliable and invalid research tools.

The importance of measurement in a research activity cannot be ignored. Measurement in the physical and biological sciences has become very accurate and sophisticated compared to measurement in social sciences and education. Wright and Stone (2004) state: "Data analysis requires the skillful interplay of theory and observation" (p. 117). The interplay of theory and observation can only be achieved through measurement. In fact, the basic assumption of data analysis requires valid and reliable observations collected through a well-formulated instrument, keeping in view the psychometric principles. Research in education and the social sciences is concerned with the most complex of all scientific subjects. Borg, Gall, and Gall (1993) lamented:

Human beings and some characteristics that are relevant to the educational process can be better measured at present than others. For example, measures of strength, manual dexterity, and visual acuity are quite precise and reliable. Measures of academic achievement are moderately accurate. However, when we attempt to measure such complex human characteristics as personality traits or creativity, our measures are subject to large errors. (p. 110)

The authors' 20 years experience in supervision of research theses at Master's, M.Phil. and Ph.D. levels and survey of more than 50 research studies conducted in the field of education and social sciences revealed that a majority of the researchers used unreliable and invalid tests and scales for data collection. Most of the studies measured attitudes of individuals towards defined stimuli through different scales. We know that attitudes involve an individual's feeling towards such things as ideas, procedures and social institutions (Wiersma, 2000). The individual's response is not only confined to the feeling of some stimuli, it further shows some degree of intensity of covert feeling. However, as Wiersma argues, "the intensity of a person's feeling usually is not dichotomous but is on a continuum between the extremes. Measurement of the attitude places an individual on this continuum" (p. 305) and varies from person to person. It even carries some degree of variation in the same person in different situations. However, behavioral scientists are skeptical because "they recognize that behavior is complex and often many factors interact to cause a psychological phenomenon. Discovering these factors is often a difficult task" (Shaughnessy, Zechmeister, & Zechmeister, 2003, p. 27).

It implies that an attitude is a learned predisposition to respond positively or negatively to a specific object, situation, institution or person. It consists of knowledge, intellectual, emotional, motivational, and behavioral (action) components. Opinion, interest, belief, and value are interchangeably used for attitude, although there are differences in the usage of these terms. Rating scales are sparingly used for measuring these constructs.

Rating scales measure an individual's attitude, personality characteristics, emotional state, interest, values, and other related factors. Scaling is defined by Cohen and Swerdlik (2002) as "the process by which a measuring device is designed and calibrated, and the way numbers or other indices and scale values are designed to different amount of traits, attributes, or characteristics being measured" (p. 193).

However, researchers apply measurement procedures to maintain objectivity and use statistical models to draw conclusions. But several questions emerge: Are the findings and conclusions reliable? What are the theoretical and practical implications of findings which are based merely on unsound research tools? What do the research students and their supervisors need to know about research tools? This article assumes, first of all, that research students need to be guided adequately in developing the research tools and second, that the supervisor's orientation in developing tools may be enhanced. Since this is the chain of knowledge transmitted from one person to another in a way as they have learned it, an attempt has been made to analyze the existing techniques in the preparation of the widely used Likert scale in measuring attitude in educational and social research.

The Likert Scale

An overview of the historical background of measurement tools shows that the Likert scale is the most commonly used scale for measuring attitudes. It was developed in 1930 by Rensis Likert, whose 1932 monograph was reprinted with a few changes in a volume edited by Murphy and Likert (1937), which contains a detailed report of applications of scales constructed by the Likert technique. Likert developed a simpler method of attitude scale construction, keeping in view the drawbacks of the Thurstone scaling method which is tedious and time-consuming (Oskamp, 1991). Most educationists involved in research know that the Likert scale is a scale with points ranging from three to seven, though five-point response categories are more commonly used. It is a uni-dimensional scale method, which has been placed in the category of "summative" scales. Neuman (2006), while discussing the commonly used scales, notes: "Likert scales are called summative-rating or additive scales because a person's score on the scale is computed by summing the number of responses the person gives" (p. 207). The development of scientific measurement enabled the quantification of even unquantifiable affective behavior. The Likert method was indeed a breakthrough which allowed the development of objective measures in social psychology (Howitt & Cramer, 2005).

Though there are a wide variety of tools for assessing behavior, "many researchers prefer to use a Likert-type scale because it is very easy to analyze statistically" (Jackson, 2006, p. 78).

The popularity of the Likert rating scale is based on the following assumptions:

- 1 Statements (items, questions etc.) can easily be constructed.
- 2 Response categories can easily be identified.
- 3 Item-wise analysis can easily be performed.
- 4 It is a uni-dimensional scale.

These characteristics lead the researchers to develop an instrument which they perceive is easier for them to use. Researchers take this simplicity as granted and frequently use the name Likert scale, but the instruments they often use in their studies are the simple fact-finding survey questionnaires instead of Likert scales. The statements are developed in such a way that almost all items elicit a response on one end of the continuum. To make it worse, the response categories are selected without knowing whether these categories are appropriate for the statement or not. The analysis of such responses (individual statements), or response of 50 respondents to a statement on a five-point scale, is displayed in Table 1. In this data, a mean score was obtained after multiplying the frequency of each response category with the respective scale assigned to it and then adding this product divided by the total number of respondents as follows:

$$\text{Mean score} = \frac{\text{Frequency of response category} \times \text{assigned scale}}{\text{Total number of respondents}}$$

Table 1

Distribution of Responses to Statement No. 1

Statement	SA	A	N	DA	SDA	Mean Score
	5	4	3	2	1	
I always feel that my friends do not care about my feelings	6	10	8	15	11	
f x scale of response	30	40	24	30	11	2.9
Responses in %	12	20	16	30	22	

The question that arises is: how can a researcher justify that a mean score on ordinal data is truly representative of the population from which the majority rated their agreement towards positive categories but have been placed in neutral category (2.9 is close to 3.00)?

Let us examine the statistical implications: This analysis violates the statistical rules, measurement rules, scaling rules on the following facts:

- (a) It is ordinal data, but has been treated as interval data (mean score calculated).
- (b) Most of the responses fall in categories "DA" and "SDA", whereas the mean value is '2.9'.
- (c) This is a negative statement and rules for assigning reverse rating have been violated. (It should be placed on the other end of the continuum of measurement while assigning a low score to the 'strongly agree' category).
- (d) The response categories do not match with the question. It must require the respondent to say: very true, true, undecided, somewhat true, or "not true".
- (e) This instrument consisted of 40 statements. In the analysis section, the researcher produced 40 tables. In another example, where there were three samples, 120 tables were formulated (40 for each sample). Even in some theses, the number of tables exceeded 300. (The researcher applied chi-square to each statement or calculated the mean in the same manner as explained in Table 1). These 40 tables could not present a unified score of the sample for which the research was conducted; rather they presented a confusion of numbers. The main aspect and domains of the variable being measured remained unclear in the analysis.
- (f) The same activity was repeated 40 times with the little change only in responses and statements.
- (g) The accomplishment of study objectives remained vague.
- (h) The criteria (formal and informal) for developing the statements for the instrument were ignored.
- (i) Facts were seriously distorted.

Application of the Likert scale method is not being used in its true spirit. As already mentioned, research findings of any study cannot be more accurate than the measure on which the findings are based (Borg et al., 1993). Borg et al. noted that "if you are interested in applying the findings of educational research to practical problems, you must know something about the principles of educational measurement and how to apply them to evaluate research studies" (p. 110).

Therefore, the main concern of this article is to disseminate the techniques for the construction of Likert scale, self-reporting, and rating scales for measuring attitudes. The method which will be discussed in subsequent sections is based upon direct responses of agreement or disagreement with attitude statements. Since the response method does not require prior knowledge of the scale values of the statements in any exact sense, a judging group is not necessary. It is sufficient for the response method if one can assume that the response "agree" to a statement indicates a more favorable attitude than the response "disagree" or vice versa. In the Likert method construction of attitude scales include the method of summated rating.

How is the Method of Summated Rating Applied?

The foremost step in developing the summated rating scale is defining the construct for which the formulation of attitude statements is required:

Defining the Focus

As in all scaling methods, the first step is to define what it is you are trying to measure. Since this is a uni-dimensional scaling method, it is assumed that the concept you want to measure is one-dimensional in nature. You might operationalize the definition as an instruction to the people who are going to create or generate the initial set of candidate items for your scale. Every test/tool/instrument or measure consists of items to which we are asked to respond in some specific way. These items have been carefully edited and selected according to some criteria. A well-constructed attitude scale consists of a number of items that have been carefully selected and edited according to the same criterion as the items contained in any standardized psychological test. The items which make up an attitude scale are called statements. A statement may be defined as "anything that is said about the psychological object. The class of all possible statements that could be made about a given psychological object is often called a universe of content or simply a universe" (Edwards, 1979, p. 10). A theoretical viewpoint for defining an attitude is given in Figure 1. The figure depicts a tri-componential viewpoint of attitude (adapted from Oskamp, 1991, p. 9).

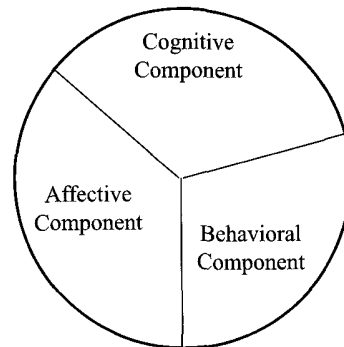


Figure 1. Tri-componential theoretical viewpoint of attitude

The cognitive component consists of the ideas and beliefs which the attitude holder has about the attitude object, whereas an affective (emotional) component consists of feelings and emotions one has towards the object. The third component includes one's action tendencies towards the object. After defining the concept, the researcher formulates the statements (Oskamp, 1991). The measurement of different affective variables demands experience and proper conceptualization of the concepts being measured. Popham (2006) in this regard states "the more experience you accumulate in creating Likert inventories, the easier it gets. After a short while you will really become quite skilled in whipping out such affective assessment devices" (p. 233).

Generating the Items

The next step is to create the set of potential scale items. These should be items that can be rated on a 1-to-5 or 1-to-7 Disagree-Agree response scale. Sometimes you can create the items by yourself based on your intimate understanding of the subject matter. But more often than not, it is helpful to engage a number of people in the item creation step. For instance, you might use some form of brainstorming to create the items. It is desirable to have as large a set of potential items as possible at this stage; about 80-100 would be best.

The pioneer theorists Thurstone and Chave (1929), Wang (1932), Likert (1932), and Edwards and Kilpatrick (1948) have suggested the following informal criteria for the constructing and editing of attitude statements (p.14):

1. Avoid statements that refer to the past rather than to the present.
2. Avoid statements that are factual or capable of being interpreted as factual.
3. Avoid statements that may be interpreted in more than one way.
4. Avoid statements that are irrelevant to the psychological object under consideration.
5. Avoid statements that are likely to be endorsed by almost everyone or by almost no one.
6. Select statements that are believed to cover the entire range of the affective scale of interest.

7. Keep the language of the statements simple, clear, and direct.
8. Keep the statements short, rarely exceeding 20 words.
9. In each statement put only one complete idea.
10. Avoid statements containing universals such as *all*, *always*, *none* and *never* which often introduce ambiguity.
11. Words such as, *only*, *just*, *merely*, and others of similar nature should be used with care and moderation in writing statements.
12. Whenever possible statements should be in the form of simple sentences rather than in the form of compound or complex sentences.
13. Avoid the use of words that may be misunderstood by those who are to be given the complete scale.
14. Avoid the use of double negative.

Besides the above-mentioned criteria it is suggested that a researcher may:

1. Avoid items that are considered too private, self-incriminating, or associated with strong social approval/ disapproval.
2. If statements are translated from English to another language or dialect, print both versions for attitudinal clarification.
3. If an item requires knowledge of more than one concept, replace it with two items, each separately testing one concept.
4. Develop an equal number of favorable and unfavorable statements.

Popham (2006) suggested that while developing the tool for measuring affective behavior, it is important to remove any cues that might force a respondent to elicit a socially desirable response.

Choosing the Response Set and Assigning the Numbers

The points or response categories are assigned numerical values ranging from 1-5 or 0-4. The numbers assigned to the response categories are arbitrary. Researchers need to understand that Likert scale measures are at the ordinal level of measurement because responses indicate a ranking only. The total of these values over the statements gives an individual's attitude score (Wiersma, 2000). The important characteristics stated by many scholars are the appropriateness of response categories for the statement (item). Some of these are given in Table 2.

Table 2
Types of Response Categories Used in Rating Scales and Likert Scales

Very good	Very satisfactory	Strongly Agree	Very supportive
Good	Satisfactory	Agree	Supportive
No opinion	Undecided	Neutral	Neutral
Poor	Unsatisfactory	Disagree	Unsupportive
Very poor	Very Unsatisfactory	Strongly disagree	Very Unsupportive
Highly appropriate	Highly favorable	Always	Strongly approve
Appropriate	Favorable	Often	Approve
Neutral	No opinion	Sometimes	Undecided
Inappropriate	Unfavorable	Rare	Disapprove
Highly Inappropriate	Highly Unfavorable	Not at all	Strongly disapprove

Since there are any number of possible sets of Likert responses, the table gives only a few as examples. A number of artificial dichotomies can be produced, keeping in mind the nature statements (items), and also mindful of the purpose for which a respondent is required to elicit a real response towards a stimulus. One of the most important characteristics of the Likert scale is the suitability of response choices; the responses must be appropriate for the items. Sometimes there is a tendency to try to force all items into a common set of responses such as 'strongly agree' to 'strongly disagree' (Wiersma, 2000).

Selecting the Items

The next step is to compute the inter-correlations between all pairs of items, based on the ratings of the judges. In making judgments about which items to retain for the final scale, there are several analyses that one can do: exclude any items that have a low correlation with the total (summed) score across all items. In most statistics packages, it is relatively easy to compute this type of Item- Total correlation. First, you create a new variable which is the sum of all the individual items for each respondent. Then, you include this variable in the correlation matrix computation (if you include it as the last variable in the list, the resulting Item-Total correlations will all be the last line of the correlation matrix and will be easy to spot). How low should the correlation be for you to throw out the item? There is no fixed rule here -- you might eliminate all items with a correlation with the total score of less than .6, for example.

For each item, get the average rating for the top quarter of judges and the bottom quarter.

Calculation of t- Values

The next step is to do a *t-test* of the differences between the mean value for the item for the top and bottom quarter judges. Add the ratings of each statement as marked by the subjects across all statements. Then take 25% of the subjects with the highest total scores

and 25% of the subjects with the lowest total scores. We assume that these two groups provide criterion groups in terms of which to evaluate the individual statement. Aiken (2003) suggested that 27% of the subjects could also significantly differentiate between two extreme groups. Now arrange the responses as displayed in Table 3.

Table 3
Calculation of t- Values for Evaluating Differences in the Mean Response to an Attitude Statement by a High group and a Low Group (Positive Statement).

Response Categories	Low Group				High Group			
	X	f	fX	fX ²	X	f	fX	fX ²
Strongly Agree	4	2	8	32	4	12	48	192
Agree	3	4	12	36	3	6	18	54
Uncertain	2	10	20	40	2	4	8	16
Disagree	1	5	5	5	1	2	2	2
Strongly Disagree	0	4	0	0	0	1	0	0
Sums		25	45	113		25	76	264
		n _L	∑fX _L	∑fX _L ²		n _H	∑fX _H	∑fX _H ²

Calculate the following:

$$\text{Mean}_L (\text{Mean of low group}) = 1.8 \quad \text{Mean}_H (\text{Mean of high group}) = 3.04$$

$$s_L^2 = \frac{\sum fX_L^2 - (\sum fX_L)^2}{n_L} = 32 \quad s_H^2 = \frac{\sum fX_H^2 - (\sum fX_H)^2}{n_H} = 32.96$$

$$t = \frac{\text{Mean}_H - \text{Mean}_L}{\sqrt{s_H^2 + s_L^2} / n(n-1)}$$

$$t = \frac{3.04 - 1.8}{\sqrt{32 + 32.96} / 25(24)} = 1.24 / .329$$

$$t = 3.76$$

Higher t-values mean that there is a greater difference between the highest and lowest judges. In more practical terms, statements with higher t-values are better discriminators, so retain these statements in the final instrument. In the end, you will have to use your judgment about which items are most sensibly retained. You want a relatively small number of statements on your final scale (e.g., 10-25), and you want them to have high Item-Total correlations and high discrimination (e.g., high t-values). Reverse the rating for negative (unfavorable) statements.

Selection of Statements

Old and current literature (Ary, Jacobs, Razavieh, & Sorensen, 2006; Baker, 1999; Best & Kahn, 1993; Edwards, 1979; Edwards & Kilpatrick, 1948; Jackson, 2006; Likert, 1932; Neuman, 2006; Oskamp, 1991; Popham, 2006; Salkind, 2006; Shaughnessy, Zechmeister, & Zechmeister, 2003; Tan, 2004; Thurstone & Chave, 1929; Wang, 1932; Wiersma, 2000) have suggested that approximately half of the selected statements should be favorable so that the 'strongly agree' response carries the '5' weight and the strongly disagree response the '1' weight. Likert experimented with different weighting of five categories, but concluded that assigning a weight of '1' (for endorsement of an item at one extreme) through '5' (for endorsement of an item at the other extreme) generally worked best (Cohen & Swerdlik, 2002). The half should consist of unfavorable statements so that the scoring system is reversed. The advantage of having both kinds of statements represented in the final scale is to minimize possible response set subjects that might be generated if only favorable or unfavorable statements were included in the scale. Approximately 2025 statements for a uni-dimensional scale are considered an appropriate number.

Reliability of the Likert Scale

The reliability of the scores on the scale can be obtained by correlating scores on the odd-numbered statements with those on the even-numbered statements. The reliability coefficient typically reported (Edwards, 1979) for scales constructed by the summated ratings is above .85, even when fewer than 20 items make up the scale.

Interpretation of Scores

The sum of the weight of all the items checked by the subject would represent the individual's total score. This weighting system means that a high score (SA to A for favorable items; SD or D for unfavorable items) indicates a positive attitude towards the object (Ary et al., 2006). The highest possible score is $5 \times N$ (the number of items); and the lowest possible score is $1 \times N$.

If a person obtains a score of 25 on a 25-item scale, we would interpret this score as indicating an unfavorable or negative attitude. In order to obtain this score, the subject would have had a 'strongly agree' response to every unfavorable statement in the scale. Similarly, we can interpret a score of 125 as indicating a favorable or positive attitude, since this score can be obtained only if the subject gave a 'strongly agree' response to every favorable statement and 'strongly disagree' to every unfavorable statement.

Discussion

Survey of literature shows that none of the published attitude scales used in most of the research and these were actually constructed by the appropriate procedures. In many cases, a set of declarative statements (interrogative ones), each with five 'agree to disagree' response categories was simply put together in an instrument without any theoretical construct in mind. The researchers only reported that content validity of the statement included in the instrument was judged in the light of study objectives by the experts. It was further observed that none of the research has reported the coefficient of reliability. Instruments are mentioned like the Likert scale, but the Likert procedure was nowhere to be found in the methodology section.

Despite the misconception and misuse, Likert scales and other summated rating scales have several advantages, because of their simplicity and veracity, and relative ease of construction. Likert scales are usually considered reliable, which accounts for their widespread popularity. "The simplicity and ease of use of the Likert scale is its real strength. When several items are combined, more comprehensive and multiple indicator measurement is possible" (Neuman, 2006, p. 210).

Neuman has stated two limitations for Likert scales:

1. A different combination of several scale items can result in the same overall score or result.
2. The response set is a potential danger.

Keeping in view these limitations, researchers have to realize that a sound research project involves doing a good job in each phase of research. The authors fully agree with Neuman's warning that "serious mistakes or sloppiness in anyone phase can do irreparable damage to the results, even if the other phases of the research project were conducted in a flawless manner" (p. 217). Obviously, instrumentation is the nucleus of the research process, and we should definitely agree with the statement of Howitt and Cramer (2005) that "the development of good quantitative technique encourages the development of that field of research since it facilitates further research" (p. 246).

References

- Aiken, L. R. (2003). *Psychological testing and assessment* (11th ed.). Boston: Pearson.
- Airasian, P. W. (2001). *Classroom assessment: Concepts and application*. (4th ed.). Boston: McGraw-Hill.
- Ary, D., Jacobs, L. C., Razavieh, A., & Sorenson, C. (2006). *Introduction to research in education* (7th ed.). Belmont, CA: Wadsworth.
- Baker, T. L. (1999). *Doing social research* (3rd ed.). Boston: McGraw-Hill.
- Best, J. W., & Kahn, J. V. (1993). *Research in education* (7th ed.). Boston: Allyn & Bacon.
- Borg, W. R., Gall, J. R., & Gall, M. D. (1993). *Applying educational research: A practical guide*. (3rd ed.). New York: Longman.
- Cohen, R. J., & Swerdlik, M. E. (2002). *Psychological testing and assessment: An introduction to tests and measurements* (5th ed.). Boston: McGraw-Hill.
- Edwards, A. L. (1979). *Techniques of attitude scale construction*. New York: Appleton.
- Edwards, A. L., & Kilpatrick, F. P. (1948). A technique for the construction of attitude scales. *Journal of Applied Psychology*, 32, 374-384.
- Howitt, D., & Cramer, D. (2005). *Introduction to research methods in psychology*. Harlow, UK: Pearson.
- Jackson, S. L. (2006). *Research methods and statistics: A critical thinking approach* (2nd ed.). Belmont, CA: Wadsworth.
- Leary, M. R. (2004). *Introduction to behavioral research methods* (4th ed.). Boston: Pearson.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archive of Psychology*, 140, 1-55.
- Murphy, G., & Likert, R. (1937). *Public opinion and the individual*. New York: Harper.
- Neuman, W. L. (2006). *Social research methods: Qualitative and quantitative approach* (5th ed.). Boston: Allyn & Bacon.
- Oskamp, S. (1991). *Attitude and opinion* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (3rd ed.). Boston: Allyn & Bacon.
- Popham, W. J. (2006). *Classroom assessment: What teachers need to know* (4th ed.). Boston: Pearson.
- Salkind, N. J. (2006). *Exploring research* (6th ed.). Boston: Pearson.
- Shaughnessy, J. J., Zechmeister, E. B., & Zechmeister, J. S. (2003). *Research methods in psychology* (6th ed.). Boston: McGraw-Hill.
- Tan, W. (2004). *Practical research methods* (2nd ed.). Singapore: Pearson.
- Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude*. Chicago: The University of Chicago Press.

- Wang, C. C. A. (1932). Suggested criteria for writing attitude statements. *Journal of Social Psychology*, 3, 367-373.
- Wiersma, W. (2000). *Research methods in education: An introduction* (7th ed.). Boston: Allyn & Bacon.
- Wright, B. D., & Stone, M. H. (2004). *Making measures*. Chicago: The Phaneron Press.

