

# The measurement process variables scales of measurement



## **INTRODUCTION**

In any kind of research undertaken, the measurement of the variables in the theoretical framework, is an important component of a research design (Sekaran, 2003). The measurement of the variables are important in order to test the hypotheses and obtain answers to complex issues (Sekaran, 2003). This assignment will address several issues related to the measurement of variables, namely, the level or scales of measurement, the criteria for good measurement which addresses the issues on reliability, validity and sensitivity, and finally, the rating scales where emphasis is given to the Likert scale.

### **The Measurement Process**

According to Trochim (2006), the measurement process can be defined as the process of observing and recording the observations while the research is being undertaken. There are two major measurement concepts in relation to the research measurement.

The first, is, the levels of measurement. There are four levels of measurement – nominal, ordinal, interval and ratio. The second, is the different types of measures used in social research. They can be classified into four broad categories of measurements. These are the survey research, scaling, qualitative research and unobtrusive measures (Trochim, 2006). The survey research involves the design and implementation of interviews and questionnaires. Under scaling, it involves the methods of developing and implementing a scale.

The non-numerical measurement approaches comes under the qualitative research where there are a variety of measurement methods that do not interfere or affect the context of the research and come under the unobtrusive research (Trochim, 2006) ([http://www. socialresearchmethods. net/kb/measure. php](http://www.socialresearchmethods.net/kb/measure.php))

A measurement process can also be considered as a process through which the kind or intensity of something is determined (Adam, 1964; Allen & Yen, 1979; Anastasi, 1982). According to Zikmund (2003), the concepts relevant to a problem must be known before the measurement process can be initiated. A concept (or construct) is a generalised idea about a class of objects, attributes, occurrences, or processes, for example, number of children, age, and sex. In order to measure the concept, it needs to be operationalised.

## **VARIABLES**

Shajahan (2005), suggested that a variable can take on different values or characteristics under varying circumstances. For example, a dependant variable is a measure of the behaviour of the subject that reflects the effects of the independent variable. An independent variable is the condition manipulated or selected by the experimenter to determine its effect on behaviour. An independent variable may have different values. According to Shajahan (2005), the subject variable is the difference between subjects that cannot be controlled but can only be selected. The variable that can vary in amount is known as quantitative variable.

Different values that entities can take may be numbers with no quantitative meaning, such as the numbers worn by football players on their jerseys. All the values for these variables have been assigned through a set of rules defining a measurement operation. These measurement operations represent different scales of measurement (Meyers, Gamst, & Guarino, p. 19)

## **SCALES OF MEASUREMENT**

Once we have learnt to operationalise concepts, then there is a need to measure them. Under this title, the different types of scales that can be applied to measure different variables will be examined.

The measurement comprises sets of rules governing the meaning of values assigned to entities. Each set of rules defines a scale of measurement.

Stevens (1951) identified four scales, which include, nominal, ordinal, interval and ratio. Each scale includes an extra feature or rule over those in the one before it (Meyers, Gamst & Guarino, p. 19).

As identified by Stevens (1951) there are four types of scales and each of these scales will be examined below.

### **Nominal Scale**

A nominal scale is the most basic method of measurement. It categorises the research subjects into mutually exclusive groups. For example, under gender, the respondents can be classified into either male or female and under business research, the male can be given a code as 1 and female as 2. These numbers have no intrinsic values except that the classification is simple and convenient. Since there isn't any third category, therefore, these

categories can be considered as collectively exhaustive (Sekaran, 2003).

Apart from being known as the nominal scale, it is also identified as categorical scale or a classification system where its only rule is that, different variables have different values (Meyers, Gamst, Guarino, p. 19). As discussed earlier, if we take the male and female as the respondents, the information that can be generated from nominal scaling is to calculate either the percentage or frequency of males and females in the sample of respondents. For example, when a survey is carried out among 300 participants, the code 1 shall be assigned to all the male participants and code 2 assigned to female participants. When the data analysis is carried out after the completion of the survey, it may show that 178 were males and 122 females. This frequency distribution tells that almost 59 per cent of the respondents were males and 41 per cent women. The nominal scaling can only provide this marginal information about the two groups. Another example of nominal scaling is the nationality of individuals. If we have the following nationalities, for instance, Chinese, Indian, Malays, American and Russian, then each respondent has to fit into any one of these five categories and the scale will allow computation of the numbers and percentage of respondents that fit into them.

## **Ordinal Scale**

According to Sekaran (2003), an ordinal scale, rank-orders the categories and indicates the differences among the various categories. The ordinal scale can be used when the categories are to be ordered, based on some preference, for example, from best to worst, first to last, less than or more than.

The ordinal scale, is an advanced form of categorisation where it uses numeral, letters and symbols to rank objects. They are most commonly used in the ranking of preferences. For example, a consumer needs to list down his preferences for several brands of beer. A typical ordinal scale in business research asks respondents to rate career opportunities, brands and companies as “ excellent,” “ good,” “ fair,” or “ poor,”. Researchers know “ excellent” is higher than “ good,” but they do not know by how much (Zikmund, 2003).

As stated in Shajahan (2005), the ordinal measurements do not provide information on how much more or less of the characteristics various objects possess. There are various kinds of descriptive statistics that can be calculated from these data, namely, mode, median and percentages.

From the above discussion, it can be said that the ordinal scale, provides more information than the nominal scale. Under the ordinal scale, the respondents are distinguished by rank-ordering them, but does not indicate the magnitude of the differences among the ranks. This deficiency is overcome by interval scaling (Sekaran, 2003).

## **Interval Scale**

An interval scale allows us to perform certain arithmetical operations on the data collected from the respondents. Interval scales not only indicate order, but also measure the order or distance in units of equal interval. According to Shajahan (2005), the distance between numerals are very meaningful because by comparing these distances, we are able to identify how far apart the objects are, in respect to the item in question. As discussed earlier, the

nominal scale is only able to differentiate groups qualitatively by classifying them into mutually exclusive and collectively exhaustive sets, while the ordinal scale rank-orders the preferences and the interval scale assists in measuring the distance between two points on a scale. By doing this it will enable us to carry out the means and standard deviations of the feedback received on the variables.

The classic example of an interval scale is the Fahrenheit temperature scale. If the temperature is  $80^{\circ}$ , it cannot be said that it is twice as hot as  $40^{\circ}$ . This is due to the fact that  $0^{\circ}$  does not indicate there is no temperature but only a relative point on the Fahrenheit scale. Due to the lack of an absolute zero, the interval does not allow the conclusion that 60 is ten times as great as six, but only the interval distance is ten times greater (Zikmund, 2003).

According to Shajahan (2005), the Likert scale is an example of an interval measurement which is used in the measurement of attitudes and personality. Virtually the entire range of statistical analysis can be applied to interval scales, for example, the various descriptive measures which include, standard deviation, median, mode, and mean. Bivariate correlation analysis, t-test, analysis of variance tests, and most multivariate techniques can be applied.

## **Ratio Scale**

Apart from having all the properties of the nominal, ordinal and interval scale, the ratio scale has an absolute zero point rather than relative quantities. Absolute zero means that the object measuring zero does not have the property in question. Because of this, it can be said that 10 hours is

twice as long as 5 hours or 4 miles is half the distance of 8 miles (Meyers, Gamst,& Guarino, p. 21). By having the absolute zero point, the ratio scale has overcome the disadvantage of the arbitrary origin point of the interval scale. Height and weight are examples of a ratio scale. Most financial research that deals with dollar values utilise ratio scales.

## **THE CHARACTERISTICS OF A GOOD MEASUREMENT**

According to Shajahan (2005), there is a certain criteria that a measurement tool has to satisfy and the most important are as explained below:

Unidimensionality - means the scale should be able to measure one characteristic at any particular time. For example, a ruler should be able to indicate the length.

Linearity - under linearity a straight-line model is followed by the scale.

Validity - the scale should be able to measure what it is capable of measuring

Reliability - the scale should be able to provide consistent results and this is a characteristic of consistency.

Accuracy - a tool should provide an accurate and precise measure of what we want to measure.

Simplicity - a scale should not be overly complicated. It should be simple as possible.



## **RATING SCALES**

After the discussion on the four different types of scales, it is important to study the methods of scaling. Basically there are two types of attitudinal scales, that is, the rating scale and the ranking scale. According to Sekaran (2003), a rating scale has several response categories and is used to obtain responses with regard to the subject studied. Ranking scales make comparisons between the subjects studied and obtain the preferred choices and ranking among them.

There are several types of rating scales which are commonly used in business research. The different types of rating scales are, dichotomous scale, category scale, Likert scale, numerical scales, semantic differential scale, itemised rating scale, fixed or constant sum rating scale, staple scale, graphic rating scale and consensus scale (Sekaran, 2003).

For the purpose of this paper, only the Likert Scale will be discussed in detail

### **Likert Scale**

Rensis Likert was a sociologist at the University of Michigan who developed the Likert scale. His main aim was to establish a means of measuring psychological attitudes in a “ scientific” way (Uebersax, 2006). The Likert scale is also known as the summated ratings method. This summated ratings method has been widely used in the measurement of attitudes due to its simplicity in administering to respondents.

McIver and Carmines (1981) (cited in Gliem & Gliem, 2003) describe the Likert scale as follows:

“ A set of items, composed of approximately an equal number of favourable and unfavourable statements concerning the attitude object, is given to a group of subjects. They are asked to respond to each statement in terms of their own degree of agreement or disagreement. Typically, they are instructed to select one of five responses – strongly agree, agree, undecided, disagree, or strongly disagree. The specific responses to the items are combined, so that individuals with the most favourable attitudes will have the highest scores while those with the least favourable (or unfavourable) attitudes will have the lowest scores. While not all summated scales are created according to Likert’s specific procedures, all such scales share the basic logic associated with “ Likert scaling” (pp. 22-23). (Gliem & Gliem, 2003).

Under the Likert scale the respondents are required to respond how strongly they agree or disagree with statements that range from the very positive to very negative

(Zikmund, 2003). As suggested by McIver and Carmines (1981) the individuals can generally choose from five alternatives ranging from strongly agree to strongly disagree. The number of alternatives may range from three to nine. For example, in a study on mergers and acquisitions:

Mergers and acquisitions will lead to a faster growth than internal expansion and lead to higher return for the organization.

Strongly Disagree Disagree Uncertain Agree Strongly Agree

(2) (3) (4) (5)

The responses over a number of items tapping a particular concept or variable (as shown in the example above) are then summated for every respondent. This is an interval scale and differences in responses between any two points on the scale remains the same.

## **Developing a Likert-Type Scale**

As suggested by Shajahan (2005), the following steps are involved in developing the Likert-scale:

The investigator assembles a large number of items considered relevant to the attitude being investigated either favourably or unfavourably.

Through the administration of a questionnaire, the items are tested on a group of respondents.

The highest score of 5 will be awarded to the respondent who provides the most favourable attitude and a lowest score of 1 to the unfavourable respondent.

The respondent's total score is obtained through summing up his scores for each statement.

The next step is to arrange these scores and find out those statements that have a high discriminatory power.

## **Advantages of a Likert Type Scale**

There are several advantages in using the Liker-type scale as suggested by Shajahan (2005). The advantages are as follows:

It is relatively easy to construct the Likert-type scale

Since each statement of the participant is included in the instrument, the Likert-scale therefore is considered to be more reliable.

Each statement is given an empirical test for discriminating ability.

The Likert-type scale can easily be used in respondent-centres and stimulus centered studies

Time taken to construct the Likert-type scale is much less.

### **Limitations of a Likert-type Scale**

Shajahan (2005) has provided some limitations to the use of the Likert-type scale. The limitations include:

Under the Likert scale we are only able to study whether the respondents are more or less favourable to a relevant subject-matter. Unfortunately, we are not able to identify how much more or less favourable the respondents are.

The five positions on the scale may not be equally spaced.

The interval or the space between “strongly agree” and “agree” may not be the same to the interval between “agree” and “undecided”

The total score of individual respondents has little clear meaning since a given total score can be secured by a variety of answer patterns.

(Shajahan, 2005)

## **The Controversy Surrounding Likert-Scale: Is It A Ordinal Scale or Interval Scale?**

There has been two different schools of thought with regards to whether Likert-scale is an ordinal scale or an interval scale. Likert-scale which commonly has a wide range of responses to a given question is used to measure attitude (Jamieson, 2004). Some researchers consider Likert-scale to fall within the definition of ordinal scale, that is, there is a rank-order but the interval or space between them are not the same. According to Blaikie (2003), some researchers tend to assume that the intervals between values are the same or equal. As suggested by Cohen et. al (2000), it is wrong to assume that the difference between “strongly agree” and “agree” is the same as between “strongly disagree” and “disagree”. It has become increasingly important to decide whether Likert-scale is an interval scale or an ordinal scale because the inferential and descriptive statistics are different for either of this scale (Cohen, Manion & Morrison, 2000; Clegg, 1998). For ordinal data, it is normal to use median and mode but the use of mean and standard deviation are not suitable for ordinal data (Blaikie, 2003; Clegg, 1998). There are different inferential statistics used for both ordinal data and interval data. For ordinal data, the usual inferential statistics used are those commonly used in non-parametric tests. Parametric tests are used for interval data.

There has been a long standing controversy between ordinal and interval scale where ordinal scale has been treated as interval scale. In a recent publication in Medical Education, two authors had used Likert-scales and interpreted their data using mean and standard deviation and carried out

parametric tests. According to Blaikie (2003), Likert-scale category constitutes interval-level scale. There is neither support nor argument to state that Likert-scale constitutes interval-level scale. According to Kuzon. Jr et. al. (1996), it is inappropriate to use parametric analysis for ordinal data and Knapp (1990) concurs with this argument stating that sample size and distribution are vital factors in determining whether it is appropriate to use parametric statistics.

### **THREE CRITERIA FOR GOOD MEASUREMENT**

Now that we have seen how to operationally define variables and apply different scaling techniques, it is imperative to ensure that the instruments that have been developed to measure a particular concept is indeed accurately measuring the variable, and we are actually measuring the concept that we set out to measure (Sekaran, 2003).

There are three major criteria for evaluating measurements – reliability, validity and sensitivity.

#### **Reliability**

When a measure is free from error and provides consistent results then the measure is considered as reliable. For example, ordinal measures are reliable if they consistently rank orders items in the same manner and reliable interval measures consistently rank order and maintain the same distance between items. As suggested by Sekaran (2003), the reliability of a measure is an indication of the stability and consistency with which the instrument measures the concept and helps assess the “goodness” of a measure.

Two dimensions underlie the concept of reliability. One is repeatability and the other, internal consistency

## **Stability of measures**

According to Sekaran (2003), the ability of a measure to remain the same over time, despite uncontrollable testing conditions or the state of the respondents themselves is indicative of its stability and low vulnerability to changes in the situation. This attests to its “goodness” because the concept is stably measured. The two tests of stability are test-retest reliability and parallel-form reliability.

### **6. 1. 1. 1 Test-Retest Reliability**

This involves the administration of the same scale or measure to the same respondents at two separate times to test for stability. If similar results are obtained where the test was administered under similar conditions each time, then the measure can be considered as stable over time. For example, if a researcher wants to find out the level of job satisfaction and finds that 70 percent of the population is satisfied with their jobs, then if a similar study were to be repeated after several weeks or months and the researcher again finds that 70 percent of the population is satisfied with their jobs, then the measure used has repeatability (Zikmund, 2003)

### **6. 1. 1. 2 Parallel -Form Reliability**

Sekaran (2003) states that when responses on two comparable sets of measures tapping the same construct are highly correlated, then there is a parallel-form reliability. Both forms have similar items and the same

response format. The only changes are the wordings in the order of sequence of the questions.

## **Internal consistency of Measures**

The internal consistency of measures is indicative of the homogeneity of the items in the measure that tap the construct. Consistency can be examined through the inter-item consistency reliability and split-half reliability tests.

### **6. 1. 2. 1 Inter-item Consistency Reliability**

This test is used to find out the consistency in the answers provided by the participants in a research to all the items in the questionnaires. Cronbach's alpha can be applied to test the inter-item consistency reliability.

### **6. 1. 2. 2 Split-Half Reliability**

Split half reliability involves creating two different scores for each group of respondents by dividing the measure into equivalent halves and then correlating the halves.

To test the goodness of an instrument reliability alone is not sufficient. According to Sekaran (2003), one could reliably measure a concept establishing high stability and consistency, but it may not be the concept that one had set out to measure. Validity ensures the ability of a scale to measure the intended concept.

## **Validity.**

In order to test the goodness of the measures, there are a few types of validity tests which includes, content validity, criterion-related validity and construct validity.



## **Content Validity**

According to Zikmund (2003), a scale should accurately measure what it is supposed to measure. The content validity increases as the number of items that represent the concept is large. In other words, content validity is a function of how well the dimensions and elements of a concept have been delineated (Sekaran, 2003). As stated in Sekaran (2003), “ A panel of judges can attest to the content of the instrument”. Kidder and Judd (1986) cite the example where a test designed to measure degrees of speech impairment can be considered as having validity if it is so evaluated by a group of expert judges (i. e, professional speech therapists) p. 206. Face validity means that on the surface it looks as though the items seem to measure what it is supposed to measure. Clear and obvious question like, “ How many children do you have?”, is considered to have face validity but in scientific researches, it is better to have strong evidence.

## **Criterion-Related Validity**

Criterion-related validity is established where the individuals are differentiated based on certain criteria that it should predict. (Sekaran, 2003). A researcher wishing to establish a criterion validity for a new measure of employee turnover, such as a measure utilising co-workers’ ratings of employee turnover, it is important to ensure that the new measure has some similarity with other traditional measures of turnover. Criterion validity can be classified as concurrent validity or predictive validity. Concurrent validity exists when the scale is able to differentiate individuals who are supposed to be different. Predictive validity enables to predict future event by using a new measure. For example, in a job recruitment process,

where the interviewees are given selection tests in order to differentiate the individual interviewees based on their future job performance than those who score low in the selection tests are considered to be poor performers while those who score high are expected to be high performers.

## **Construct Validity**

According to Zikmund (2003), construct validity is established by the degree to which a measure confirms a network of related hypotheses generated from theory based on the concepts. By applying the statistical analysis of the data, construct validity could be achieved. Construct validity establishes that it is capable of demonstrating that the empirical evidence obtained through a measure is consistent with the theoretical logic of the concepts. In other words, according to Zikmund (2003), there is evidence of construct validity where the measure provides results the way it is supposed to in a pattern of inter-correlation with other variables.

To achieve construct validity, a researcher must already have determined the meaning of measure by establishing what basic researchers call the convergent and discriminant validity (Zikmund, 2003). Convergent validity is established when the scores obtained with two different instruments measuring the same concept are highly correlated. Discriminant validity is established when, based on theory, two variables are predicted to be uncorrelated and the scores obtained by measuring them are empirically found to be similar (Sekaran, 2003)

Some of the ways in which the above form of validity can be established are through correlational analysis, factor analysis and the multitrait, multimethod matrix of correlations.

## **CONCLUSION**