Module 5: Measurement and Scaling Techniques

Objectives

Studying this unit, you will be able to:

- Reliability and validity of sound measurement scale

Multi-item scale should be evaluated for accuracy and applicability. This involves an assessment of reliability, validity and generalisability of the scale. Approaches to assessing reliability include **test–retest reliability, alternative-forms reliability and internal consistency reliability.**

Validity can be assessed by examining **content validity, criterion validity and construct validity.**

Before we can examine reliability and validity we need an understanding of measurement accuracy; it is fundamental to scale evaluation.

Measuring parameters such as **height, weight, length, etc**. does not pose any problem to the investigator as standardized measuring devices such as weighing machines and foot-scales are available. But measuring **abstract properties such as opinion, attitude, belief, values, morale, motivation, etc. is not an easy task as these cannot be measured directly.** They can be assessed only through carefully designed logical questions through questions through questionnaire, interview, etc.

5.3 Measurement accuracy

A measurement is a number that reflects some characteristic of an object. A measurement is not the true value of the characteristic of interest but rather an observation of it. A variety of factors can cause measurement error, which results in the measurement or observed score being different from the true score of the characteristic being measured. The true score model provides a framework for understanding the accuracy of measurement. According to this model,

XO=XT+XS+XR

Where XO = the observed score or measurement

XT = the true score of the characteristic

XS = systematic error

XR = random error

The Observed score is the actual score on the exam and True score is the person's actual ability. Error is the difference between observed and true scores.

Note that the total measurement error includes the systematic error, XS, and the random error, XR.

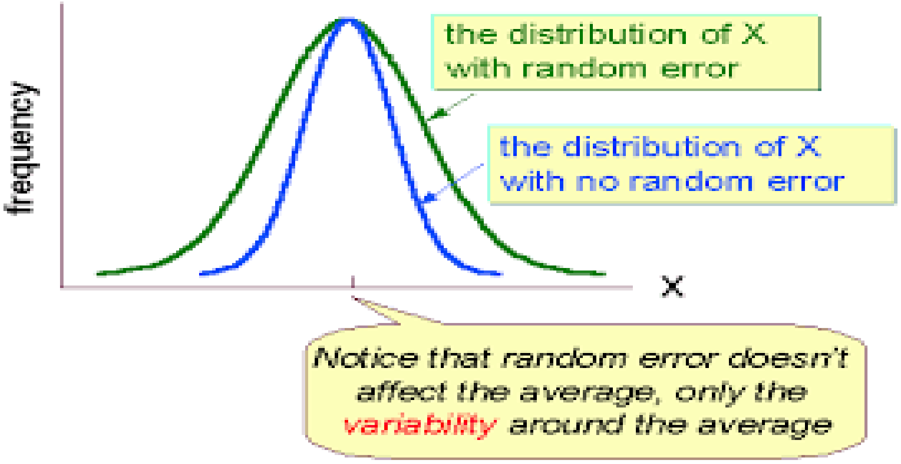
The distinction between systematic and random error is crucial to our understanding of reliability and validity.

Random error, on the other hand, is not constant. It represents transient factors that affect the observed score in different ways each time the measurement is made, such as short-term transient personal factors or situational factors

Random error is caused by any factors that randomly affect measurement of the variable across the sample.

For example, each person's mood can inflate or deflate their performance on any occasion.

In a particular test, some children may be feeling in a good mood and others may be depressed. If mood affects their performance on the measure, it may artificially inflate the Observed Score for some children and artificially deflate them for others. The important thing about random error is that it does not have any consistent effects across the entire sample. Instead, it pushes observed scores up or down randomly. This means that if we could see all of the random errors in a distribution they would have to sum to 0. There would be as many negative errors as positive ones.

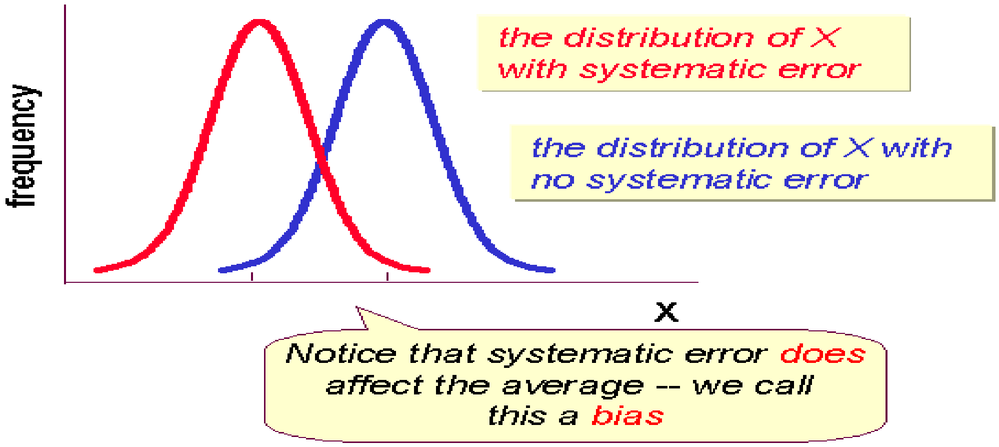


Random errors will affect the reliability but may not affect the overall accuracy of a result.

Systematic error affects the measurement in a constant way. It represents stable factors that affect the observed score in the same way each time the measurement is made, such as mechanical factors.

Systematic error is caused by any factors that systematically affect measurement of the variable across the sample.

For example, if there is a loud traffic going by just outside of a classroom where students are taking a test, this noise is liable to affect all of the children's scores. In this case, the entire group is affected so it affects the accuracy. But do not affect the reliability. Systematic errors tends to be consistently either positive or negative because of this, **Systematic error is sometimes considered to be bias in measurement.**



Potential sources of error in measurement

1. Other relatively stable characteristics of the individual that influence the test score, such as intelligence, social desirability and education
2. Short-term or transient personal factors, such as health, emotions and fatigue
3. Situational factors, such as the presence of other people, noise and distractions
4. Sampling of items included in the scale: addition, deletion or changes in the scale items
5. Lack of clarity of the scale, including the instructions or the items themselves
6. Mechanical factors, such as poor printing, overcrowding items in the questionnaire and poor design

7. Administration of the scale, such as differences among interviewers
8. Analysis factors, such as differences in scoring and statistical analysis

5.3.1Reliability

Reliability refers to the extent to which a scale produces consistent results if repeated measurements are made. Systematic sources of error do not have an adverse impact on reliability, because they affect the measurement in a constant way and do not lead to inconsistency.

In contrast, random error produces inconsistency, leading to lower reliability. Reliability can be defined as the extent to which measures are free from random error, XR. If XR = 0, the measure is perfectly reliable.

Reliability is assessed by determining the proportion of systematic variation in a scale.

This is done by determining the association between scores obtained from different administrations of the scale. If the association is high, the scale yields consistent results and is therefore reliable. Approaches for assessing reliability include the test–retest, alternative-forms and internal consistency methods.

In **test–retest reliability**, participants are administered identical sets of scale items at two different times, under as nearly equivalent conditions as possible. The time interval between tests or administrations is typically two to four weeks. The degree of similarity between the two measurements is determined by computing a correlation coefficient. The higher the correlation coefficient, the greater the reliability. Several problems are associated with the test–retest approach to determining reliability.

First, it is sensitive to the time interval between testing. Other things being equal, the longer the time interval, the lower the reliability. Second, the initial measurement may alter the characteristic being measured. For example, measuring participants' attitude towards low-alcohol beer may cause them to become more health conscious and to develop a more positive attitude towards low-alcohol beer. Third, it may be impossible to make repeated measurements (e.g. the research topic may be the participant's initial reaction to a new product). Fourth, the first measurement may have a carryover effect to the second or subsequent measurements. Participants may attempt to remember answers they gave the first time. Fifth, the characteristic being measured may change between measurements. For example, favourable information about an object between measurements may make a participant's attitude more positive. Finally, the test–retest reliability coefficient can be inflated by the correlation of each item with itself. These correlations tend to be higher than correlations between different scale items across administrations.

Hence, it is possible to have high test–retest correlations because of the high correlations between the same scales items measured at different times, even though the correlations between different scale items are quite low. Because of these problems, a test–retest approach is best applied in conjunction with other approaches, such as alternative-forms reliability.

In **alternative-forms reliability**, two equivalent forms of the scale are constructed. The same participants are measured at two different times, usually two to four weeks apart, with a different scale form being administered each time. The scores from the administrations of the alternative scale forms are correlated to assess reliability. The two forms should be equivalent with respect to content, i.e. each scale item should attempt to measure the same items. There are two major problems with this approach. First, it is time-consuming and expensive to construct an equivalent form of the scale. Second, it is difficult to construct two equivalent forms of a scale. The two forms should be equivalent with respect to content. In a strict sense, it is required that the alternative sets of scale items should have the same means, variances and inter correlations. Even if these conditions are satisfied, the two forms may not be equivalent in content. Thus, a low correlation may reflect either an unreliable scale or non-equivalent forms.

Internal consistency reliability is used to assess the reliability of a summated scale where several items are summed to form a total score. In a scale of this type, each item measures some aspect of the construct measured by the entire scale, and the items should be consistent in what they indicate about the construct. This measure of reliability focuses on the internal consistency of the set of items forming the scale.

The simplest measure of internal consistency is **split-half reliability**. The items on the scale are divided into two halves and the resulting half scores are correlated. High correlations between the halves indicate high internal consistency. The scale items can be split into halves based on odd- and even-numbered items, or randomly. The problem is that the results will depend on how the scale items are split. A popular approach to overcoming this problem is to use the coefficient alpha. **The coefficient alpha, or Cronbach's alpha**, is the average of all possible split-half coefficients resulting from different ways of splitting the scale items. **This coefficient varies from 0 to 1**, and a value of 0.6 or less generally indicates unsatisfactory internal consistency reliability. An important property of coefficient alpha is that its value tends to increase with an increase in the number of scale items.

5.3.2 Validity: The validity of a scale may be considered as the extent to which differences in observed scale scores reflect true differences among objects on the characteristic being measured, rather than systematic or random error. Perfect validity requires that there be no measurement error (XO = XT, XR = 0, XS = 0). Researchers may assess content validity, criterion validity or construct validity.

Content validity, sometimes called **face validity**, is a subjective but systematic evaluation of how well the content of a scale represents the measurement task at hand. The researcher or someone else examines whether the scale items adequately cover the entire domain of the construct being measured.

It is the extent to which a measuring instrument provides adequate coverage of the study. There are again two forms of validity-face validity and sampling validity. Face validity is a logical type depending on the investigator's subjective evaluation. For example, the investigator may prepare an inventory consisting of 15 statements to know individuals' opinion on globalization. The investigator then, evaluates each statement to assess whether the statements can really extract the opinion on globalization and may also get it confirmed from a specialist. It is a poor way of determining validity.

Sampling validity refers to the represented character of the content of the instrument. It is an appropriate sample and sampling technique to represent adequately the content of the population. Many a time, the methods used may not adequately measure the real content.

Given its subjective nature, content validity alone is not a sufficient measure of the validity of a scale, yet it aids in a common-sense interpretation of the scale scores.

Assume that a marketing researcher desires to know which TV serial is most popular to insert his/her company's advertisement and may attach a device to the TV sets of the respondents to record which serials are mostly viewed by the respondents. Based on the record the researcher may prefer a particular popular serial for the company's commercial. But the researcher must remember that there are many people who switch over to other serials/programmes whenever the commercials appear on the screen thus nullifying the results of observation. That is, the serial may be observed to be popular but there is no guarantee that all those who view the serial would pay attention to the advertisement also.

Criterion validity

It relates to the ability to predict some outcome or estimate the existence of some current condition. If the instrument is capable of predicting future performance, it can be said that it has predictive validity; if it is able to relate to other measures of known validity is concurrent validity.

A more formal evaluation can be obtained by examining criterion validity. Criterion validity reflects whether a scale performs as expected in relation to other selected variables (criterion variables) as meaningful criteria. If, for example, a scale is designed to measure loyalty in customers, criterion validity might be determined by comparing the results generated by this scale with results generated by observing the extent of repeat purchasing. Based on the time period involved, **criterion validity can take two forms, concurrent validity and predictive validity**.

Concurrent validity is assessed when the data on the scale being evaluated (e.g. loyalty scale) and the criterion variables (e.g. repeat purchasing) are collected at the same time. The scale being developed and the alternative means of encapsulating the criterion variables would be administered simultaneously and the results compared. **Predictive validity is concerned with how well a scale can forecast a future criterion.** To assess predictive validity, the researcher collects data on the scale at one point in time and data on the criterion variables at a future time. **For example, attitudes towards how loyal customers feel to a particular brand could be used to predict future repeat purchases of that brand**. The predicted and actual purchases (which could be tracked on CRM databases or scanned purchases) are compared to assess the predictive validity of the attitudinal scale.

For example, if there is a significant correlation between the scores in the admission test and the performance (grades obtained) in the first semester then it can be concluded that the admission test has predictive validity. However, the statistical association between predicted outcome and subsequent outcome exhibited need not be a conclusive proof of the instrument’s predictive validity; because the exhibited performance may be influenced by extraneous factors. In the present example, the students who scored low in the admission tests may get better grades in the semester examinations because of their of their hard work or special tuition.

Construct validity It is the degree to which scores on a test can be accounted for by explanatory constructs of a theory. There are procedures to test construct validity.

It addresses the question of what construct or characteristic the scale is, in fact, measuring. When assessing construct validity, the researcher attempts to answer theoretical questions about why the scale works and what deductions can be made concerning the underlying theory. Thus, construct validity requires a sound theory of the nature of the construct being measured and how it relates to other constructs.

Construct validity is the most sophisticated and difficult type of validity to establish. Construct validity includes convergent, discriminant and nomological validity.

Convergent validity is the extent to which the scale correlates positively with other measurements of the same construct. It is not necessary that all these measurements be obtained by using conventional scaling techniques. **Discriminant validity** is the extent to which a measure does not correlate with other constructs from which it is supposed to differ. It involves demonstrating a lack of correlation among differing constructs.

Nomological validity is the extent to which the scale correlates in theoretically predicted ways with measures of different but related constructs. A theoretical model is formulated that leads to further deductions, tests and inferences. Gradually, a nomological net is built in which several constructs are systematically interrelated. Eg, For **example**, a comparison of human aging with memory loss.

For example, a researcher seeks to provide evidence of construct validity in a multi-item scale, designed to measure the concept of 'self-image'. These findings would be sought:

- High correlations with other scales designed to measure self-concepts and with reported classifications by friends (convergent validity).

- Low correlations with unrelated constructs of brand loyalty and variety seeking (discriminant validity).

- Brands that are congruent with the individual's self-concept are preferred, as postulated by the theory (nomological validity).

- A high level of reliability.

Note that a high level of reliability was included as evidence of construct validity in this example. This illustrates the relationship between reliability and validity.

Factor analysis

It is the most powerful method of construct validation. This is a statistical technique designed to determine the basic components of a measure. It is explained in the next chapter data analysis.

It is to be borne in mind that all the three types of validity must be taken into account in selecting an instrument

Summary:

The relationship between reliability and validity can be understood in terms of the true score model. If a measure is perfectly valid, it is also perfectly reliable. In this case, XO = XT, XR = 0 and XS = 0. Thus, perfect validity implies perfect reliability. If a measure is unreliable, it cannot be perfectly valid, since at a minimum XO = XT + XR. Furthermore, systematic error may also be present, i.e. XS ≠ 0. Thus, unreliability implies invalidity. If a measure is perfectly reliable, it may or may not be perfectly valid, because systematic error may still be present (XO = XT + XS). In other words, a reliable scale can be constructed to measure 'customer loyalty' but it may not necessarily be a valid measurement of 'customer loyalty'. Conversely, a valid measurement of 'customer loyalty' has to be reliable. Reliability is a necessary, but not sufficient, condition for validity. Generalisability refers to the extent to which one can generalise from the observations at hand to a universe of generalisations. The set of all conditions of measurement over which the investigator wishes to generalise is the universe of generalisation. These conditions may include items, interviewers and situations of observation.

A good rating scale should have the following characteristics:

Minimal response bias, participant interpretation and understanding, discriminating power, ease of administration, ease of use by participants, credibility and usefulness of results. As a general rule, using the scaling technique that will yield the highest level of information feasible in a given situation will permit using the greatest variety of statistical analyses. Also, regardless of the type of scale used, whenever feasible, several scale items should be used to measure the characteristic of interest. This provides more accurate measurement than a single-item scale. In many situations, it is desirable to use more than one scaling technique or to obtain additional measures using mathematically derived scales.

&&&&&&&&&&&&&&