

**Chapter 7 Scale Reliability and Validity**

The previous chapter examined some of the difficulties with measuring constructs in social science research. For instance, how do we know whether we are measuring “compassion” and not the “empathy”, since both constructs are somewhat similar in meaning? Or is compassion the same thing as empathy? What makes it more complex is that sometimes these constructs are imaginary concepts (i.e., they don’t exist in reality), and multi-dimensional (in which case, we have the added problem of identifying their constituent dimensions). Hence, it is not adequate just to measure social science constructs using any scale that we prefer. We also must test these scales to ensure that: (1) these scales indeed measure the unobservable construct that we wanted to measure (i.e., the scales are “valid”), and (2) they measure the intended construct consistently and precisely (i.e., the scales are “reliable”). Reliability and validity, jointly called the “psychometric properties” of measurement scales, are the yardsticks against which the adequacy and accuracy of our measurement procedures are evaluated in scientific research.

A measure can be reliable but not valid, if it is measuring something very consistently but is consistently measuring the wrong construct. Likewise, a measure can be valid but not reliable if it is measuring the right construct, but not doing so in a consistent manner. Using the analogy of a shooting target, as shown in Figure 7.1, a multiple-item measure of a construct that is both reliable and valid consists of shots that clustered within a narrow range near the center of the target. A measure that is valid but not reliable will consist of shots centered on the target but not clustered within a narrow range, but rather scattered around the target. Finally, a measure that is reliable but not valid will consist of shots clustered within a narrow range but off from the target. Hence, reliability and validity are both needed to assure adequate measurement of the constructs of interest.

## Reliability

Reliability is the degree to which the measure of a construct is consistent or dependable. In other words, if we use this scale to measure the same construct multiple times, do we get pretty much the same result every time, assuming the underlying phenomenon is not changing? An example of an unreliable measurement is people guessing your weight. Quite likely, people will guess differently, the different measures will be inconsistent, and therefore, the “guessing” technique of measurement is unreliable. A more reliable measurement may be to use a weight scale, where you are likely to get the same value every time you step on the scale, unless your weight has actually changed between measurements.

Note that reliability implies consistency but not accuracy. In the previous example of the weight scale, if the weight scale is calibrated incorrectly (say, to shave off ten pounds from your true weight, just to make you feel better!), it will not measure your true weight and is therefore not a valid measure. Nevertheless, the miscalibrated weight scale will still give you the same weight every time (which is ten pounds less than your true weight), and hence the scale is reliable.

What are the sources of unreliable observations in social science measurements? One of the primary sources is the observer’s (or researcher’s) subjectivity. If employee morale in a firm is measured by watching whether the employees smile at each other, whether they make jokes, and so forth, then different observers may infer different measures of morale if they are watching the employees on a very busy day (when they have no time to joke or chat) or a light day (when they are more jovial or chatty). Two observers may also infer different levels of morale on the same day, depending on what they view as a joke and what is not. “Observation” is a qualitative measurement technique. Sometimes, reliability may be improved by using quantitative measures, for instance, by counting the number of grievances filed over one month as a measure of (the inverse of) morale. Of course, grievances may or may not be a valid measure of morale, but it is less subject to human subjectivity, and therefore more reliable. A second source of unreliable observation is asking imprecise or ambiguous questions. For instance, if you ask people what their salary is, different respondents may interpret this question differently as monthly salary, annual salary, or per hour wage, and hence, the resulting observations will likely be highly divergent and unreliable. A third source of unreliability is asking questions about issues that respondents are not very familiar about or care about, such as asking an American college graduate whether he/she is satisfied with Canada’s relationship with Slovenia, or asking a Chief Executive Officer to rate the effectiveness of his company’s technology strategy – something that he has likely delegated to a technology executive.

So how can you create reliable measures? If your measurement involves soliciting information from others, as is the case with much of social science research, then you can start by replacing data collection techniques that depends more on researcher subjectivity (such as observations) with those that are less dependent on subjectivity (such as questionnaire), by asking only those questions that respondents may know the answer to or issues that they care about, by avoiding ambiguous items in your measures (e.g., by clearly stating whether you are looking for annual salary), and by simplifying the wording in your indicators so that they not misinterpreted by some respondents (e.g., by avoiding difficult words whose meanings they may not know). These strategies can improve the reliability of our measures, even though they will not necessarily make the measurements completely reliable. Measurement instruments must still be tested for reliability. There are many ways of estimating reliability, which are discussed next.

Inter-rater reliability . Inter-rater reliability, also called inter-observer reliability, is a measure of consistency between two or more independent raters (observers) of the same construct. Usually, this is assessed in a pilot study, and can be done in two ways, depending on the level of measurement of the construct. If the measure is categorical, a set of all categories is defined, raters check off which category each observation falls in, and the percentage of agreement between the raters is an estimate of inter-rater reliability. For instance, if there are two raters rating 100 observations into one of three possible categories, and their ratings match for 75% of the observations, then inter-rater reliability is 0.75. If the measure is interval or ratio scaled (e.g., classroom activity is being measured once every 5 minutes by two raters on 1 to 7 response scale), then a simple correlation between measures from the two raters can also serve as an estimate of inter-rater reliability.

Test-retest reliability . Test-retest reliability is a measure of consistency between two measurements (tests) of the same construct administered to the same sample at two different points in time. If the observations have not changed substantially between the two tests, then the measure is reliable. The correlation in observations between the two tests is an estimate of test-retest reliability. Note here that the time interval between the two tests is critical. Generally, the longer is the time gap, the greater is the chance that the two observations may change during this time (due to random error), and the lower will be the test-retest reliability.

Split-half reliability . Split-half reliability is a measure of consistency between two halves of a construct measure. For instance, if you have a ten-item measure of a given construct, randomly split those ten items into two sets of five (unequal halves are allowed if the total number of items is odd), and administer the entire instrument to a sample of respondents. Then, calculate the total score for each half for each respondent, and the correlation between the total scores in each half is a measure of split-half reliability. The longer is the instrument, the more likely it is that the two halves of the measure will be similar (since random errors are minimized as more items are added), and hence, this technique tends to systematically overestimate the reliability of longer instruments.

Internal consistency reliability . Internal consistency reliability is a measure of consistency between different items of the same construct. If a multiple-item construct measure is administered to respondents, the extent to which respondents rate those items in a similar manner is a reflection of internal consistency. This reliability can be estimated in terms of average inter-item correlation, average item-to-total correlation, or more commonly,

Cronbach’s alpha. As an example, if you have a scale with six items, you will have fifteen different item pairings, and fifteen correlations between these six items. Average inter-item correlation is the average of these fifteen correlations. To calculate average item-to-total correlation, you have to first create a “total” item by adding the values of all six items, compute the correlations between this total item and each of the six individual items, and finally, average the six correlations. Neither of the two above measures takes into account the number of items in the measure (six items in this example).

## Validity

Validity , often called construct validity, refers to the extent to which a measure adequately represents the underlying construct that it is supposed to measure. For instance, is a measure of compassion really measuring compassion, and not measuring a different construct such as empathy? Validity can be assessed using theoretical or empirical approaches, and should ideally be measured using both approaches. Theoretical assessment of validity focuses on how well the idea of a theoretical construct is translated into or represented in an operational measure. This type of validity is called translational validity (or representational validity), and consists of two subtypes: face and content validity. Translational validity is typically assessed using a panel of expert judges, who rate each item (indicator) on how well they fit the conceptual definition of that construct, and a qualitative technique called Q-sort.

Empirical assessment of validity examines how well a given measure relates to one or more external criterion, based on empirical observations. This type of validity is called criterion-related validity , which includes four sub-types: convergent, discriminant, concurrent, and predictive validity. While translation validity examines whether a measure is a good reflection of its underlying construct, criterion -related validity examines whether a given measure behaves the way it should, given the theory of that construct. This assessment is based on quantitative analysis of observed data using statistical techniques such as correlational analysis, factor analysis, and so forth. The distinction between theoretical and empirical assessment of validity is illustrated in Figure 7.2. However, both approaches are needed to adequately ensure the validity of measures in social science research.



Note that the different types of validity discussed here refer to the validity of the measurement procedures , which is distinct from the validity of hypotheses testing procedures , such as internal validity (causality), external validity (generalizability), or statistical conclusion validity. The latter types of validity are discussed in a later chapter.

Face validity . Face validity refers to whether an indicator seems to be a reasonable measure of its underlying construct “on its face”. For instance, the frequency of one’s attendance at religious services seems to make sense as an indication of a person’s religiosity without a lot of explanation. Hence this indicator has face validity. However, if we were to suggest how many books were checked out of an office library as a measure of employee morale, then such a measure would probably lack face validity because it does not seem to make much sense. Interestingly, some of the popular measures used in organizational research appears to lack face validity. For instance, absorptive capacity of an organization (how much new knowledge can it assimilate for improving organizational processes) has often been measured as research and development intensity (i.e., R&D expenses divided by gross revenues)! If your research includes constructs that are highly abstract or constructs that are hard to conceptually separate from each other (e.g., compassion and empathy), it may be worthwhile to consider using a panel of experts to evaluate the face validity of your construct measures.

Content validity. Content validity is an assessment of how well a set of scale items matches with the relevant content domain of the construct that it is trying to measure. For instance, if you want to measure the construct “satisfaction with restaurant service,” and you define the content domain of restaurant service as including the quality of food, courtesy of wait staff, duration of wait, and the overall ambience of the restaurant (i.e., whether it is noisy, smoky, etc.), then for adequate content validity, this construct should be measured using indicators that examine the extent to which a restaurant patron is satisfied with the quality of food, courtesy of wait staff, the length of wait, and the restaurant’s ambience. Of course, this approach requires a detailed description of the entire content domain of a construct, which may be difficult for complex constructs such as self-esteem or intelligence. Hence, it may not be always possible to adequately assess content validity. As with face validity, an expert panel of judges may be employed to examine content validity of constructs.

Convergent validity refers to the closeness with which a measure relates to (or converges on) the construct that it is purported to measure, and discriminant validity refers to the degree to which a measure does not measure (or discriminates from) other constructs that it is not supposed to measure. Usually, convergent validity and discriminant validity are assessed jointly for a set of related constructs. For instance, if you expect that an organization’s knowledge is related to its performance, how can you assure that your measure of organizational knowledge is indeed measuring organizational knowledge (for convergent validity) and not organizational performance (for discriminant validity)? Convergent validity can be established by comparing the observed values of one indicator of one construct with that of other indicators of the same construct and demonstrating similarity (or high correlation) between values of these indicators. Discriminant validity is established by demonstrating that indicators of one construct are dissimilar from (i.e., have low correlation with) other constructs. In the above example, if we have a three-item measure of organizational knowledge and three more items for organizational performance, based on observed sample data, we can compute bivariate correlations between each pair of knowledge and performance items. If this correlation matrix shows high correlations within items of the organizational knowledge and organizational performance constructs, but low correlations between items of these constructs, then we have simultaneously demonstrated convergent and discriminant validity (see Table 7.1).



**https://courses.lumenlearning.com/suny-hccc-research-methods/chapter/chapter-7-scale-reliability-and-validity/**