Chapter 6

Measurement of Constructs

Theoretical propositions consist of relationships between abstract constructs. Testing theories (i.e., theoretical propositions) require measuring these constructs accurately, correctly, and in a scientific manner, before the strength of their relationships can be tested. Measurement refers to careful, deliberate observations of the real world and is the essence of empirical research. While some constructs in social science research, such as a person's age, weight, or a firm's size, may be easy to measure, other constructs, such as creativity, prejudice, or alienation, may be considerably harder to measure. In this chapter, we will examine the related processes of conceptualization and operationalization for creating measures of such constructs.

Conceptualization

Conceptualization is the mental process by which fuzzy and imprecise constructs (concepts) and their constituent components are defined in concrete and precise terms. For instance, we often use the word "prejudice" and the word conjures a certain image in our mind; however, we may struggle if we were asked to define exactly what the term meant. If someone says bad things about other racial groups, is that racial prejudice? If women earn less than men for the same job, is that gender prejudice? If churchgoers believe that non-believers will burn in hell, is that religious prejudice? Are there different kinds of prejudice, and if so, what are they? Are there different levels of prejudice, such as high or low? Answering all of these questions is the key to measuring the prejudice construct correctly. The process of understanding what is included and what is excluded in the concept of prejudice is the conceptualization process.

The conceptualization process is all the more important because of the imprecision, vagueness, and ambiguity of many social science constructs. For instance, is "compassion" the same thing as "empathy" or "sentimentality"? If you have a proposition stating that "compassion is positively related to empathy", you cannot test that proposition unless you can conceptually separate empathy from compassion and then empirically measure these two very similar constructs correctly. If deeply religious people believe that some members of their society, such as nonbelievers, gays, and abortion doctors, will burn in hell for their sins, and forcefully try to change the "sinners" behaviors to prevent them from going to hell, are they acting in a prejudicial manner or a compassionate manner? Our definition of such constructs is not based on any objective criterion, but rather on a shared ("inter-subjective") agreement between our mental images (conceptions) of these constructs.

While defining constructs such as prejudice or compassion, we must understand that sometimes, these constructs are not real or can exist independently, but are simply imaginary creations in our mind. For instance, there may be certain tribes in the world who lack prejudice and who cannot even imagine what this concept entails. But in real life, we tend to treat this concept as real. The process of regarding mental constructs as real is called *reification*, which is central to defining constructs and identifying measurable variables for measuring them.

One important decision in conceptualizing constructs is specifying whether they are unidimensional and multidimensional. **Unidimensional** constructs are those that are expected to have a single underlying dimension. These constructs can be measured using a single measure or test. Examples include simple constructs such as a person's weight, wind speed, and probably even complex constructs like self-esteem (if we conceptualize self-esteem as consisting of a single dimension, which of course, may be a unrealistic assumption). **Multidimensional** constructs consist of two or more underlying dimensions. For instance, if we conceptualize a person's academic aptitude as consisting of two dimensions – mathematical and verbal ability – then academic aptitude is a multidimensional construct. Each of the underlying dimensions in this case must be measured separately, say, using different tests for mathematical and verbal ability, and the two scores can be combined, possibly in a weighted manner, to create an overall value for the academic aptitude construct.

Operationalization

Once a theoretical construct is defined, exactly how do we measure it? **Operationalization** refers to the process of developing **indicators** or items for measuring these constructs. For instance, if an unobservable theoretical construct such as socioeconomic status is defined as the level of family income, it can be operationalized using an indicator that asks respondents the question: what is your annual family income? Given the high level of subjectivity and imprecision inherent in social science constructs, we tend to measure most of those constructs (except a few demographic constructs such as age, gender, education, and income) using multiple indicators. This process allows us to examine the closeness amongst these indicators as an assessment of their accuracy (reliability).

Indicators operate at the empirical level, in contrast to constructs, which are conceptualized at the theoretical level. The combination of indicators at the empirical level representing a given construct is called a **variable**. As noted in a previous chapter, variables may be independent, dependent, mediating, or moderating, depending on how they are employed in a research study. Also each indicator may have several **attributes** (or levels) and each attribute represent a **value**. For instance, a "gender" variable may have two attributes: male or female. Likewise, a customer satisfaction scale may be constructed to represent five attributes: "strongly dissatisfied", "somewhat dissatisfied", "neutral", "somewhat satisfied" and "strongly satisfied". Values of attributes may be **quantitative** (numeric) or **qualitative** (non-numeric). Quantitative data can be analyzed using quantitative data analysis techniques, such as regression or structural equation modeling, while qualitative data require qualitative data analysis techniques, such as coding. Note that many variables in social science research are qualitative, even when represented in a quantitative manner. For instance, we can create a customer satisfaction indicator with five attributes: strongly dissatisfied, somewhat dissatisfied, neutral, somewhat satisfied, and strongly satisfied, and assign numbers 1 through 5 respectively for these five attributes, so that we can use sophisticated statistical tools for quantitative data analysis. However, note that the numbers are only labels associated with

respondents' personal evaluation of their own satisfaction, and the underlying variable (satisfaction) is still qualitative even though we represented it in a quantitative manner.

Indicators may be reflective or formative. A **reflective indicator** is a measure that "reflects" an underlying construct. For example, if religiosity is defined as a construct that measures how religious a person is, then attending religious services may be a reflective indicator of religiosity. A **formative indicator** is a measure that "forms" or contributes to an underlying construct. Such indicators may represent different dimensions of the construct of interest. For instance, if religiosity is defined as composing of a belief dimension, a devotional dimension, and a ritual dimension, then indicators chosen to measure each of these different dimensions will be considered formative indicators. Unidimensional constructs are measured using reflective indicators (even though multiple reflective indicators may be used for measuring abstruse constructs such as self-esteem), while multidimensional constructs are measured as a formative combination of the multiple dimensions, even though each of the underlying dimensions may be measured using one or more reflective indicators.

Levels of Measurement

The first decision to be made in operationalizing a construct is to decide on what is the intended level of measurement. **Levels of measurement**, also called **rating scales**, refer to the values that an indicator can take (but says nothing about the indicator itself). For example, male and female (or M and F, or 1 and 2) are two levels of the indicator "gender." In his seminal article titled "On the theory of scales of measurement" published in *Science* in 1946, psychologist Stanley Smith Stevens (1946) defined four generic types of rating scales for scientific measurements: nominal, ordinal, interval, and ratio scales. The statistical properties of these scales are shown in Table 6.1.

Scale	Central Tendency	Statistics	Transformations
Nominal	Mode	Chi-square	One-to-one (equality)
Ordinal	Median	Percentile, non-parametric statistics	Monotonic increasing (order)
Interval	Arithmetic mean, range, standard deviation	Correlation, regression, analysis of variance	Positive linear (affine)
Ratio	Geometric mean, harmonic mean	Coefficient of variation	Positive similarities (multiplicative, logarithmic)
Note: All higher-order scales can use any of the statistics for lower order scales.			

Table 6.1. Statistical properties of rating scales

Nominal scales, also called categorical scales, measure categorical data. These scales are used for variables or indicators that have mutually exclusive attributes. Examples include gender (two values: male or female), industry type (manufacturing, financial, agriculture, etc.), and religious affiliation (Christian, Muslim, Jew, etc.). Even if we assign unique numbers to each value, for instance 1 for male and 2 for female, the numbers don't really mean anything (i.e., 1 is not less than or half of 2) and could have been easily been represented non-numerically, such as

M for male and F for female. Nominal scales merely offer *names* or *labels* for different attribute values. The appropriate measure of central tendency of a nominal scale is mode, and neither the mean nor the median can be defined. Permissible statistics are chi-square and frequency distribution, and only a one-to-one (equality) transformation is allowed (e.g., 1=Male, 2=Female).

Ordinal scales are those that measure *rank-ordered* data, such as the ranking of students in a class as first, second, third, and so forth, based on their grade point average or test scores. However, the actual or relative values of attributes or difference in attribute values cannot be assessed. For instance, ranking of students in class says nothing about the actual GPA or test scores of the students, or how they well performed relative to one another. A classic example in the natural sciences is Moh's scale of mineral hardness, which characterizes the hardness of various minerals by their ability to scratch other minerals. For instance, diamonds can scratch all other naturally occurring minerals on earth, and hence diamond is the "hardest" mineral. However, the scale does not indicate the actual hardness of these minerals or even provides a relative assessment of their hardness. Ordinal scales can also use attribute labels (anchors) such as "bad", "medium", and "good", or "strongly dissatisfied", "somewhat dissatisfied", "neutral", or "somewhat satisfied", and "strongly satisfied". In the latter case, we can say that respondents who are "somewhat satisfied" are less satisfied than those who are "strongly satisfied", but we cannot quantify their satisfaction levels. The central tendency measure of an ordinal scale can be its median or mode, and means are uninterpretable. Hence, statistical analyses may involve percentiles and non-parametric analysis, but more sophisticated techniques such as correlation, regression, and analysis of variance, are not appropriate. Monotonically increasing transformation (which retains the ranking) is allowed.

Interval scales are those where the values measured are not only rank-ordered, but are also equidistant from adjacent attributes. For example, the temperature scale (in Fahrenheit or Celsius), where the difference between 30 and 40 degree Fahrenheit is the same as that between 80 and 90 degree Fahrenheit. Likewise, if you have a scale that asks respondents' annual income using the following attributes (ranges): $0 to 10,000, $10,000 to 20,000, $20,000 to 30,000, and so forth, this is also an interval scale, because the mid-point of each range (i.e., $5,000, $15,000, $25,000, etc.) are equidistant from each other. The intelligence quotient (IQ) scale is also an interval scale, because the scale is designed such that the difference between IQ scores 100 and 110 is supposed to be the same as between 110 and 120 (although we do not really know whether that is truly the case). Interval scale allows us to examine "how much more" is one attribute when compared to another, which is not possible with nominal or ordinal scales. Allowed central tendency measures include mean, median, or mode, as are measures of dispersion, such as range and standard deviation. Permissible statistical analyses include all of those allowed for nominal and ordinal scales, plus correlation, regression, analysis of variance, and so on. Allowed scale transformation are positive linear. Note that the satisfaction scale discussed earlier is not strictly an interval scale, because we cannot say whether the difference between "strongly satisfied" and "somewhat satisfied" is the same as that between "neutral" and "somewhat satisfied" or between "somewhat dissatisfied" and "strongly dissatisfied". However, social science researchers often "pretend" (incorrectly) that these differences are equal so that we can use statistical techniques for analyzing ordinal scaled data.

Ratio scales are those that have all the qualities of nominal, ordinal, and interval scales, and in addition, also have a "true zero" point (where the value zero implies lack or non-availability of the underlying construct). Most measurement in the natural sciences and engineering, such as mass, incline of a plane, and electric charge, employ ratio scales, as are

some social science variables such as age, tenure in an organization, and firm size (measured as employee count or gross revenues). For example, a firm of size zero means that it has no employees or revenues. The Kelvin temperature scale is also a ratio scale, in contrast to the Fahrenheit or Celsius scales, because the zero point on this scale (equaling -273.15 degree Celsius) is not an arbitrary value but represents a state where the particles of matter at this temperature have zero kinetic energy. These scales are called "ratio" scales because the ratios of two points on these measures are meaningful and interpretable. For example, a firm of size 10 employees is double that of a firm of size 5, and the same can be said for a firm of 10,000 employees relative to a different firm of 5,000 employees. All measures of central tendencies, including geometric and harmonic means, are allowed for ratio scales, as are ratio measures, such as studentized range or coefficient of variation. All statistical methods are allowed. Sophisticated transformation such as positive similar (e.g., multiplicative or logarithmic) are also allowed.

Based on the four generic types of scales discussed above, we can create specific rating scales for social science research. Common rating scales include binary, Likert, semantic differential, or Guttman scales. Other less common scales are not discussed here.

Binary scales. Binary scales are nominal scales consisting of binary items that assume one of two possible values, such as yes or no, true or false, and so on. For example, a typical binary scale for the "political activism" construct may consist of the six binary items shown in Table 6.2. Each item in this scale is a binary item, and the total number of "yes" indicated by a respondent (a value from 0 to 6) can be used as an overall measure of that person's political activism. To understand how these items were derived, refer to the "Scaling" section later on in this chapter. Binary scales can also employ other values, such as male or female for gender, full-time or part-time for employment status, and so forth. If an employment status item is modified to allow for more than two possible values (e.g., unemployed, full-time, part-time, and retired), it is no longer binary, but still remains a nominal scaled item.

Have you ever written a letter to a public official	Yes	No
Have you ever signed a political petition	Yes	No
Have you ever donated money to a political cause	Yes	No
Have you ever donated money to a candidate running for public office	Yes	No
Have you ever written a political letter to the editor of a newspaper or magazine	Yes	No
Have you ever persuaded someone to change his/her voting plans	Yes	No

Table 6.2. A six-item binary scale for measuring political activism

Likert scale. Designed by Rensis Likert, this is a very popular rating scale for measuring ordinal data in social science research. This scale includes Likert items that are simply-worded statements to which respondents can indicate their extent of agreement or disagreement on a five or seven-point scale ranging from "strongly disagree" to "strongly agree". A typical example of a six-item Likert scale for the "employment self-esteem" construct is shown in Table 6.3. Likert scales are summated scales, that is, the overall scale score may be a summation of the attribute values of each item as selected by a respondent.

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
I feel good about my job	1	2	3	4	5
I get along well with others at work	1	2	3	4	5
I'm proud of my relationship with my supervisor at work	1	2	3	4	5
I can tell that other people at work are glad to have me there	1	2	3	4	5
I can tell that my coworkers respect me	1	2	3	4	5
I feel like I make a useful contribution at work	1	2	3	4	5

Table 6.3. A six-item Likert scale for measuring employment self-esteem

Likert items allow for more granularity (more finely tuned response) than binary items, including whether respondents are neutral to the statement. Three or nine values (often called "anchors") may also be used, but it is important to use an odd number of values to allow for a "neutral" (or "neither agree nor disagree") anchor. Some studies have used a "forced choice approach" to force respondents to agree or disagree with the LIkert statement by dropping the neutral mid-point and using even number of values and, but this is not a good strategy because some people may indeed be neutral to a given statement and the forced choice approach does not provide them the opportunity to record their neutral stance. A key characteristic of a Likert scale is that even though the statements vary in different items or indicators, the anchors ("strongly disagree" to "strongly agree") remain the same. Likert scales are ordinal scales because the anchors are not necessarily equidistant, even though sometimes we treat them like interval scales.

How would you rate your opinions on national health insurance?						
	Very much	Somewhat	Neither	Somewhat	Very much	
Good	□	□	□	□	□	Bad
Useful	□	□	□	□	□	Useless
Caring	□	□	□	□	□	Uncaring
Interesting	□	□	□	□	□	Boring

Table 6.4. A semantic differential scale for measuring attitude toward national health insurance

Semantic differential scale. This is a composite (multi-item) scale where respondents are asked to indicate their opinions or feelings toward a single statement using different pairs of adjectives framed as polar opposites. For instance, the construct "attitude toward national health insurance" can be measured using four items shown in Table 6.4. As in the Likert scale, the overall scale score may be a summation of individual item scores. Notice that in Likert scales, the statement changes but the anchors remain the same across items. However, in semantic differential scales, the statement remains constant, while the anchors (adjective pairs)