Bottom line: Regardless of how control variables affect variables, the point is to identify rival hypotheses (control variables) that may affect study conclusions.

6. A useful acronym for remembering these threats to internal validity is "Mis Smith:" maturation, instrumentation, selection, statistical regression, mortality, imitation, testing, and history. The classic source for these distinctions is Donald Campbell and Julian Stanley, *Experimental and Quasi-experimental Designs for Research* (Chicago: Rand McNally), 1963.
7. Even the classic, randomized research design is subject to some of these validity threats. Threats to external validity (generalizability) are a problem in any setting, and problems of history and mortality also may be present. It is not a given that the experimental and control groups experience the same intervening events (history), and they may have different rates of attrition (mortality). Testing might affect both groups, too. To address the problem of testing, a modification of the classic, randomized research design is the Solomon four-group design:

	Pretest	Program	Posttest
Group 1:	R O1	X	O2
Group 2:	R O3		O4
Group 3:	R	X	O5
Group 4:	R		O6

In this design, groups 3 and 4 allow the researcher to control for the impact of pretesting on groups 1 and 2.

8. For example, in a simple posttest design with no comparison group, the evaluation of anger management should consider whether any intervening effects (history) occurred that could have affected students' levels of anger and violence. Analysts will want to examine and account for possible maturation, statistical regression, and sample bias effects on the results. They also will want to assess subjects' knowledge of other efforts used elsewhere (which could give rise to imitation or rivalry) and ensure that the assessment method is accurate (minimizing effects of instrumentation).

Conceptualization and Measurement

CHAPTER OBJECTIVES

After reading this chapter, you should be able to

- Appreciate the challenge of measuring abstract concepts
- Implement methods for measuring abstract concepts
- Distinguish between different levels of measurement
- Apply a variety of Likert scales
- Create index variables
- Understand criteria for assessing measurement validity

Measurement is a foundation of science and knowledge. How well phenomena are measured affects what we know about them, and rigor in measurement increases the validity of analytical work. This chapter discusses key concepts of measurement and shows how to apply these measurements in analytical work such as program evaluation. This chapter also shows how to make index variables.

MEASUREMENT LEVELS AND SCALES

A ***scale*** is defined as the collection of attributes used to measure a specific variable. For example, the variable "gender" is commonly measured on a

scale defined by the specific attributes "male" and "female." Scales are important because they define the nature of information about variables. For example, we can measure incomes by asking respondents for their exact income or by asking them to identify their income using prespecified income brackets. Scales vary greatly—some are unique to the variables they measure, such as the Richter scale, which measures the strength of earthquakes; others are used for many different purposes, such as response scales found in survey questionnaires. Managers should be familiar with different types of scales so that they can adapt them to their needs.

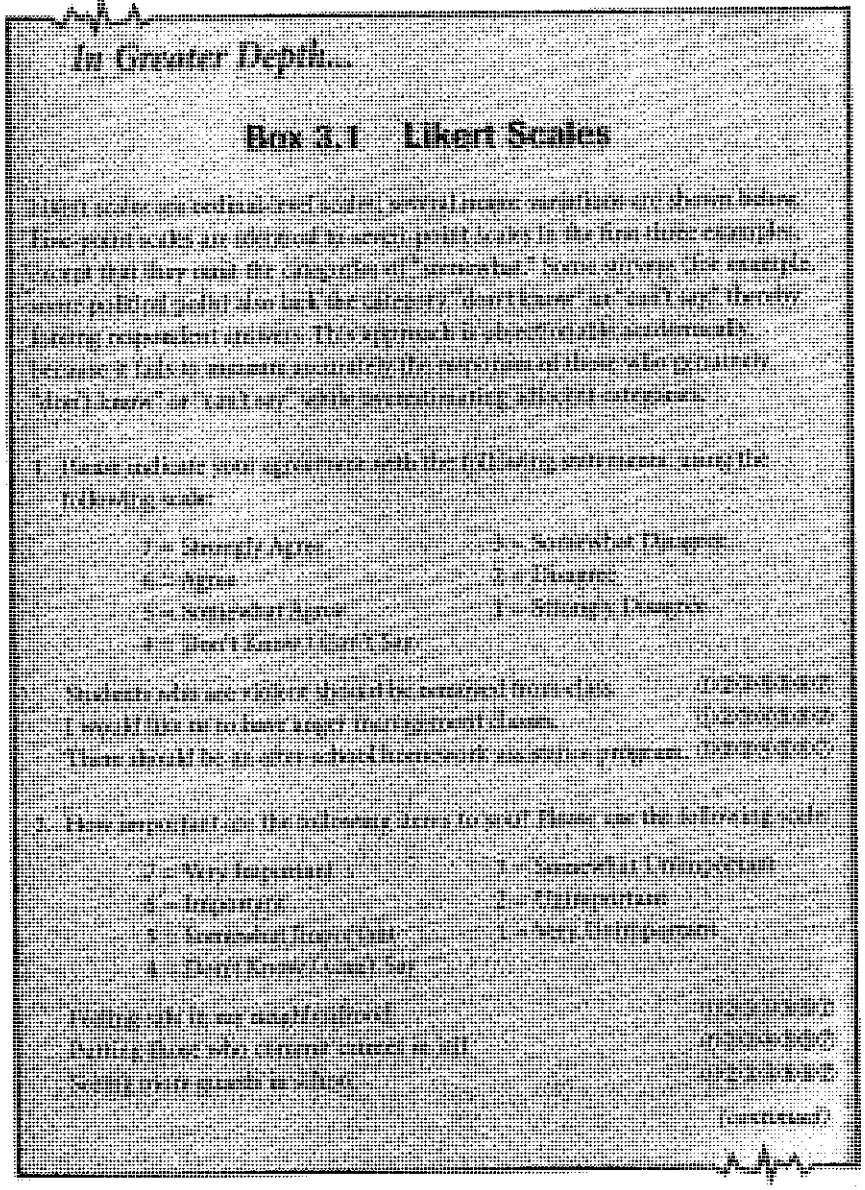
Measurement scales are distinguished by their level of measurement. There are four levels of measurement: *nominal, ordinal, interval,* and *ratio.* A variable that has, for example, an ordinal-level measurement scale is commonly referred to as an ordinal-level variable or, simply, as an ordinal variable. The importance of the ***measurement level*** is threefold: (1) it determines the selection of test statistics (highly relevant to subsequent chapters), (2) it affects the amount of information collected about variables, and (3) it affects how survey and other types of questions are phrased.

A ***nominal-level scale*** exhibits no ordering among the categories. It provides the least amount of information. For example, the variable "gender" has a nominal scale because there is no ordering among the attributes "men" and "women." We cannot say that "men" are more than "women," regardless of any coding scheme that assigns numbers to these categories; they are nominal categories, only. "Region" is another common nominal scale: no ordering exists among the values of North, South, East, and West.

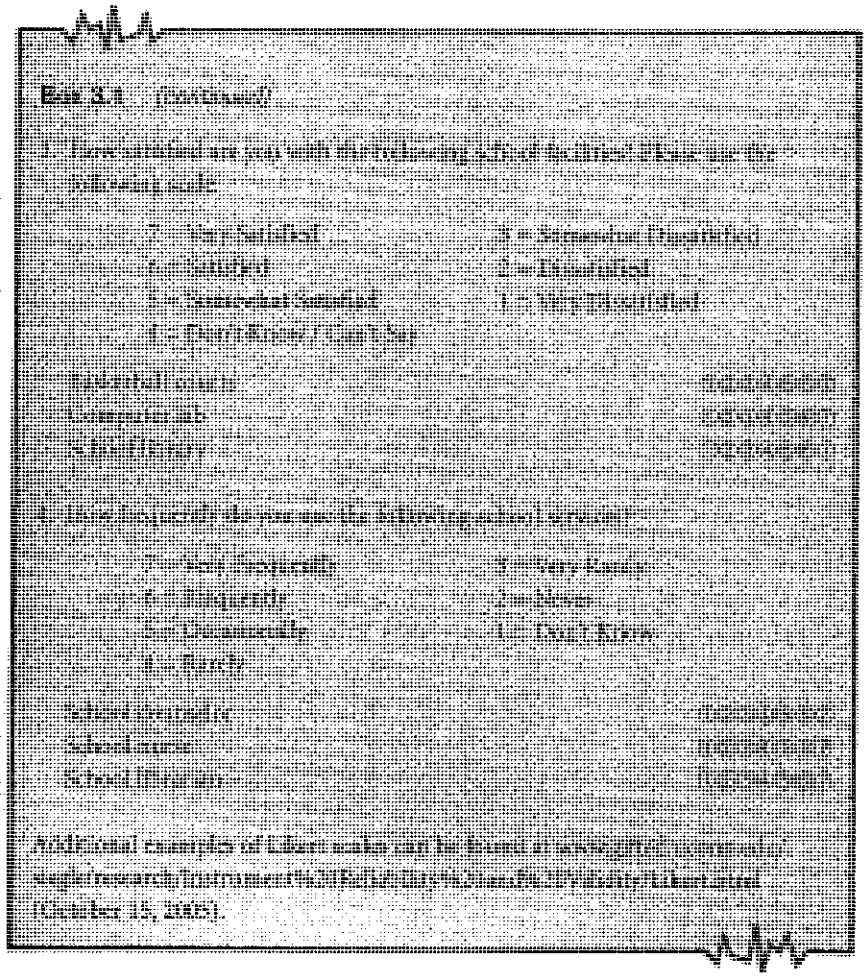
By contrast, an ***ordinal-level scale*** exhibits order among categories (hence, the name *ordinal*), though without exact distances between successive categories. "Order" means that categories can be compared as being "more" or "less" than one another. For example, assume that we measure teenage anger by asking adolescents whether they feel irritated, aggravated, or raging mad. Clearly someone feeling "raging mad" is more angry than someone who feels only "aggravated," who in turn is more angry than someone who feels "irritated." "Distance" means that we can measure how much more one category is than another. Ordinal scales lack distance. Although we can say that "raging mad" is more angry than "aggravated," we cannot say *how much* more angry "raging mad" is than "aggravated. Collectively, ordinal- and nominal-level variables are called ***categorical*** (or ***discrete***) ***variables.***

Likert scales are a common type of ordinal scale. Developed in 1932 by Professor Rensis Likert, these scales are now a staple in surveys that measure attitudes. The responses used on Likert scales come in many variations, such as Strongly Agree, Agree, Somewhat Agree, Don't Know, Somewhat Disagree, Disagree, and Strongly Disagree. Survey respondents are read statements (for

example, "I feel safe at school") and are then asked, after each statement, to respond by selecting one of the responses. Likert scales demonstrate order and the absence of distance between categories: "strongly agree" is a higher level of agreeing than just "agree," but we cannot say how much more. Likert


In Greater Depth...

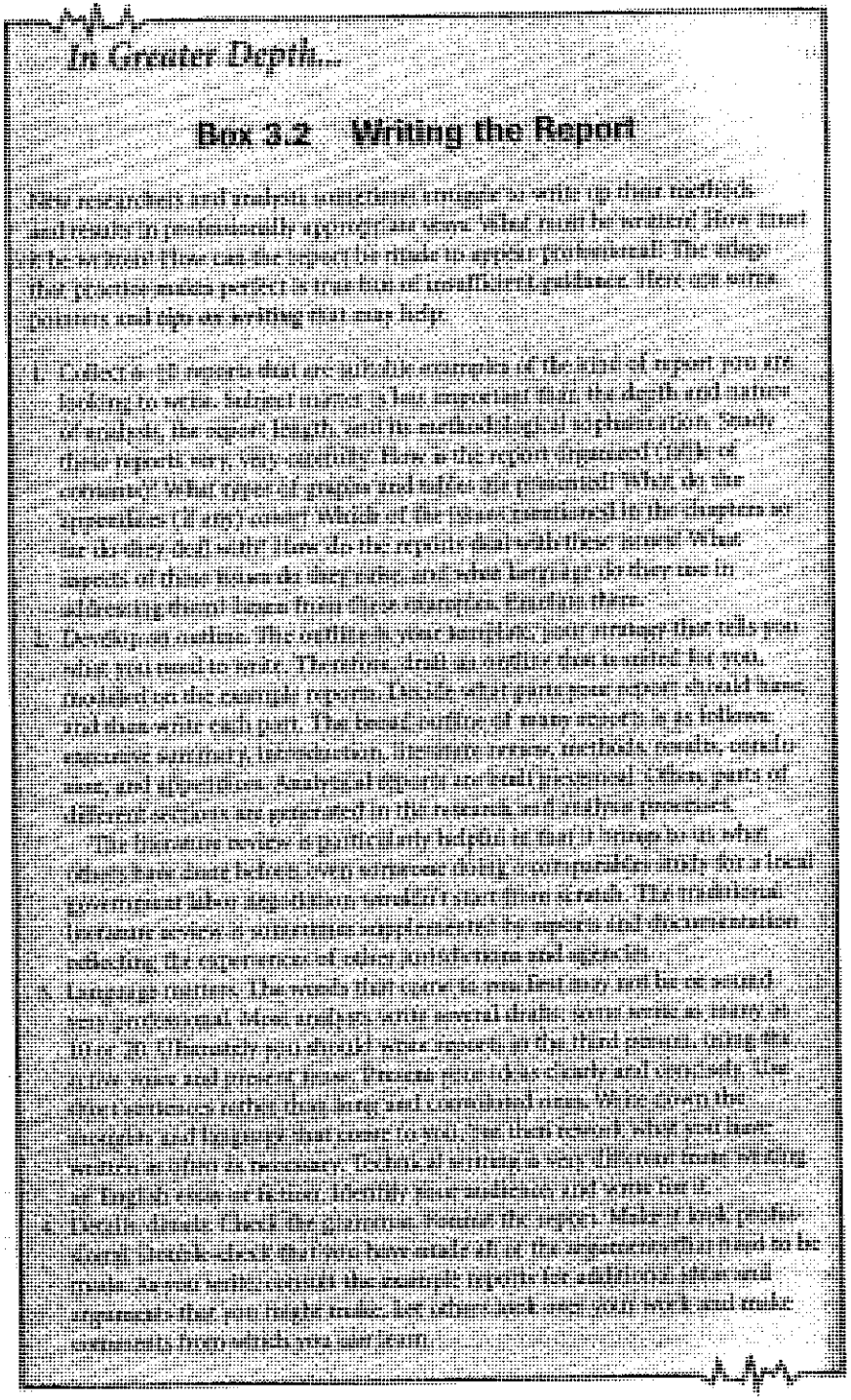
Box 3.1 Likert Scales



scales are so popular because they have been widely tested and are easy to use and adapt. Box 3.1 provide further examples of these important scales.[1]

Interval- and ***ratio-level scales*** exhibit both order and distance among categories. We can say that someone who exercises daily does so seven times more often than someone who exercises only weekly. Someone who earns $75,000 per year makes exactly three times that of someone making $25,000. The *only* difference between interval and ratio scales is that the latter have a true "zero" (for example, income can be zero, but IQ cannot). The distinction between ratio- and interval-level variables is typically of little relevance to public and nonprofit administration and policy analysis. Variables with interval- and ratio-level scales are also called ***continuous variables.***[2]

In general, we strongly prefer having more information about variables than less. Therefore, continuous-level scales are preferred over ordinal-level



scales, which in turn are preferred over nominal-level scales. We also prefer ordinal and nominal scales that have more rather than fewer categories. Of course, a variable such as "gender" cannot be made ordinal or continuous, and it can have only two categories. Likewise, some variables can only be ordinal and cannot be continuous.[3]

The development of measures and scales is a precise task. We must avoid scales that are incomplete, ambiguous, or overlapping. An *incomplete scale* might omit "zero" as a response category when asking respondents how many fist fights they witnessed. An *ambiguous scale* is one that asks respondents to answer a question about the presence of violence "on a scale of 1 to 10" without defining each value. Respondents may have different definitions of any specific value, such as the value of "6." In an *overlapping scale,* at least one response is covered by more than one category. An example of such a scale is one that measures income with brackets $20,000–$40,000 and $40,000–$60,000. It is better to use $20,000–$39,999 and $40,000–$59,999. Another example is measuring "fist fights" and "scuffles" as separate categories.

Problems with measurement scales can affect the validity of one's findings. Measurement scales should be complete, unambiguous, and have unique categories for each response. Of course, other measurement challenges also exist, such as well-known problems of using leading (or biased) survey questions as well as samples that are biased or restricted in some way; these matters are discussed further in Chapter 5.[4] We now turn to another measurement topic, that of measuring abstract concepts.

CONCEPTUALIZATION

Many important matters of public and nonprofit management and analysis involve abstract concepts, such as notions of democracy, effectiveness, volunteerism, citizen satisfaction, and, yes, high school violence and anger. The rigor with which study concepts are defined (such as the level of anger or high school violence) enhances the validity of our efforts. ***Measurement validity*** simply means that something measures or reflects what it is intended to. This is identified as step 4 in the six-step model of program evaluation (see Chapter 2). A research task needs to be clear about what is being studied. For example, how should we measure the concept "high school violence?"

In this regard, variables must be distinguished from concepts. Whereas variables belong to the realm of directly observable phenomena, concepts belong to the realm of ideas. ***Concepts*** are abstract ideas that are observed indirectly, through variables. ***Processes of concept measurement*** typically have two steps. First we need to be clear about the meaning of the concept and, in particular, identify all of the relevant dimensions of the concept. This is called ***conceptualization***. Then we need to identify and define the variable(s) that will be used to measure the concept and its dimensions. This is called ***operationalization***.

This process is best explained through an example. Suppose we want to measure the concept "student anger." First we need to be clear about what "student anger" means. What is the essence of this concept?[5] How might we best define it in the context of our study? Because our program aims to reduce and control manifestations of anger and violence that can be disruptive, we might define "student anger" as "a strong emotion of displeasure by students that may be triggered by, or directed toward, specific or general grievances." Of course, this is not the only way to define student anger, and certainly some other definitions might be better. Perhaps you know of a better definition? We can justify this definition, however, through criteria that are commonly used for this purpose: consistency with generally understood meanings of the concept (here, anger), consistency with expert understandings and studies, and being relevant and central to the program and its evaluation.

Next, and still part of conceptualization, we need to ask whether any discernible, distinct dimensions of this concept should be considered and measured separately? Assume that we identify the following three dimensions to the concept "anger": (1) emotions of anger, (2) thoughts (cognition) of anger, and (3) physical rage. Each dimension stands alone and can be measured separately. For example, some students might have thoughts of anger but little emotion associated with these thoughts, and vice versa. Some may have rage and emotion but little cognition. These are different dimensions of anger. Only after the dimensions of "student anger" have been identified can the analytical task shift toward developing a process for measuring these dimensions (that is, operationalization).

It is not a given that this concept will or should always have these dimensions. Complex concepts and those that are key to the research design are usually conceptualized with greater rigor than those that are simple or less key to the program or evaluation. In the case of evaluating the anger management program, "student anger" is an important concept and one that managers and analysts will want to examine carefully. Yet this concept might not be of much importance in a study about, say, student achievement. Such a study might choose to measure "student anger" in only a cursory way, perhaps as just a single item (for example, "How angry do you usually feel?"). Decisions about study rigor and the importance of specific study concepts drive thoroughness.

When study concepts are conceptualized with rigor and thoroughness, analysts need to determine how many concept dimensions they will identify and measure. An imprecise guiding principle is that analysts should be true and comprehensive with regard to their concepts. Typically two to five dimensions are used in rigorous conceptualizations, usually based on (1) the

consensus of past studies, (2) whether concept definitions include disparate facets or dimensions, (3) program needs that might suggest dimensions, and (4) practical constraints in the ability to collect data.[6] Our example of conceptualizing "student anger" reflects judgments that it is a key study concept requiring rigor, that the three identified dimensions reflect a comprehensive and appropriate understanding of the concept, and that the concept is relevant to program management.

Analysts must justify their choices about the conceptualization and operationalization of study concepts. An important perspective is that no correct number of dimensions or variables exists, only bad or lacking ones.[7]

Another example involves the conceptualization of "high school violence." Assume that after defining this concept, conducting a brief literature review, and talking with program officials, we reach a consensus that high school violence has three dimensions: (1) use of weapons, (2) inappropriate physical contact (occurring without weapons and not involving sanctioned physical contact during sports activities), and (3) verbal assaults. These are seen not as degrees of violence, but as three different dimensions (or types) of violence. Students can have physical contact without necessarily using weapons or involving verbal assault. These dimensions of violence can be measured separately.

OPERATIONALIZATION

As we discussed earlier, the development of specific measures is called operationalization. This process develops the specific variables that will be used to measure a concept. Three approaches to operationalization are (1) to develop separate measures for each dimension, (2) to develop a single set of measures that encompass the dimensions, or (3) to develop a single measure. These three strategies reflect a *declining* order of rigor.[8]

The *first* strategy is the most comprehensive approach—measuring each dimension separately. By way of example, Table 3.1 lays out the basic measurement strategy for conceptualizing and operationalizing the three dimensions of high school violence.

Whereas the table shows a mix of objective data and subjective assessments, this is not always the case nor is it always necessary. Student perceptions might be assessed through a survey in which students are asked to evaluate such statements. Typically 5–10 questions are used to measure each dimension. For example, the following questions might be used to assess student perceptions of inappropriate physical contact at school (dimension 2):

> Please tell me which of the following you experienced in or around school, during the last month, which were not part of any normal sports activity:
>
> I was involved in a fight or scuffle.
> I was physically injured in a fight or scuffle.
> I was pushed or tripped by someone who, I believe, tried to injure me.
> I was touched sexually in ways that were unwanted by me.
> I was struck by someone with an object (such as a stick or stone).
> I was assaulted in some way but not injured.
> I was physically hurt in some other way (please specify). . . .

Table 3.1 Measuring High School Violence

Dimension	Measurement
1: Use of weapons	Number of students caught using weapons Student perception of presence of weapons (guns, knives, other)
2: Physical contact	Number of fights and scuffles reported to administrators Number of inappropriate physical contacts reported to administrators (sexual and nonsexual) Student perception of fights, scuffles, and inappropriate physical contact (sexual and nonsexual)
3: Verbal assaults and threats	Number of harassment allegations brought to administrators Student perception of verbal assaults and threats

This is certainly not the only way to assess student perceptions of inappropriate physical contact. Some assessments might ask different questions. Unwanted sexual contact is included as a form of violence. As indicated in the table, objective data might also be included for each dimension. In the same way, survey items would be developed to measure the other two dimensions, "use of weapons" and "verbal assaults and threats." The correct and complete development of study measures finishes the process of operationalization. Chapter 5 provides additional guidance on developing survey questions and collecting data.

The *second*, less rigorous (but still comprehensive) approach is to develop questions that each measure a different aspect of high school violence, without specifying and developing these questions into measures of the three dimensions. Such an approach might be necessary because of data limitations (for example, limited space on surveys) or because other study concepts are more important. Although less thorough than the first approach, the following question might be considered a measure of high school violence:

> Please indicate whether you strongly agree, agree, don't know, disagree, or strongly disagree with each of the following statements: . . .

At least one of my classmates has carried a gun to school.
Some students in my class regularly carry knives to school.
Students in my class regularly get involved in fights and scuffles.
There is inappropriate sexual contact or gesturing occurring in my class.
People try to hurt others in my class through tripping, pushing, or shoving.
People in my class threaten each other with physical violence.
People in my class vandalize each other's property.
People in my class regularly insult each other.

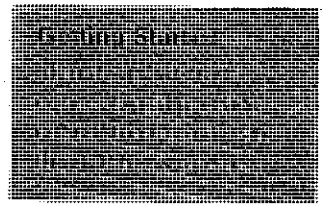
Note that the list encompasses the three dimensions identified earlier. Whether this measure suffices depends on the manager's needs for more specific information and on validation (discussed later in this chapter). Sometimes this second approach develops into the first approach as analysts give more careful consideration to the distinct dimensions of the concept.

The *third* approach is decidedly nonrigorous, using a single survey item to measure the concept. For our current example, such an item might read as follows:

Please indicate whether you strongly agree, agree, don't know, disagree, or strongly disagree with the following statement:

My high school is a violent place.

While not biased, this item does not provide any information about specific aspects of the phenomenon. As noted earlier, this approach is typically used when the concept is of quite minor importance to the program or evaluation. In our example, however, we want more information than would be obtained from this single item.



Finally, an important question is whether any best set of measures exists for measuring a concept. The *theorem of the interchangeability of indicators* states that if several measures are equally valid indicators of a concept, then any subset of these measures will be valid as well. In other words, there are many valid ways to measure a given concept. The analyst's task is to choose one approach and then justify that that approach is valid. The challenge of justification is discussed later in this chapter.

INDEX VARIABLES

An ***index variable*** is a variable that combines the values of other variables into a single indictor or score. For example, the consumer price index is a

Table 3.2 Creating an Index Variable

Observation	Measure 1	Measure 2	Measure 3	Measure 4	Index
567	1	2	2	4	9
568	4	1	1	1	7
569	4	2	2	4	12
570	5	5	5	5	20
571	1	2	—	1	—
572	1	1	1	1	4

variable that combines the prices of common consumer goods and services into a single score. Index variables are common, for example, measuring the economic outlook, infant and child health, environmental quality, political stability, volunteerism and giving, culture in cities, and so on. Managers and analysts frequently encounter index variables in their work.

Index variables are also commonly used to empirically measure abstract concepts and multifaceted, encompassing phenomena. In the preceding sections, we developed a strategy for measuring different dimensions of high school violence. Some variables measure violence that involves weapons, other variables measure inappropriate physical contact, and still other variables measure verbal assaults and threats. How can these disparate measures be combined into one aggregate measure of high school violence?

The logic of index variable construction is simple: the values of the measurement variables are simply summed. The term *measurement variable* refers to the (observed) variables that make up the index; it has no bearing on any measurement scale or data collection strategy and is used to distinguish these variables from the index variable. When respondents score low on measurement variables, the resulting index score is also low, and vice versa. Table 3.2 shows how an index variable is created by simply adding up the values of the measurement variables that constitute the dimension or concept. Thus, when respondents score high on measurement variables, the resulting index score is high. When one or more of the measurement variables are missing from an observation, the value of the index variables for that observation is missing, too, as shown for observation 571. Note that whereas measurement variables might be ordinal (for example, measured on a five-point Likert scale), the resulting index variable often is continuous. In the example in Table 3.2, the index variable can range from a minimum of 4 to a maximum of 20. Of course, statistical software does the addition.

This logic is applied to other indexes, too. For example, the consumer price index is based on price changes for a bundle of goods. The sum of all prices is determined in each period, and the periods are then compared with each other. An index of municipal cultural activity might sum the number

of performances, renowned organizations, and cultural facilities (museums, theaters, and the like).

To continue our example of high school violence, in our second approach we simply sum the values of each of the survey items and in this way construct an index measure of high school violence. The values of these items are summed for each observation, in exactly the same manner as in Table 3.2. But in our first, more rigorous approach, we follow a two-step process for creating the index measure. In the first step, we construct index variables for each of the three separate dimensions (use of weapons, inappropriate physical contact, and verbal assaults and threats). In the second step, we sum the values of these three index variables, for each observation, which then results in a new, "super" index of "high school violence." The latter index is clearly grounded in the three dimensions of high school violence.

A practical problem with index variables is that individual components sometimes have different scales or ranges. For example, if one variable can range from 0 (min) to 10 (max), and the other from 0 (min) to 1,000 (max), then the former will likely not have much impact on the aggregate measure, the index. Especially if most values of the latter variable are between, say, 300 and 800, then adding the values of the first variable, ranging between 0 and 10, will not much affect the aggregate score. To address this problem, analysts can rescale each of the variables being summed, so that each has the same range, such as 0 to 100. One way to do this, in the preceding case, would be to multiply each value of the first variable by 10 and divide each value of the second variable by 10. However, other approaches exist.[9]

Although index measures are not very difficult to make, a key issue is their validation. The resulting index variable must be established as a valid measure of the underlying concept being measured. The next section discusses how we go about doing that.

MEASUREMENT VALIDITY

It is always important to think about the validity of what we do. Earlier, in Chapter 2, we discussed validity with regard to drawing study conclusions; here, we discuss it narrowly with regard to measurement. Measures must be shown to be valid measures of the phenomena and concepts that they measure. Measurement validity simply means that variables really measure what they are said to measure. Considerable thought has gone into the different strategies that can be used to establish measurement validity. Analysts are not expected to use all or even most of these strategies, but they are expected to justify their variables in some way.[10]

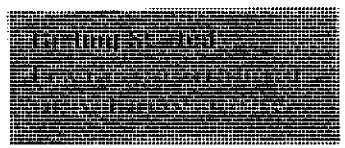
An important form of validation is theoretical—a persuasive argument that the measures make sense. One argument is that the measures are

reasonable, common-sense ways of measuring the underlying concept. This is called ***face validity***. Measuring gender by asking respondents whether they are male or female is a reasonable, common-sense method. Some respondents may erroneously indicate the wrong gender, but such numbers will likely be few and not affect study conclusions in any material way. In the case of high school violence, however, the justification is more elaborate. But again we can argue that the measures used are reasonable, common-sense ways of measuring the specific variables and underlying concept.

Regarding index variables, another argument is that they should encompass a broad range of aspects. For example, variables measuring "physical exercise" should not be skewed in some biasing way, perhaps underemphasizing individual sports in favor of team sports. Whether "student anger" is measured in a comprehensive or simple way, it should not be biased against certain forms of student anger. This form of validity is called ***content validity***. The very simple operationalization given earlier ("how angry do you usually feel?") avoids this problem by not specifying any specific form of anger. In the case of high school violence, we measure a broad range of aspects, especially those that ought to be included in such a study.

Empirical evidence can also be mustered in several ways. First, variables can be validated by comparing them with other measures or sources. For example, the measure "physical contact without weapons" might be triangulated by records of the school nurse (treatment of scrapes and bruises) and a student survey. Although such correlation does not prove that the measure is valid, certainly the lack of correlation would raise some eyebrows. Comparison with external sources is sometimes called ***criterion (or external) validity*** (not be confused with threats to external validity, discussed in Chapter 2). Some researchers also refer to this as *triangulation*. When the variable correlates as expected, additional validity is provided.

Second, we might ask respondents on the same survey about physical contact without weapons and compare that response with other responses, such as regarding physical injuries incurred at school. Such comparison against internal sources is called ***construct (or internal) validity***. Although this comparison does not provide absolute proof (respondents may receive injuries at school for reasons unrelated to high school violence), it may provide some reassurance and, hence, a measure of validity. Certainly a lack of correlation would require further inquiry and explanation.



Third, regarding index variables, the variables used to measure a concept should be strongly associated (or correlated) with each other. This is because each index variable measures different but related dimensions. When variables are not highly related, analysts should consider whether, perhaps,

one or more of the variables measure some other concept. The correlation of measurement variables is called ***internal reliability*** (or internal consistency, not be confused with "threats to internal validity," discussed in Chapter 2). ***Cronbach alpha*** (also called *alpha* or *measure alpha*) is a statistical measure of internal reliability that is often cited in research articles that use index variables.[11] Although you need not be concerned about the exact calculation of this measure,[12] alpha can range from 0 to 1, where a 1 indicates perfect correlation among the measurement variables, and a 0 indicates the lack of any correlation among the measurement variables. Values between 0.80 and 1.00 are desired, and they indicate high reliability among the measurement variables. Values between 0.70 and 0.80 indicate moderate (but acceptable) reliability. Alpha values below 0.70 are poor and should cause analysts to consider a different mix of variables. While index variables with alpha scores below 0.70 should be avoided, values between 0.60 and 0.70 are sometimes used when analysts lack a better mix of variables. Analysts usually collect a few more variables than are minimally needed because they cannot know, prior to reliability analysis, which variable mix will have a sufficiently high alpha score to lend empirical support for the index measure. This is especially relevant for one-dimensional measures of complex concepts, such as the less rigorous measure of high school violence discussed earlier in this chapter.

Finally, descriptive analysis is used to examine the range of values of (index) variables. If most values of a variable are "high," then little will be known about those who score "low." Being mindful of this problem helps analysts avoid inappropriate generalizations to categories (for example, subpopulations) about which little empirical information has been collected. For example, if most of our respondents indicate that high school violence is a serious problem, then little will be learned about factors associated with high school violence among those who perceive it to be low, including, quite possibly, strategies causing some schools to have low levels of high school violence. Descriptive analysis is also used to examine whether observations with missing values in their index variables create a pattern of bias, perhaps systematically excluding some group or groups of observations, for example, such as minorities or pregnant teenagers for whom some items may have been irrelevant or in some way troublesome.

In sum, a plethora of strategies exists for assessing measurement validity. Analysts are not expected to use all of these approaches, but they should use some strategies to justify their measures. In scientific research, this usually requires some up-front consideration because, after data have been collected, it may be too late to collect more observations as needed for validation.

An obvious and final question is this: what is an analyst to do if one or more of the strategies described in this chapter show variables to be less valid than hoped for? Perhaps the measures of internal and external validity provide mixed results, and the alpha measure is marginal at best. If this happens, the analyst needs to add a caveat to his or her results. However, with foresight and planning, analysts usually gather a broad range of variables so that adequate supporting evidence from face and construct validity are available.

SUMMARY

The four measurement levels of variables are nominal, ordinal, interval, and ratio. A general guideline is that measurement scales are preferred that give as much information as possible about variables. Nominal-level scales exhibit no order among attributes, ordinal-level scales exhibit order but no distance between attributes, and interval- and ratio-level scales exhibit both order and distance. Variables with interval- and ratio-level scales are sometimes called continuous variables, and variables with nominal- and ordinal-level scales are called categorical or discrete variables. A variable's measurement level is also important in the selection of statistical tests, discussed in later chapters. Likert scales are commonly used ordinal-level variables in surveys. There are many different types of Likert scales, assessing degrees of importance, satisfaction, agreement, and frequency, for example. Index variables sum the values of disparate variables and are used to measure concepts.

Rigor in measurement increases the validity of analytical work. When working with abstract concepts, analysts need to carefully identify the different dimensions of their concepts and then develop appropriate ways to measure each. Measures used by other studies can help guide analysts in this task, but they often must develop and validate their own measures.

Measures should be valid, and this chapter offers strategies for determining measurement validity. Four types of validity are face validity, content validity, criterion validity, and construct validity. Additionally, Cronbach alpha is used for index variables as a measure of their internal reliability. Analysts should examine their measures for validity and provide caveats to their results as necessary.

KEY TERMS

Categorical variables (p. 42)
Concepts (p. 46)
Conceptualization (p. 46)
Construct validity (p. 53)
Content validity (p. 53)
Continuous variables (p. 44)
Criterion validity (p. 53)
Cronbach alpha (p. 54)
Discrete variables (p. 42)
Face validity (p. 53)
Index variable (p. 50)
Internal reliability (p. 54)
Interval-level scales (p. 44)
Likert scales (p. 42)

Notes

1. Other types of ordinal-level scales exist, too, but they are much less common. For example, *Guttman scales* are based on a series of statements with increasing or decreasing intensity, for example, "I feel safe around my classmates," "I avoid classmates who are violent," and " I bring a knife to school to defend myself against my classmates." The scale assumes a consistent pattern in answering these statements. That is, those who agree with the last statement are unlikely to agree with the first statement too. A statistical coefficient is calculated that measures the extent to which such a consistent pattern exists. Guttman scales have become less popular in recent years, due to their rigidity and complexity. *Thurstone scales* use judges to assess and order a large number of such statements, from which a scale is then composed. The cumbersomeness of using panels also makes Thurstone scales unpopular.

   Somewhat more common are *semantic differential scales*, especially in psychological studies. These scales assume that people think in opposing pairs as they assess situations, such as "How do you feel about anger management classes as a method for reducing high school violence?" Respondents are asked to indicate a point on each line that indicates their feeling:

Good	----------	Bad
Smart	----------	Dumb
Respectful	----------	Disrespectful

2. Some texts refer to both interval and ratio scales as interval scales, which may cause confusion. Other texts refer to both as metric scales and often refer to nominal and ordinal variables as nonmetric variables. In this context, the term *metric* has no bearing on the metric system of measurement. We avoid using the terms *metric* and *nonmetric* here, to prevent any such confusion.
3. The following question is sometimes raised: how many categories must an ordinal-level variable have in order to be considered an interval variable? This question misses the point that the key theoretical distinction between ordinal and continuous variables is whether the distances between categories can be determined. Even so, in practice ordinal-level variables with seven or more categories are sometimes analyzed with statistics that are appropriate only for interval-level variables. This practice has many critics, but it is done, because interval-level statistics more readily address control variables and also because ordinal-level statistics sometimes don't work well with large tables. Nonetheless, the practice is controversial and it is best to analyze ordinal variables with statistics that are appropriate for ordinal-level variables, discussed later.
4. Measurement validity is also discussed at the end of this chapter.
5. This can be regarded as an example of asking questions of basic research—see Section II introduction.
6. Many scientific studies in public administration and public policy use one to five dimensions per concept (thus some concepts have only one dimension), and operationalization is often limited to five to eight variables per dimension. A practical consideration is that, when working with existing data (also called secondary data), analysts often must use whatever variables are available. Conceptualization and operationalization may then be wanting, to say the least. Analysts must acknowledge study limitations (caveats) and argue that the analysis adds value and is the best available.
7. In addition to the strategies discussed here, empirical approaches such as *factor analysis* can be used to justify the number of dimensions (see Chapter 16).
8. Some researchers also use this term to include procedures for data collection, but these procedures are discussed in Chapter 5.
9. Variability is another factor. A better way of dealing with this factor is through standardization, a process discussed in Chapter 6.
10. Recall the other threats to internal and external validity discussed in Chapter 2. Some problems of validity deal with sample bias, such as a biased selection of administrative records or survey respondents. Other problems deal with testing and instrumentation, such as biased or leading questions on survey questionnaires. Guidelines for dealing with these problems are discussed in Chapter 5.
11. You can find such articles in public administration, for example, by Googling "cronbach" "public administration".
12. See note 8 in Chapter 12.