Chapter 7

Measurement

Measurement is the foundation of scientific inquiry. In order to test our hypotheses, we must observe our theoretical concepts at the operational level. In simple words, we must measure what we have defined. But there are different levels of measurement, which provide differing amounts of information about the theoretical construct. There are also some basic issues about the adequacy of measurement which we must address.

Levels Of Measurement

Depending on our operational definition, a measurement can give us differing kinds of information about a theoretical concept. However, at the minimal level a measure must provide the ability to detect the presence or absence of the theoretical construct. All levels of measurement give this ability, or they wouldn't be measurements at all. If the presence or absence of a theoretical construct is the only information that the measurement provides, we call the measure nominal. A second level of measurement adds the idea of quantity, or an underlying dimension, to the measure's ability to detect. At this level, we can not only detect the presence of the theoretical construct, we can also make comparative statements about its quantity, like "more of…" or "higher than…." If the measurement contains only detection and comparative ordering information, we call it ordinal. At the next higher level, a measurement adds the idea of units, so that we can make absolute (rather than simple comparative) statements about the similarity or difference between measurements. That is, we can state the number of units by which observations are measured to be different. This level of measurement is called interval. Finally, if the measure is interval, but also contains an absolute zero category or scale point, we can make statements of proportion ("only one half of") or ratios ("twice

as much…") about the magnitude of our measurements. We call this highest level of measurement ratio-level.

Nominal Measurement

A nominal measure makes only a single, simple distinction: between the presence or absence of the theoretical concept within the unit of analysis. It's a simple black-or-white kind of view that we will use to categorize the observed units. For instance, we might operationally define the theoretical concept "Gender of Participant" by classifying all participants in an experiment as either Male or Female.

Let us carefully analyze what we actually do when we assign these labels. We choose to ignore all gradations of "masculinity" or "femininity". We will merely characterize every subject as having "present" or "absent" the characteristic "Maleness", OR as having "present" or "absent" the characteristic "Femaleness". What we actually do then is to measure the participant on one of two nominal variables - Maleness or Femaleness. We only need to rate the participant as "absent" or "present" on one of the two, because "present" on the characteristic Maleness, for instance, IMPLIES "absence" on the characteristic "Femaleness", and vice versa. Because "absence" on Maleness implies "presence" on Femaleness, the categories in a nominal measure (or any other level of measurement, for that matter) are called mutually exclusive. This means that it must not be possible for any single unit of analysis to be a member of more than one category. Furthermore, the categories in any variable at any level of measurement must be exhaustive: every unit of analysis we encounter must be able to be assigned to one of the nominal categories. Mutual exclusivity and exhaustiveness therefore constitute the minimal requirements for measurement, whether measurement be nominal, ordinal, interval or ratio. Observe also that there is no ordering of the two categories—Female is not bigger or smaller than Male, and Male is not greater than or less than Female—they are simply mutually exclusive.

"Gender of Participant" is an example of the simplest nominal measurement, called dichotomous or binary measurement. In this kind of operational definition, the variable may take on only one of two possible values.

Of course, theoretical concepts can have more than two nominal response categories. If so, the construct is properly called a "nominal factor", as it really consists of a number of simple nominal variables. Again, an example is probably the best way to explain this. Suppose we are conducting a political communication study, and we want to determine the political affiliation of each respondent. We can define the theoretical concept "Political Party Affiliation" as a nominal factor with the response categories of Democrat, Republican, and Independent. These three categories actually require that we characterize each respondent as "absent" or "present" on two nominal variables to correctly represent each person's political affiliation. We'll call the nominal variables "Democratic Party Membership" and "Republican Party Membership". A respondent's party affiliation is then described as the particular combination of presence or absence of each of these two variables, as is shown in Table 7-1.

If a person is scored as "absent" on the nominal variable "Democratic Party Membership" AND also as "absent" on the variable "Republican Party Membership", it is implied that the person is an "Independent". "Present" on either one of these implies membership in that party. Again

Table 7-1 The Nominal Components of the Factor "Political Party Affiliation"

Nominal Category	*Nominal Variables*	
	Democratic Party Membership	*Republican Party Membership*
Democratic	Present	Absent
Republican	Absent	Present
Independent	Absent	Absent

notice how this measurement scheme is mutually exclusive. Furthermore, it is exhaustive as all observations that are not assigned to Democrat or Republican will logically be assigned to the category Independent. In addition, again there is no underlying order to these three categories of Party Affiliation.

We can extend this idea to theoretical concepts with any number of nominal categories. We will always need G - 1 nominal variables to represent a nominal factor with G categories.

In general, we recommend that nominal measurements be avoided, as they provide the least amount of information about the theoretical concept, and they can be quite cumbersome to analyze if you define a nominal factor with a large number of categories. But sometimes they are the only realistic possibility. It is very easy to measure the "gender of subject" by categorizing each subject as either male or female; it is much harder to determine the degree of masculinity or femininity. We may not have the time to administer a complicated psychological gender scale to each participant in our research, so we may decide to settle for the nominal male-female distinction. Other times it may appear that we do not have any choice but to use nominal variables such as Party Affiliation. Later on in this chapter we will provide an example of how we may well be able to replace such nominal variables with variables at higher levels of measurement and obtain more information about the concept we are trying to measure.

Ordinal Measurement

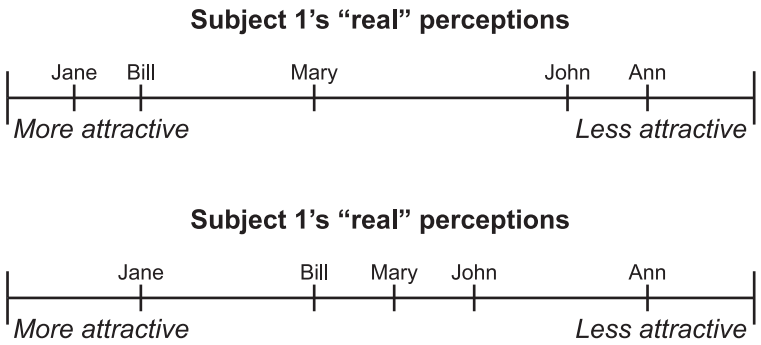
As we saw above, the categories in a nominal variable cannot be arranged in any order of magnitude. But if we add this idea of ordering by quantity to the definition of the categories, we can improve the sensitivity of our observations.

Let's look at another simple example. Suppose we wish to measure the theoretical concept "age" using a dichotomized variable. Every unit in our sample will be categorized as either OLD or YOUNG. This measurement of age is an example of an ordinal-level measurement. In addition to viewing the OLD and the YOUNG categories as mutually exclusive and exhaustive, we can also think of the YOUNG category as lesser in age than the OLD category. Furthermore, if we want to add the category MIDDLE-AGED, we can place it in between YOUNG and OLD with some justification. Contrast this with the situation of adding a new category called SOCIALIST to the set of categories in the "Political Party" concept discussed in the previous section. The new category can be slotted anywhere in the set of already existing categories, which indicates that "Political Party" is truly a nominal factor, with no inherent quantities which can be used to arrange the order of the categories. For the category MIDDLE AGED there is only one position that makes sense: right between YOUNG and OLD.

Ordinal measurements allow us to make comparative distinctions between observations along some dimension. For instance, suppose we ask participants in an interpersonal communication experiment to rank the physical attractiveness of a number of conversational partners by sorting a stack of photographs so that the most attractive partner is on top and the least attractive is on the bottom. We can now say that second photograph in the pile is more attractive to the subject than all the photos in the pile below it, but less attractive than the photo on the top of the pile. We can assign an "attractiveness" score to each photograph by numbering it, starting at the top of the pile (1=most attractive, 2=second most attractive, etc.). This is called a rank order measurement. This measurement takes on comparative degrees of difference, and this distinguishes an ordinal measure from a nominal one, in which we have only a single distinction of difference (in nominal measurement, an observation is the same as others in its category, and different from all other categories).

This feature allows us to introduce the general idea of comparative similarity in observations. In the photo ranking example, we can conclude that adjacent photographs in the sorted pile are similar to each other in attractiveness, and that photos near the top of the pile are very different in attractiveness from those at the bottom of the pile. This distinction gives us more information about attractiveness than a nominal measurement scheme. This information can be used to great advantage during statistical tests which establish relationships between ordinal variables.

But one thing which we cannot do with ordinal data is determine the absolute distance between adjacent categories. For example, suppose we knew the "real" attractiveness score of each photograph for two subjects. In Figure 7-1, we've placed the photos on the "real" attractiveness scale according to each subject's true evaluation. Although both subjects "real" evaluation of their conversational partners are quite different, they will rank the partners' comparative attractiveness



Subject 1's Ordinal Ranking:

1 - Jane
2 - Bill
3 - Mary
4 - John
5 - Ann

Subject 2's Ordinal Ranking:

1 - Jane
2 - Bill
3 - Mary
4 - John
5 - Ann

FIGURE 7-1 Linear operational linkage between two variables: Income and newspaper readership

identically.

An ordinal measurement will give us only relatively fuzzy comparative distinctions among observations. To get truly sensitive measurement, we must add absolute measurement units to the operational definition. These units add more information to the measurement, in the form of magnitudes between adjacent categories. If we can do this, we are measuring at the interval level.

Interval Measurement

When we can not only rank order observations, but can also assign them numerical scores which register the degree of distance between observations or points on the measurement scale, we have improved the level of measurement to interval. In interval measurement, equal numerical distances imply equal dissimilarity.

In the example shown in Figure 7-2, for Subject #1, Jane is one unit more attractive than Bill, and John is one unit more attractive than Ann. We conclude that the difference between Jane and Bill's attractiveness is identical to the difference between John and Ann's, as the interval between each of the pairs is identical: one unit. Furthermore, we see that Mary is equidistant between Jane and Ann for Respondent #2, and that the difference in attractiveness between Ann and Jane is twice as large as the difference between Mary and Jane. And we can compare Jane's attractiveness to each of the subjects, and see that Subject #1 found her more attractive (gave her 12 units) than did Subject #2 (who only gave her 10 units), although both ranked her as the most attractive partner.

Once again, the additional information provided by interval measurement will make it easier for us to detect relationships between variables. But there are still some statements which we cannot make with an interval measurement. We cannot say, for example, that Mary (with 6 attractiveness units) is three times more attractive than Ann (who received 2) for Subject #2. To make comparative statements based on ratios, we must move to yet another higher level of measurement.

Ratio Measurement

If the measurement classes include an absolute zero point, corresponding to the complete absence of the theoretical concept, we have the highest level of measurement: measurement at the ratio level. A survey respondent's age, expressed in "number of years since birth", is a good example of a ratio measurement. The classes of that variable contain a zero (if this year is the year of birth), and we can make statements like "Tom is twice as old as Vanessa" if the ratio of their ages is 2:1, or

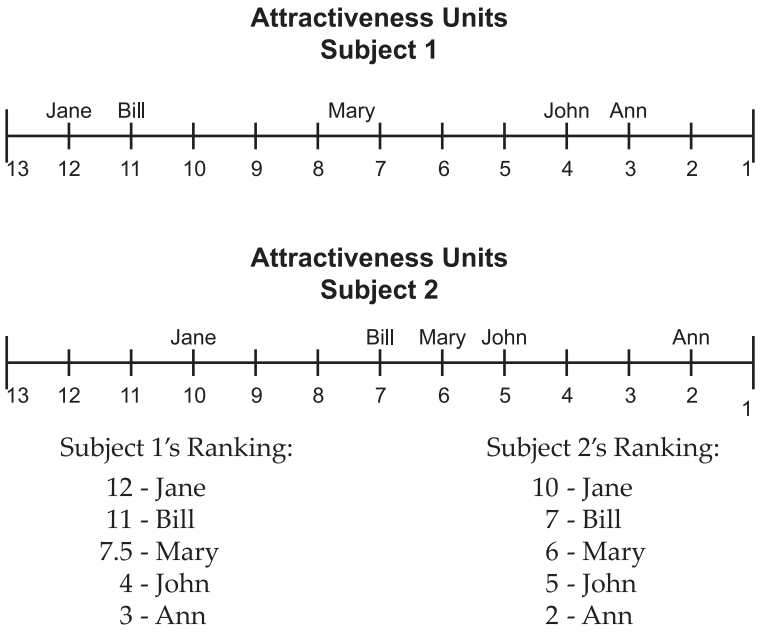


FIGURE 7-2 Interval Measurements

2.0. It is important to point out that interval level measurement classes may also contain a zero. In this case the zero is an "arbitrary" zero; it does not denote the absence of whatever characteristic is being observed. For instance, the Fahrenheit scale for measuring temperature has a zero-point which does not, however, indicate the "absence" of heat. So if today's temperature is 60 degrees and yesterday's was 30, we cannot say that it is twice as warm today as it was yesterday. Ratio-level measurements are the most desirable, as they contain the most information about each observation of the theoretical construct.

In the behavioral sciences, it is customary to treat interval-level measurements as if they are ratio-level. Normally, this does not distort observations, as long as the theoretical concepts include the implication of an absolute zero point, even if it is not part of the actual operational definition. The attractiveness scale used in the example above does not have an actual zero point on the scale, but we can visualize a zero point for attractiveness (no attractiveness whatsoever). Being able to treat variables as ratio variables gives us access to many extremely useful statistics (which we will discuss later) that assume ratio level measurement in their computations.

Choosing the Level of Measurement

This choice is easy: always construct your operational definition at the highest level of measurement which is practically possible. This will give the maximum sensitivity in your measurement, and you can use more powerful statistics to test your hypotheses.

The first thing to keep in mind when you construct an operational definition is this: use a measurement definition that is interval or ratio, if at all possible. It is often tempting to "simplify" measurement of a theoretical concept by dropping down a level of measurement or two. A simple example will illustrate the dangers in succumbing to this temptation. Let's look at several ways of measuring the construct of "age".

The best way is to simply request that the experimental subject or survey respondent report the number of years since their birth. This is a ratio-level measurement. It preserves very small differences between observations, and permits both comparative and quantitative comparisons of observations or groups of observations. But we might ask the "age" question in another way. Suppose we ask the respondent to categorize his or her age in this way:

What is your age?

a) under 18

b) 18-21
c) 22-29
d) 30-39
e) 40-64
f) 65 or above

This is a common method of requesting age information. But it has reduced the level of measurement from ratio to ordinal. The category boundaries are not equal in range, so we can't infer the magnitude of differences in age by category membership. Suppose two respondents differ by one category. How different are their ages? If the categories are (b) and (c), the average age difference will be the difference between the midpoints of each category, or 25.5 - 19.5 = 6 years. But if the categories are (d) and (e), the average difference will be 52 - 34.5 = 17.5 years. This varying difference between category boundaries makes quantitative comparisons useless. We can certainly say that the person in the higher age category is older than the person in the lower category, but we can't say how much older.

The difference in size of the categories also means that there may be more variation in age within a category (all of whose members are considered identical in age) than there is between categories. Persons could differ by up to 24 years in category (e), while they could differ by only 3 years in category (b). Having wide boundaries for the categories also means that we've lost sensitivity in expressing age differences—a 40-year-old respondent will be considered identical in age to a 64-year-old. Categorizing age this way reduces the precision of measurement and thereby reduces our ability to detect relationships between our theoretical concepts. And finding relationships is the basic reason that we're conducting research.

We can do even more damage to our ability to accurately observe the theoretical concept by the following. Suppose we again "simplify" the measurement of age to the following question:

Are you a(n):
a) Child
b) Adolescent
c) Adult

Now, not only are the categories different in size, as they were in the previous example, but we have also added fuzzy boundaries between the categories. We would consider this measurement to still be ordinal-level, since it is possible to order the categories along a single dimension. But the ordering is weak, as the boundary between adolescent and adult, for example, is not clear. One 18-year-old might consider herself an adult, while another 18-year-old might classify himself as an adolescent.

Why would we ever use this form of measurement? As mentioned above, the primary reason is for convenience. But before using this kind of measurement, we should make an attempt to move up to interval or ratio level measurement if at all possible.

One way to do this is to consider the real nature of the theoretical concepts we're measuring, and see if we're really expressing the central ideas of our investigation. Often there is a tendency to use simple nominal categories that describe the surface features of concept, rather than the more specific dimensions that truly matter. Whenever possible we must specify the theoretical constructs in a way that will allow us to use the highest level of measurement.

Let's look at an example. Suppose that we initially think about whether the political affiliation of survey respondents is associated with their readership of particular magazines. The almost reflexively obvious way to do this is to ask each respondent these questions:

To which political party do you belong?
a) Democratic
b) Republican
c) Neither

and

Which one of these magazines do you regularly read?
a) *Time*
b) *Newsweek*

c) *U.S. News and World Report*
d) *Commentary*
e) *New Republic*
f) None of the above

Both theoretical concepts (party affiliation and magazine readership) are measured at the nominal level, so we receive the minimum amount of information about each. For example, we know that Democrats can range from very conservative to very liberal, and that the political beliefs of some Republicans are almost identical to those of some Democrats, but none of this information will be captured. Furthermore, we can't be sure how a respondent will define a magazine that is "regularly read". One respondent might consider glancing at the headlines each week as regular reading, while another might decide that only magazines which are read cover-to-cover qualify. Such a difference will result in substantial amounts of error in the reporting of actual exposure to the content of these magazines.

The problem can be solved by considering the details of the theoretical process that we are investigating. As we consider these, perhaps we will realize that what we are really trying to understand is something more specific than general party affiliation and simple magazine readership. We might conclude that we really want to investigate the relationship between political liberalism or conservatism and exposure to political news and commentary. We can then replace the nominal political party categories with a set of questions about political views. If these questions are expressed as scales, as described in the next section, we then have measurement at ordinal or interval level to replace the nominal political party categories. Similarly, we might replace the nominal checklist of magazines with a measurement question like, "How many hours in an average week do you devote to reading political news and commentary?" This is a ratio-level measurement which is much more sensitive than the nominal categories. As a result, we can use more sensitive statistics which require interval-level measurement, rather than simpler nominal level statistics.

In the process of defining higher-level measurement, we have also honed our theoretical thinking. It is often true that improvements in defining theoretical concepts leads to better measurements, as well. As we saw in Chapter 2, there is a strong interplay between the "verbal world" of theoretical concepts and the "measurement world" of operational definitions. Insightful thinking about theoretical or operational definitions will probably result in improvements to both.

Scaling

Scaling is a term used to describe the way that an operational definition can be conceptualized to provide numerical measurement. Usually the term is applied only to ordinal or interval level measures, as nominal scaling is really just a matter of classification within a set of categories, as we saw above. There are a vast number of different scaling techniques and procedures, and scaling represents a whole area of study by itself. Here, we'll just outline some of the more common types of scaling.

Counting frequencies

Perhaps the simplest scaling involves natural measures like the counting of instances of occurrence of events. Such occurrence is absolute in nature and can be measured in terms of its "frequency". Scales reflecting measures of frequency are at the ratio level of measurement, and thus are very desirable. Typical operational definitions might involve counting the number of different cartoons presented by a network on Saturday morning; counting the number of times (the frequency) that an employee communicates with his boss in a week; measuring the frequency with which stories on a particular topic appear in a newspaper; counting the number of retrieval requests for a particular document in an electronic database.

We can also use established metrics such as temperature scales and electrical units like volts and ohms (these are often useful in physiological measurements), measurements of sound amplitude in the decibel scale, distances in miles or meters, measures of time, such as hours and minutes, and so forth. These are units of measurement that are arbitrary in nature, rather than absolute. When we count the number of "inquiries" to a database, the unit of measurement is an "inquiry" - we count the number that occur. However, when the unit of measurement is arbitrary we need to

establish first what the unit of measurement is going to be. To measure distance, for instance, we need to agree upon some arbitrary way of expressing that distance: in terms of meters and yards, kilometers and miles. These units are not absolute, but their definitions have been established over a long time period, are widely accepted and they are standardized so that reliable measurement is possible. They are usually ratio scales, too. Although these metrics are attached to theoretical concepts of their own (like "displacement in space" for distance metrics), they can be used as partial indicators of more abstract communication concepts. For example, distance may be used to measure (at least partially) the degree of difficulty in obtaining information from a library, as in this question:

"How many miles do you have to drive or walk to obtain a book from your nearest public library?"

_____

The skin's resistance to electrical currents can be measured in ohms; as a consequence, arousal due to, for instance, exposure to erotic video programs, is often measured in terms of changes in skin resistance as measured in ohm units. In both of these examples we'd measure the amount of the arbitrarily defined units.

Measuring Magnitude

There are also scaling procedures which are associated more specifically with behavioral research. Perhaps the most common of these are the magnitude types of scales, of which the Likert scale is a typical example. In this measurement procedure, verbal "anchors", which define the extremes of the dimension being measured, are provided to allow a range of responses to some specific question. The experimental subject or respondent is then provided with a statement and is asked to choose some point on the scale which represents his or her judgment of magnitude. Figure 7.3 contains some examples of different magnitude scales:

In these three examples "Infrequently" and "Frequently", "Strongly Agree" and "Strongly Disagree", "George Washington" and "Adolf Hitler" represent the extreme "poles" or "anchors" of the underlying dimensions of, respectively, frequency of communication, agreement, and democratic leadership.

If the scale anchors are a set of adjectives which are antonyms (adjectives which are logical opposites), the resulting set of scales, such as shown in Figure 7-4, is sometimes called a semantic differential scale:

It is a mistake to assume that the measurement obtained from magnitude scales such as the ones above is at the interval or ratio level because we have no way of determining that the distances between adjacent scale points are really equal. In fact, there is considerable evidence that the "psychological distance" between scale points in the middle of magnitude scales is smaller than it is near the end points. There is a general reluctance on the part of respondents to use the extreme ends of magnitude scales. While the difference between the adjacent scale points representing "neutral" and "mildly agree" (scale points 5 and 6 on a 9-point scale) might be seen as slight, the

"I communicate with my mother…"

1 2 3 4 5 6 7

Infrequently
Frequently

"There is too much violence on television"

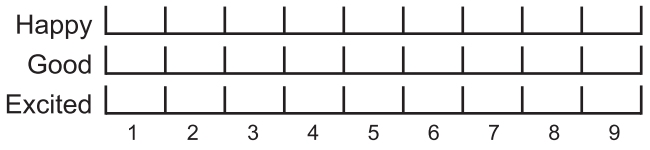
1. Strongly agree
2. Agree
3. Neither agree nor disagree
4. Disagree
5. Strongly disagree

"If George Washington represents 100 and Adolf Hitler represents 0 on a scale of democratic leadership, where does our current president fall?"

_____

FIGURE 7.3 Examples of Magnitude Scales

distance between "agree" and "strongly agree" (scale points 8 and 9) is perceived as much greater, even though these are also adjacent scale points.



If magnitude scales are analyzed with ordinal statistics like those described later in this book, this lack of interval distances makes no difference. But magnitude scales are frequently selected to measure theoretical concepts which require analysis with statistics that assume at least interval data. Research on magnitude scales has shown that the assumption of interval-level measurement does not give seriously incorrect results in most cases. To handle the more severe non-interval problem at the end points of the scales, some researchers add extra scale points to either end of the scale, and then collapse the two end points into a single value. For example, if a researcher wished to minimize the problem in a scale with 7 points, she would use a 9-point scale, then consider responses 1 and 2 as identical, and 8 and 9 as identical. The result is a 7-point scale (going from 2 to 8) which is closer to interval-level.

There are scaling methods which directly address the problem of constructing scales with equal intervals between adjacent points. Thurstone scaling is a procedure in which the scale consists of a series of questions to which the subject responds with "yes-no" or "agree-disagree" answers. The questions are chosen so that they represent a set of intervals that appear similar in magnitude to respondents. A set of Thurstone scale questions to measure the degree of "Reliance on Informal Communication" in an organization might look like the ones depicted in Exhibit 7-1:

If such a scale is well-constructed, a respondent's position on the dimension being measured

Exhibit 7-1 Example of Thurstone Scaling

1. "Most of what I know about what happens in management comes from rumors."

[Agree] [Disagree]

(This question represents scale point 3, representing high reliance on informal communication.)

2. "Information about our company's long-term decisions is just as likely to be passed down in conversation with friends in management as it is to show up in the company newsletter."

[Agree] [Disagree]

(This question represents scale point 2, reliance on both formal and informal communication.)

3. "We have a very efficient system of meetings, newsletters, and briefings which management uses to keep us fully informed about the business decisions that they make."

[Agree] [Disagree]

(This question represents scale point 2, reliance on both formal and informal communication.)

can be determined by the scale value of the question at which the respondent switches from agreement to disagreement.

The questions used on a Thurstone scale are selected from a much larger set of potential questions by a rather laborious procedure which is too detailed to discuss here. (See the References and Additional Readings section for some readings which detail scaling procedures). For much communication research, the improvement in measurement represented by equal-appearing interval scales does not justify the extra effort, so Thurstone scales are seen only rarely.

Guttman or cumulative scales use a similar format in which the scale consists of a series of questions. But Guttman scaling also provides a way of determining if the scale is unidimensional,

Exhibit 7-2 Example of Guttman Scaling

A: Consistency in Scoring

"I think the following contains pornographic material:"	Subject A	Subject B	Subject C	Scale Value
Adult movies rated XXX	[Yes]	[Yes]	[Yes]	4
Playboy magazine	[Yes]	[Yes]	[No]	3
Lingerie advertisements	[Yes]	[No]	[No]	2
New York Times	[No]	[No]	[No]	1

B: Inconsistencies in Scoring

"I think the following contains pornographic material:"	Subject A	Subject B	Subject C	Scale Value
Adult movies rated XXX	[Yes]	[Yes]	[Yes]	4
Playboy magazine	[Yes]	[No]	[Yes]	3
Lingerie advertisements	[Yes]	[Yes]	[No]	2
New York Times	[No]	[No]	[Yes]	1

C: Coding Inconsistencies

"I think the following contains pornographic material:"	Subject A	Subject B	Subject C	Scale Value
Adult movies rated XXX	[+]	[+]	[+]	4
Playboy magazine	[+]	[+]	[+]	3
Lingerie advertisements	[+]	[-]	[+]	2
New York Times	[+]	[+]	[-]	1

that is, if it is measuring only a single theoretical concept. If the statements are ordered in magnitude, we should see a consistent pattern of responses in which all questions which are below a critical magnitude for the respondent are answered in the same way, and all questions above this point are answered in the opposite fashion. Exhibit 7-2 contains a hypothetical set of questions to measure the threshold of what constitutes pornography for three subjects, and shows how these three subjects could have respondent to the questions:

Exhibit 7-2(A) shows how we can score responses: Subject A would receive a score of 2 on this scale, Subject B a score of 3, and Subject C would score 4. Subject C would be thought to have the highest threshold for pornography of these three individuals.

But suppose the subjects responded as in Exhibit 7-2 (B). This kind of response is possible if the scale is not unidimensional, or if the scale language is not being interpreted the same by each respondent, or if the respondents are simply not responding in a reliable manner. Person B, for instance, would not be expected to rate lingerie advertisements as pornographic once Playboy magazine has been rated as not pornographic. To do so might mean that another dimension, such as accessibility by minors, plays a role in these judgments. In that case the scale would not be just measuring the single dimension of a person's threshold for pornography, which is what we would expect the scale to do. Instead the scale might be measuring that person's threshold for pornography depending upon whether minors do or do not have access to the content and would not be unidimensional.

Cumulative scaling provides a statistic, called the Coefficient of Reproducibility which indicates the degree to which the pattern of responses are consistent with those which would be ex-

pected in a perfect unidimensional scale. Its formula is:

CR = 1.0 - (Number of Inconsistencies / Number of Choices)

In the example in Exhibit 7-2 there are 12 choices (3 subjects x 4 questions). The inconsistencies are marked in part C of that figure as (-).For person B the inconsistency lies in the fact that after rating Playboy as nonpornographic, lingerie ads are rated as pornographic again. Similarly C rates the New York Times as pornographic after having said that lingerie advertisements are not. Using the data from Part C we determine the value of the Coefficient of Reproducibility to be:

CR = 1.0 - (2 / 12) = .8333

The higher this coefficient, the more confident you can be that the scale is measuring only a single theoretical concept (that is, that every respondent is interpreting the language similarly) and that the respondents are replying in a reliable and logical fashion.

Thurstone and Guttman procedures can be combined to create an equal-appearing interval, unidimensional scale. But the amount of effort required to create such a scale would probably only be expended where measurement is critical. In many cases Likert-type magnitude scales are sufficient to meet typical measurement demands.

Reliability

Establishing the reliability of the measurement is critical to good scientific observation and allows us to increase confidence in our findings.

Remember that one of the basic requirements of science is that independent observers measuring the same theoretical concept will always see the same thing, regardless of when or where the measurements are made. While the concept "freedom" may mean many things to many people, this should never be the case with a scientific measurement. All measurements have to exhibit two basic characteristics: stability and consistency. To the degree that they do, we call them reliable measures.

Stability

A stable measure will yield identical measurement results whenever it encounters an identical amount of the theoretical concept. To illustrate this, let's consider a thermometer scale which is a measure of the theoretical concept "temperature". If the measurements of a thermometer are stable, the thermometer will give identical results whenever it encounters the same temperature. To test this stability, we'll take the thermometer and place it in a jar of ice water for several minutes, then record its scale reading. Suppose the thermometer reads 0 degrees Celsius. Then we'll take the thermometer out of the water and let it return to room temperature. After a while, we'll again place it in the ice water, let it remain a few minutes, and read the scale once more. If the thermometer scale is a reliable measure of temperature, it will read 0 degrees once again. But suppose it now reads 2 degrees C instead of zero. If we repeat the procedure several times, we find that the thermometer reads -1 degree C., 3 degrees C., 0 degrees C., and -2 degrees C., on subsequent trials. This thermometer is exhibiting some instability in measurement, and thus it is somewhat unreliable.

Any measure used in communication research can also be tested for stability in a similar fashion. This procedure is often called test-retest reliability. Suppose we have a measurement instrument which quantifies the theoretical concept "communication apprehension", defined as the amount of fear or nervousness a person experiences before giving a public presentation. We would expect the measure to give identical results if it was given to the same person on two consecutive days (or some similar short time interval within which we can be fairly sure that nothing happened to change the amount of apprehension). To test the stability of the measure, we might select a random sample of college students, and give them the test in two consecutive class meetings. Each person will then have two scores, one for the first response to the instrument and one for the second. If the two scores are identical for all persons, the measure is perfectly reliable.

But since perfect reliability is a very unusual situation in behavioral research, we need to use some mathematical way of expressing the amount of stability shown by a measure. One way would be to simply take the difference between the two scores for each person, and average this difference

over all persons. The resulting "average stability error" will give some indication of the test reliability, expressed in the same units as are used by the measurement. More commonly, stability is computed with a correlation coefficient (described in Chapter 20) which ranges from +1.0 for perfect reliability to 0.0 when no consistent pattern of relationship can be found between the second measurement and the first. This reliability index has the virtue of being standardized, that is, of having the same range and meaning for any measurement instrument, regardless of its actual measurement units, be they temperature degrees, apprehension scale points, or any other metric.

Some kinds of measures have characteristics that make test-retest reliability checks inappropriate. Usually this happens when there is something about the first measurement which affects the second. In the case of the communication apprehension scale, it is possible that during the retest subjects might remember the scale questions and their answers from the first test administration. The consistency in answers introduced by memory will falsely inflate the stability estimate. In these cases, a multiple sample (sometimes called a dual sample or split sample) check may be used.

We already know that two (or more) random probability samples drawn from the same population will have the same characteristics (subject to some sampling error, which decreases as the number of observations in the samples increases). A stable measuring instrument applied to each sample should give readings which are identical, at least within the range of expected sampling error. We can thus apply the measure to two or more random samples from the same population, and compare the results.

Describing the actual statistics for quantifying the degree of multiple sample stability will have to wait until we introduce the ideas of sampling distributions and inferential tests for differences between two or more samples in much more detail in later chapters.

Consistency

Stability is only one characteristic of reliable measurement. Reliability, in addition, demands that our operational definition describe a measurement procedure which behaves in a consistent fashion. There are two major kinds of consistency.

Inter judge or inter coder reliability determines the consistency with which the measurement rules, categories, or procedures defined in the operational definition are applied by human judges. In a content analysis of newspaper stories, for example, we might want to determine the amount of coverage devoted to a content category called "positive economic news". The amount of coverage is to be measured by calculating the number of square inches devoted to this type of coverage in each newspaper in our sample. However, the nature of "positive economic news" is open to interpretation by the person doing the measurement. To the extent that two coders differ in their judgment of ambiguous stories, the measure will be unreliable.

To assess the amount of unreliability, we can give two or more coders the same stories to measure. The reliability of the "positive economic news" variable can then be determined by finding the extent of agreement among the coders. The average correlation among the coders might be used to characterize the inter coder reliability. We might set a lower limit of .80 for reliability as a limit below which we will not consider the measurement of the variable as reliable enough to be useful. If the reliability figure is 1.0, the variable "positive economic news" is perfectly reliable, and we can trust the observations made with this operational definition. But if it is .65, we will have to take some corrective action. The first thing we can do is to improve the operational definition. This can be accomplished by being more specific in what, to us, constitutes positive economic news. This will make it easier for the coders to recognize the concept being measured and thus to agree upon its presence and amount. A second thing we can do is to improve our measurement procedure. For instance, we might train coders more thoroughly to improve their ability to recognize "positive economic news". Or we might use more diligent persons as coders.

Another kind of consistency is important when a measure is made up of more than one item, indicator or question. This is the internal consistency of the indicator items. If the items are all supposed to be measuring the same theoretical concept, they should perform in predictable ways. To the degree that they do not, the measure constructed from them is unreliable.

We have already shown an example of one kind of internal consistency test in our discussion of the Coefficient of Reproducibility in Guttman scaling. Since each test item is supposed to be measuring a different magnitude of the same concept, we can predict the pattern of responses which we should see, and use a numerical measure to compute the degree to which we actually see the

pattern.

One common way to combine measurement indicators is simply to add scale scores. Suppose we wished to measure a concept called "positive facial expression" in a nonverbal communication experiment. We ask the subjects in an experiment to rate a number of photographs on a series of Likert-type scales, such as the ones in Exhibit 7-3:

Exhibit 7-3 Examples of Likert Scaling

Pleasant

1	2	3	4	5	6	7
Very					Not at all	

Happy

1	2	3	4	5	6	7
Very					Not at all	

Appealing

1	2	3	4	5	6	7
Very					Not at all	

Favorable

1	2	3	4	5	6	7
Very					Not at all	

To compute the measure of overall positive facial expression, our operational definition instructs us simply to sum all the scale values. We expect that each of these items should be answered similarly; that is, photographs scoring high on one scale should score high on the others, since all scales indicate similar aspects of the theoretical concept. If we see experimental subjects consistently rating photographs in the "very happy" range at the same time they rate them as "not at all appealing", we lack internal consistency within the items.

There are a number of ways to measure internal consistency, some of them requiring very sophisticated statistical procedures. They are discussed extensively in some of the texts we've listed at the end of this chapter. We'll mention some to guide you in your search. A very common indicator of internal consistency is Cronbach's Alpha. One can also use correlation coefficients or factor analysis to determine the degree of similarity in scale responses.

A note of caution: it is not always necessary to have internal consistency in measurement items in order to have a reliable measure. Internal consistency is required only when a series of individual items is used to measure a common aspect of a theoretical construct. If each item measures a different aspect of the concept, then the series of items does not necessarily have to be answered in a consistent pattern. In the example above, if we consider "happy" and "favorable" as somewhat independent components of positive facial expression, it would not be necessary for the subjects to rate the photographs similarly on each scale. A expression could be simultaneously happy and unfavorable or favorable and unhappy. This would mean that measures of internal consistency could be low.

However, if the items are actually independent, as in the above example, we probably should not be simply adding together the scale values. To do so is to add possibly unrelated units (apples to oranges), which may give us misleading/distorted results (is one unit of "happy" really equal to one unit of "favorable"?). Low internal consistency figures may indicate that we should examine our theoretical definition, to make sure that it truly is unidimensional. If it is not, we'd be well advised treating the concept as a multi-dimensional construct.

The perceptive reader may wonder why we would bother with internal consistency at all, since it just measures the similarity of response to duplicate measurement items. Why not just get rid of the duplication and use a single item? Instead of using four scales to measure positive facial expression, why not use one scale that just asks for the amount of such expression?

The reason is simple—using multiple items increases reliability. Any single item may be somewhat unreliable by itself, but in conjunction with a number of similar items, may produce a reliable measure. Two different subjects in our experiment may rate a photograph slightly differently on the "happy" scale and also slightly differently on the "favorable" scale, etc., but these small differences will tend to cancel over the whole set of scales (See Table 7-2).

Table 7-2 shows that, although the different subjects have somewhat different ratings of the photograph on the individual scales, their overall summed ratings of the positiveness of facial expression are identical.

Reliability may also increase with multiple indicators because each item can be described or defined in much more concrete terms than a single operational measure of the overall concept. For example, it is more specific to ask to what degree the photograph shows a "happy", a "favorable" or

a "pleasant" expression than to ask for "positive" expressions. Providing multiple items increases our ability to obtain accurate responses as the respondents do not have to define "positive" for themselves. Making such a definition might introduce a large amount of difference in responses because of individual interpretation.

Table 7-2 Increasing Reliability with Multiple Items

	Subject 1	Subject 2	Subject 3	Subject 4
Pleasant	6	5	5	7
Happy	5	6	6	6
Appealing	6	6	7	5
Favorable	4	4	3	3
Positive Facial Expression (Sum)	**21**	**21**	**21**	**21**

Validity

Our measurement must be not only reliable, it must also be valid. As we saw in Chapter 2, the degree to which the operational definition reflects the same meaning as the theoretical definition determines the most critical kind of validity, measurement validity. (This is also sometimes called empirical validity). The amount of measurement validity cannot be determined by any numerical method. It relies on a self-evident overlap between "verbal world" theoretical definitions and "measurement world" operational definitions. Furthermore, the self-evident overlap must be generally agreed upon by independent observers. It is not enough that an operational definition shows measurement validity to the researcher who constructs it; it must also exhibit the same measurement validity to other researchers and critics.

Some kinds of validity can be inferred by observing the pattern of relationships between measurements. The basic logic of these kinds of validity tests is explained below.

Concurrent Validity

If my operational definition provides valid measurement, the results it gives should covary strongly (or agree) with the results given by other operational definitions of the same concept or measures of a related concept. This is called concurrent or convergent validity. If I construct a measure of "reading ability", I expect that it will correlate highly with a high-school student's SAT-Verbal score which reflects general verbal ability. If it does not, there is some reason to question the validity of my measurement. Valid measurements may not correlate perfectly with other measures of the same concept, however, for two reasons other than a mismatch between the theoretical and operational definitions: First, the measurements of the concepts being compared are probably not perfectly reliable. An unreliable measure cannot correlate perfectly with any measure, as it contains some random "noise". Second, the two theoretical definitions of the concept may differ somewhat. My definition of verbal ability may emphasize vocabulary, while the SAT-Verbal definition may emphasize logical relationships in language. Since the "verbal world" of the two theoretical definitions does not overlap perfectly, the "measurement world" of their respective operational definitions cannot overlap perfectly either. Because of these limitations, we usually expect to find a moderate, but not perfect, convergent validity.

Discriminant Validity

If my operational definition provides valid measurement, the results it gives should NOT covary strongly with the results given by measures of different concepts. This is called discriminant

validity. My measure of "verbal ability" should not correlate strongly with a student's SAT-Math scores. If the relationships between my "verbal ability" measure and the SAT-Verbal and the SAT-Math scores are comparable in size, I can conclude that I have an indiscriminate measure, and thus one which is at least partially invalid. Perhaps my operational definition really measures intellectual abilities of all kinds, and thus is only partially a valid measure of "verbal ability".

Construct Validity

The object of assessing convergent and discriminant validity is to determine that an operational definition provides measurement of only the defined concept, and not measurement of any other concept. But the degree of measurement of the target concept can vary. That is, it is possible that an operational definition measures only part of the meaning outlined in the theoretical definition. The degree to which the operational definition taps the full meaning of the theoretical definition is called construct validity. This is obviously closely related to measurement validity, and the two terms are sometimes used interchangeably. However, we will consider construct validity as a quantifiable idea, like convergent validity.

One way to assess construct validity is by means of multiple indicators. If the meaning of the theoretical concept or construct is somewhat abstract, one way to measure it is by operationally defining a number of different measurement indicators which get at different parts of the meaning. In Chapter 2 we saw an example of this in the definition of "source credibility." This concept was defined as a combination of formal education, experience, job title, mode of dress, objectivity, and other more concrete concepts. If we operationally define these concepts, so that they are measured as separate indicators of source credibility, we can combine them to form a measurement of the more abstract concept of "source credibility". The question of construct validity then becomes this: how well do my indicators, when combined, represent the full meaning of the theoretical concept? To answer this question, I can ask a sample of respondents to rate a number of communication sources on the above indicators, and also on a summary scale of "source credibility". The covariance between the total set of indicators and the summary scale is an estimate of the construct validity. The statistical methods necessary to calculate this estimate will be addressed much later in this book. They involve multiple correlation and other advanced statistics.

Figure 7-5 graphically illustrates the relationships between theoretical meanings and operational measurement which have been outlined here.

Summary

Chapter 2 of this book provided an introduction to the explication of theoretical concepts: the process of producing theoretical and operational definitions. In this chapter we have extended this discussion of operational definitions to include detailed descriptions of how different strategies in operationalization will yield different levels of measurement, how these levels of measurement can be quantified by the different types of scaling methods and how we can assess the adequacy of measurement.

The level of measurement is an important topic due to the varying amounts of information provided by the different levels of measurement. Nominal measurement represents the lowest form of measurement: the various categories in a nominal factor are merely mutually exclusive and exhaustive. An observation assigned to a given category is considered identical to all the other observations in that category and not identical to the observations in other categories. Ordinal measurement adds to this equivalence/nonequivalence the dimension of magnitude: the categories in an ordinal variable can be ordered as representing "more" or "less" of a particular attribute. Interval measurement adds to this the notion of equal intervals, be they of an absolute or of an arbitrary nature. The presence of equal intervals allows us to extend statements of "more" to "How many more units more". Finally, ratio measurement incorporates an explicit or implicit absolute zero indicating the absence of whatever it is we attempt to measure. The presence of zero means that different observations can be compared in statements such as "twice as much as", or "only half as much as".

The process of converting ordinal or better levels of measures to numerical measurement is called scaling. We distinguished between two general categories of scaling: measurement in terms of natural or established metrics which requires either interval or ratio measurement, and the scal-

ing of magnitude, which applies generally to measurement which is inherently ordinal in nature. Specific examples of magnitude scaling are Likert scales and the Semantic Differential scale. The Thurstone and Guttman types of scales can be considered to be attempts to introduce equal or nearly equal intervals in what is essentially ordinal measurement.

Regardless of level of measurement the most important criterion of evaluation of any measurement scheme has to be the adequacy of measurement. This adequacy can be evaluated in terms of two different dimensions: the reliability of measurement and the validity of measurement.

The reliability of measurement refers to how well a measurement scheme measures, and its ability to do that can be expressed in terms of its stability and its consistency. A measure is considered to be stable whenever it gives identical results whenever an identical observation is encountered. A stretchable latex ruler used to measure the same table repeatedly would not be stable, as it would

likely stretch differently on each application. A steel ruler, however, would give stable results. The latex ruler would also not be consistent; different people using this ruler would probably ob-

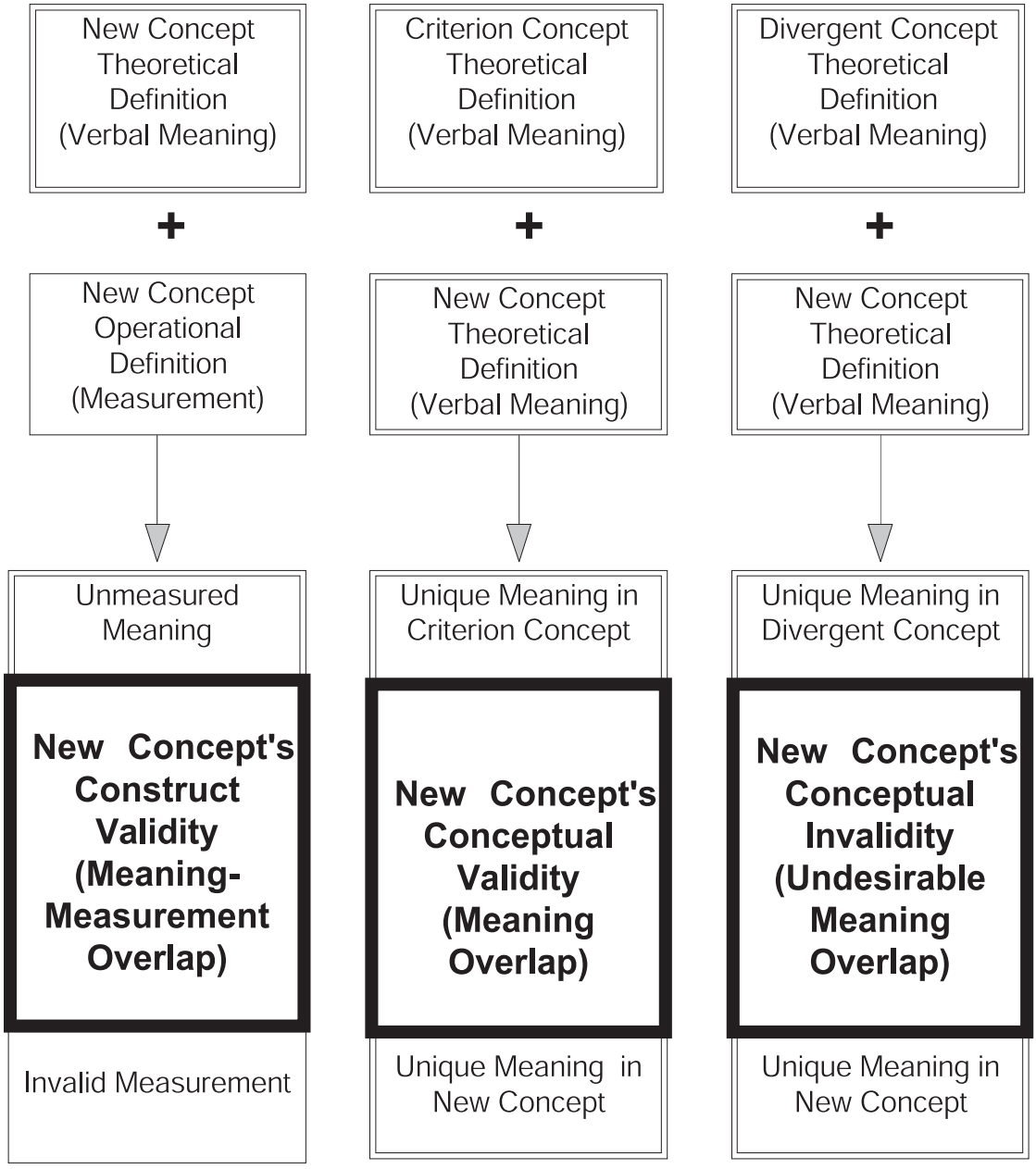


FIGURE 7-5 Types of Validity

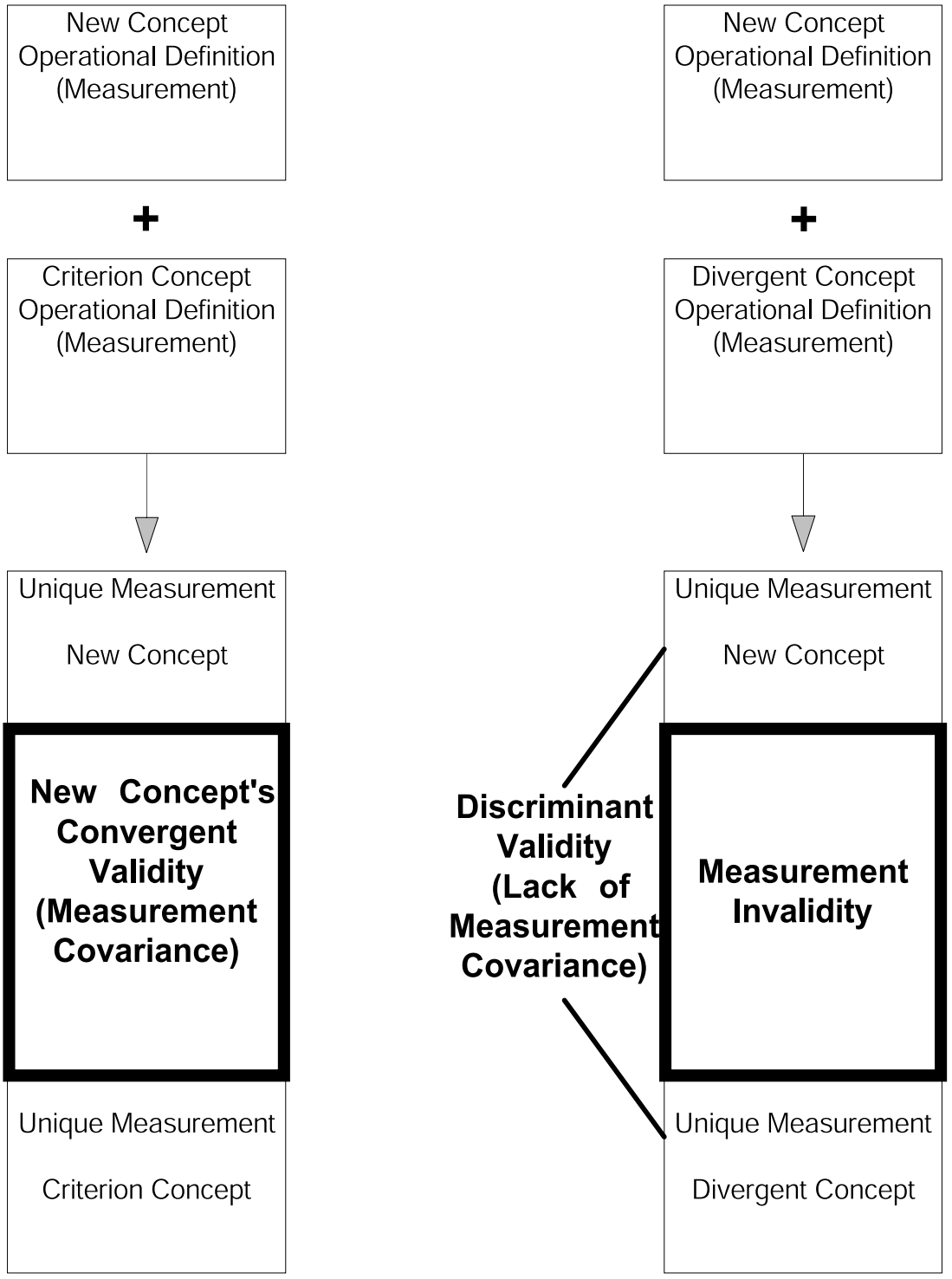


FIGURE 7-5 cont.

serve different results. The steel ruler, again, would yield identical measures for different users. Internal consistency is another aspect of reliability that can be determined whenever multiple scales are used to assess a single theoretical construct. Internal consistency refers to the extent to which such scales yield similar results.

In addition to assessing how reliably we can make measurements, we should also be concerned about validity: whether we indeed measure what we intend to measure. One way to assess validity is through measurement validity or empirical validity. There is no quantitative index for

this type of validity; it is based on agreement by independent observers that there is sufficient overlap between the meaning contained in the theoretical definition and the measures of the operational definition. In addition, some quantitative measures of validity exist. Convergent validity measures the extent to which our measurement agrees with other measures that purport to measure the same meaning. Measures of discriminant validity, on the other hand, are based on the premise that our measurement should disagree with other measures that purport to measure some other, different, concept.

References And Additional Readings

Babbie, E.R. (1992). *The practice of social research* (6th ed.). Belmont, CA: Wadsworth. (Chapter 5, "Conceptualization and Measurement"; Chapter 7, "Indexes, Scales, and Typologies").

Campbell, D.T & Fiske, D.W. (1959). *Convergent and discriminant validation by the multitrait-multimethod matrix*. Psychological Bulletin, 56, 81-105.

Gorden, R. (1977). *Unidimensional scaling of social variables-concepts and procedures*. New York: Free Press. (Chapter 1, "Scaling Theory").

Guttman, L. (1974). *The basis for scalogram analysis*. In G.M. Maranell (Ed.), Scaling: A sourcebook for behavioral scientists. Chicago: Aldine.

Kerlinger, F.N. (1986). *Foundations of behavioral research* (3rd ed.) New York: Holt, Rinehart and Winston. (Chapter 29, "Objective Tests and Scales").

Likert, K. (1974). *The method of constructing an attitude scale*. In G. Maranell (Ed.), Scaling - a sourcebook for the behavioral sciences (pp. 233-243). Chicago: Aldine.

Miller, D.C. (1983). *Handbook of research design and social measurement* (4th ed.). New York: Longman. (Part 4, "Selected Sociometric Scales and Index").

Osgood, C.E., Suci, G.J. & Tannenbaum, P.H. (1957) *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Oppenheim, A.N. (1966). *Questionnaire design and attitude measurement*. New York: Basic Books. (Chapter 4, "Checklists, Rating Scales, and Inventories"; Chapter 6, "Attitude-Scaling Methods").

Smith, M. J. (1988). *Contemporary communication research methods*. Belmont, CA: Wadsworth. (Chapter 4, "Collecting and Measuring Data").

Thurstone, L. & Chave, E. (1929). *The measurement of attitude*. Chicago: University of Chicago Press.

Torgerson, W. S. (1958) *Theory and method of scaling*. New York: Wiley. (Chapter 1, "The Importance of Measurement in Science"; Chapter 2, "The Nature of Measurement"; Chapter 3, "Classification of Scaling Methods").