# Contents in Brief

## Preface

<table>
<thead>
<tr>
<th>Part</th>
<th>CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part I</td>
<td>FOUNDATIONS OF EMPirical RESEARCH</td>
</tr>
<tr>
<td>1</td>
<td>The Scientific Approach</td>
</tr>
<tr>
<td>2</td>
<td>Conceptual Foundations of Research</td>
</tr>
<tr>
<td>3</td>
<td>Elements of Research</td>
</tr>
<tr>
<td>4</td>
<td>Ethics in Social Science Research</td>
</tr>
<tr>
<td>Part II</td>
<td>DESIGN AND STRUCTURE OF RESEARCH</td>
</tr>
<tr>
<td>5</td>
<td>Research Designs: Experiments</td>
</tr>
<tr>
<td>6</td>
<td>Research Designs: Cross-Sectional and Quasi-Experimental Designs</td>
</tr>
<tr>
<td>7</td>
<td>Measurement</td>
</tr>
<tr>
<td>8</td>
<td>Sampling and Sample Designs</td>
</tr>
<tr>
<td>Part III</td>
<td>DATA COLLECTION</td>
</tr>
<tr>
<td>9</td>
<td>Observational Methods</td>
</tr>
<tr>
<td>10</td>
<td>Survey Research</td>
</tr>
<tr>
<td>11</td>
<td>Questionnaire Construction</td>
</tr>
<tr>
<td>12</td>
<td>Qualitative Research</td>
</tr>
<tr>
<td>13</td>
<td>Secondary Data Analysis and Sources</td>
</tr>
<tr>
<td>Part IV</td>
<td>DATA PROCESSING AND ANALYSIS</td>
</tr>
<tr>
<td>14</td>
<td>Data Preparation and Analysis</td>
</tr>
<tr>
<td>15</td>
<td>The Univariate Distribution</td>
</tr>
<tr>
<td>16</td>
<td>Bivariate Analysis</td>
</tr>
<tr>
<td>17</td>
<td>Control, Elaboration, and Multivariate Analysis</td>
</tr>
<tr>
<td>18</td>
<td>Index Construction and Scaling Methods</td>
</tr>
<tr>
<td>19</td>
<td>Inferences</td>
</tr>
</tbody>
</table>

## Appendices

| A | Introduction to SPSS |
| B | Writing Research Reports |
| C | Σ: The Summation Sign |
| D | Random Digits |
| E | Areas under the Normal Curve |
| F | Distribution of t |
| G | Critical Values of F |
| H | Distribution of \( \chi^2 \) |

## Glossary

| Subject Index |
| 531 |
| Author Index |
| 537 |
INDEX CONSTRUCTION
The Purpose of the Index
The Sources of the Data
The Base of Comparison
Methods of Aggregation and Weighting
Index Construction: Examples
Attitude Indexes

SCALING METHODS
Likert Scales
Other Composite Measures
Guttman Scaling
Guttman Scale Application: An Example
Factor Analysis
How's the nation (or the city) doing?" Candidates for public office, voters, and social scientists have asked this question in any number of situations, not just during elections. The answers they seek are often found in special "thermometers" designed for that purpose: the Consumer Price Index, which measures the ups and downs of inflation; the FBI Uniform Crime Report, which serves as a measure of social order; and the World Bank's Social Indicators of Development, which provides data indicating how the world's nations have progressed and how they compare to each other on a number of dimensions. One group of concerned social scientists, centered at the Fordham Institute for Innovation in Social Policy in Tarrytown, New York, has developed an index, described in its report "The Index of Social Health," that measures the progress made toward solving social problems. The institute's researchers keep track of changes in 16 socioeconomic indicators, including juvenile suicide, high school dropout rates, and out-of-pocket health costs for the elderly. They then calculate a single, summary figure, ranging from 0 to 100. The higher the figure, the healthier the nation is. In their 1992 report, 9 of the 16 indicators either stayed the same or declined; the rise in the other 7 was sufficient to raise the total index—a sign that total social health is improving. Dr. Marc L. Miringoff, the institute's director and author of this report, notes that the index often rises during presidential election years, a response, he believes, to the increase in funding typical of those periods.1

How social scientists construct such indexes, and why they do so, are the major topics discussed in this chapter.


There are several reasons why social scientists employ scales and indexes, some of which are quite practical. First, indexes and scales enable researchers to represent several variables by a single score, a quality that reduces the difficulties of dealing with complex data. Second, the quantitative measurements obtained with scales and indexes are amenable to more precise statistical analysis. Finally, using indexes and scales increases the reliability of measurement itself. This last feature is based on the fact that a score on a scale or an index is based on responses to more than a single question or item. To clarify this issue, let's consider one of the more pedestrian scales used regularly—exams and their grades (or scores). Few students would like to have an exam grade determined solely by their answer to a single multiple-choice or true-false question. In the first place, it is unlikely that the total universe of course material can be covered by a single question and answer. Second, if a student misinterprets or makes a mistake on a one-question exam, that one error could mislead the professor in his or her judgment of how much knowledge the student has absorbed. If there are several questions, the grade will be more accurate because the student will have additional opportunities to demonstrate how much was learned.

Another illustration is the measurement of sexual harassment. How can we predict the effect of harassment on a person's health—or calculate the amount of compensation due if a case goes to court—if we cannot gauge the severity of the behavior considered to be harassing? Therefore, some instrument has to be devised to differentiate between, say, an offensive remark made by a superior and that superior's preventing an employee's promotion if sex is denied.

Following this line of argument, it becomes obvious that researchers use multiple-item scales and indexes to increase the reliability and precision of their measurements. The reasons for doing so are theoretical, methodological, and practical.

Scales differ from indexes by their greater rigor. Whereas researchers in all the sciences construct indexes by simply accumulating scores, when social scientists are constructing scales, they pay special attention to tests of validity and reliability. These tests can be applied because most scales incorporate the principle of unidimensionality in their construction. According to this principle, the items comprising a scale reflect a single dimension and can be placed on a continuum presumed to apply to one and only one concept.

A scale's unidimensionality does, however, permit its use for a variety of purposes and in conjunction with various techniques. Some scales are employed to identify questions or items that do not belong to the set of items in which they were originally placed. Other scaling techniques permit us to rank items by their level of difficulty or intensity. In addition, some scaling methods produce interval-level scales and thereby avoid the limitations imposed by nominal or ordinal data.

Before constructing a new scale, the literature should be surveyed to ascertain whether an appropriate scale is already available. The additional readings at the end of this chapter represent a comprehensive list of sources of information on the scales currently available in the social sciences.

INDEX CONSTRUCTION

If you combine two or more items or indicators to create a composite measure, you are constructing an index. The familiar Consumer Price Index (CPI) is a composite measure because it presents the changes taking place in the retail prices of eight major product categories—food, housing, apparel, transportation, medical care, personal care, reading and recreation, and other goods and services—as a group, together. The approximately
400 commodities and services included in the CPI were selected because researchers consider the changes in the prices of these items to be representative of the price trends displayed by related items. These items include the cost of commodities ranging from the price of hamburgers, gasoline, and men's work gloves to home computers, and services such as haircuts, legal advice, and interest on mortgages. The U.S. Department of Labor collects the data in 50 urban areas. These areas have been selected on the basis of their representativeness in terms of urban characteristics—size, climate, population density, income level, and so forth—that affect the way in which families spend their money. Within each city, prices are recorded as they appear in the shops where families typically purchase goods and services. For each item, the prices reported by the various sources are combined and weighted to ascertain average price changes for the area. The Department of Labor prepares an index of these figures monthly for the country as a whole and for each of five major cities, and quarterly for other cities.2

Four major issues influence the construction of indexes: the purpose for which the index is being compiled, the sources of the data, the base of comparison, and methods for aggregating and weighting the data.

The Purpose of the Index

Two essential questions must be answered before the process of index construction can begin: "What are we attempting to measure?" and "How are we going to use the measure?" Logically, if A is an index (or indicator) of X, A may still be only one of several indexes of X. Thus, some kind of supporting evidence is needed to make the case that the values of A correspond to the values of X in a more precise and valid way than other indexes do. Most often, X is a broad concept, such as public welfare or technological change or political participation. Such concepts consist of a complex combination of phenomena, some of which may be subject to differing interpretations. Accordingly, no single indicator will be capable of covering all the dimensions involved, and a number of indicators will have to be selected. Each possible indicator will serve a specific purpose, set forth and explained prior to construction of the index.

The Sources of the Data

Social scientists can employ either obtrusive or unobtrusive (or both) methods of data collection when constructing indexes. What source of data should be used depends on the purpose of the index and the research design employed. In all cases, the investigator must ascertain whether the data relate strictly to the phenomenon being measured. This decision involves application of the validity and reliability standards discussed in Chapter 7.

The Base of Comparison

Indexes are expressed in the form of a proportion, a percentage, or a ratio in order to simplify comparison of the data. Each of these measures can be calculated either in terms of the actual data collected or reported, or in terms of another year's data, which is used as a numerical base for adjusting nonequivalent figures in order to enable their comparison. A proportion is defined as the frequency of observations in any given category (f) divided by the total number of observations (N), or f/N. A proportion becomes a percentage when multiplied by 100 [(f/N) x 100]; a percentage, by definition, may range from zero to 100. A ratio is a fraction that expresses the relative magnitude of any two sets of frequencies. To find the ratio between two frequencies, divide the first frequency by the second. For example, if a group consists of 500 females and 250 males, the ratio of females to males is found by dividing 500 by 250, or 2/1.

Table 18.1 illustrates the use of these measures. It shows the frequencies, proportions, and percentages of selected criminal offenses in a New Jersey city reported by the sources of the official data.3 The data represented in the first three columns, representing the frequency, proportion, and percentage of crimes in New Jersey (CNJ), is compiled annually by the office of the state attorney general. The data for the other three columns, representing the frequency, proportion, and percentage of crimes in the city, were compiled from municipal court dockets (MCD).

An examination of the table reveals serious ambiguities in the crime data. First, there are differences in the amounts of officially reported crimes: The state’s attorney general reports more offenses than court docket data. This is consistent with the fact that the jurisdiction providing the information reported in the MCD columns is different from the jurisdiction providing the information reported in the CNJ columns: CNJ offenses are those known to police and hence to the attorney general, whereas court dockets report only those cases in which offenders have been identified, arrested, and booked in response to official complaints. Given that many offenders are not apprehended, we would expect the attrition of cases reported as we move up through the legal system. Second, the data for the "simple assault" category are highly problematic because the official records indicate that municipal courts heard more cases than were known to the police in a city where the courts and the police share the same jurisdiction. Because such a situation is quite improbable, an index based on these data will be misleading.


SHIFTING THE BASE. In order to make meaningful comparisons of different indexes, or of the same index during different years or between regions, we are often required to shift the base of the index. When we shift a base, we are converting or standardizing the data in such a way as to make them comparable. This procedure is often methodologically necessary because only by converting the data to a uniform base can social scientists analyze trends or changes in the phenomena studied. For example, we might shift the base of an index number series from one year to another. Uniformity would be achieved by setting the values for that year at 100. As we shall see, the Cost of Living Index employs such a procedure. It uses the prices for a selected year as the base for comparing prices for the next 5 or 10 years.

In the following example (see Table 18.2), we will be converting the values of the index of, let’s say, small business start-ups based on 1995 data into a new index, one based on 1990 data. We do this because we have hypothesized that interest rates for small-business loans in an earlier year, 1990, may have had a greater long-term impact on start-ups than did interest rates in the later year, 1995. To obtain the new index (values based on 1990 data) for 1995, we divide the original figure (based on the 1995 index) for 1990 by 70 (the number of indexed start-ups for 1990) and then multiply by 100. This results in (70/70) × 100 = 100. The new value for 1991 equals (80/70) × 100 = 114.3, and so forth, until all the original figures have been converted into the new series.

Methods of Aggregation and Weighting

A common method for constructing indexes is the computing of aggregate values. The aggregates can be either simple or weighted, depending on the purpose of the index.

SIMPLE AGGREGATES. Table 18.3 illustrates the construction of a simple aggregate price index. The prices of each commodity (C) in any given year are added to give the index for that year. As noted earlier, it is convenient to designate some year as a base, which is set equal to 100. In this example, all the indexes are expressed in the last row as a percentage of the 1995 figure, obtained by dividing each of the numbers by the aggregate value in the base year ($20.13) and multiplying by 100. Symbolically,

\[ PI = \frac{\sum P_r}{\sum P_o} \times 100 \]  

(18.1)

where \( PI \) = price index
\( p = \) the price of an individual commodity
\( o = \) the base period according to which the price changes are measured
\( n = \) the period being compared with the base

The formula for a particular year (for instance 1999, with 1995 being the base) is

\[ PI_{1999} = \frac{\sum P_{1999}}{\sum P_{1995}} \times 100 \]  

(18.2)

Thus,

\[ PI_{1999} = \frac{6.10 + 7.18 + 7.90 + 6.80}{3.21 + 5.40 + 6.62 + 4.90} \times 100 \]

\[ = \frac{27.98}{20.13} \times 100 = 139.00 \]

WEIGHTED AGGREGATES. Simple aggregates may conceal the relative influence exerted by each indicator used in the index. To prevent such misrepresentation, weighted aggregates are often used. To construct a weighted aggregate price index for the data displayed in Table 18.3, list the quantities of the specific commodities (i.e., consumption) and determine the aggregate value of those goods, that is, how much the total is worth in each year’s current prices. This means that each unit price is multiplied by the number of units of the commodity marketed, produced, or consumed, with the resulting values summed for each period. Symbolically,

\[ PI = \frac{\sum P_r q}{\sum P_o q} \times 100 \]  

(18.3)

where \( q \) represents the quantity of the commodity marketed, produced, or consumed during the base year, that is, the quantity weight or multiplier. The procedure, using the quantities in 1995 as multipliers, is illustrated in Table 18.4. Because the total value changes while the components of the aggregate do not, these changes must be due to
Table 18.4

Construction of Aggregative Index Weighted by Consumption in 1995

<table>
<thead>
<tr>
<th>Commodity</th>
<th>Consumption</th>
<th>Value of 1995 Quantity at Price of Specified Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>800</td>
<td>$2,568</td>
</tr>
<tr>
<td>$C_2$</td>
<td>300</td>
<td>1,620</td>
</tr>
<tr>
<td>$C_3$</td>
<td>450</td>
<td>2,979</td>
</tr>
<tr>
<td>$C_4$</td>
<td>600</td>
<td>2,940</td>
</tr>
<tr>
<td>Aggregate value</td>
<td>$10,107</td>
<td>$11,877</td>
</tr>
<tr>
<td>Index</td>
<td>100.0</td>
<td>117.5</td>
</tr>
</tbody>
</table>

price changes. Thus, the aggregative price index measures the changing value of a fixed aggregate of goods.

Index Construction: Examples

Let us look first at a simple index developed to evaluate the statistics textbooks used in the social sciences according to students’ instructional needs. The index, the Statistics Textbook Anxiety Rating Test (START), uses seven factors keyed to the students’ needs. These factors are related to possible deficiencies in the students’ math backgrounds and the corresponding anxieties they arouse:

1. Reviews basic algebraic operations
2. Contains a section on notations
3. Includes exercise answers
4. Explains exercise answers
5. Does not use definitional formulas
6. Uses relevant examples
7. Addresses student statistics or math anxiety explicitly

The index works as follows: Textbooks are given a score on each factor—1 if the book meets the criterion, 0 if it does not. Summing all the scores yields a composite score ranging from 0 to 7. When the index was used to evaluate 12 popular textbooks, the scores ranged from 0 to 4.

Another well-known example is the Sellin and Wolfgang Index of Delinquency. In order to evaluate crime control policies, policy makers require at least three major types of information: data on the incidence of crimes, data on the response of the criminal justice system, and data on the offenders’ sociodemographic characteristics. With respect to the incidence of crimes, a major problem is that offenses vary in nature and magnitude. Some result in death, others inflict losses of property, and still others merely cause inconvenience. One common way of comparing, say, one year’s incidence of crime with another’s has been simply to count offenses, disregarding the differences in their character. Given the range of crimes and the varying degrees of their seriousness, such unweighted indexes are misleading. A police report showing an overall decrease or increase in the total number of offenses committed says little if significant changes have occurred in the types of offenses committed. For example, a 10 percent decrease in auto theft but a 30 percent increase in rape could lead to a decline in an unweighted crime index because reported auto thefts are usually much greater in absolute numbers than are reported rapes.

In a pioneering attempt to tackle this problem with respect to delinquency, Thorsten Sellin and Marvin Wolfgang developed a weighting system by distributing 141 carefully prepared accounts of different crimes to three samples of “evaluators”: police officers, juvenile court judges, and college students. The accounts included combinations of circumstances, such as death or hospitalization of the victim, type of weapon, and value of property stolen, damaged, or destroyed—for example, “The offender robs a person at gunpoint,” “The victim struggles and is shot to death,” “The offender forces open a cash register in a department store and steals five dollars,” “The offender smokes marijuana.” The evaluators were asked to rate each of these accounts on a “category scale” and a “magnitude estimating scale.” Sellin and Wolfgang used these ratings to construct the weighting system. For example, a crime with the following “attributes” would be given the following number of points:

A house is forcibly entered 1
A person is murdered 26
The spouse receives a minor injury 1
Between $251 and $2,000 is taken 2
Total score 30

With such an index, policy makers and social scientists can make meaningful comparisons over time and between different communities because the seriousness of the crimes committed are taken into account in addition to their frequencies.

Attitude Indexes

To construct an attitude index, researchers prepare a set of questions, selected a priori. Numerical values (e.g., 0 to 4 or 1 to 5) are assigned arbitrarily to the item or question responses. The values assigned by the respondents are added to obtain total scores. The scores are then interpreted as indicators of the respondents’ attitudes. Consider the following five statements designed to measure alienation:

1. Sometimes I have the feeling that other people are using me.
   □ Strongly agree □ Disagree
   □ Agree □ Strongly disagree
   □ Neither agree nor disagree

2. We are just so many cogs in the machinery of life.
   □ Strongly agree □ Disagree
   □ Agree □ Strongly disagree
   □ Neither agree nor disagree

3. The future looks very dismal.
   □ Strongly agree □ Disagree
   □ Agree □ Strongly disagree
   □ Neither agree nor disagree

4. More and more, I feel helpless in the face of what’s happening in the world today.
   □ Strongly agree □ Disagree
   □ Agree □ Strongly disagree
   □ Neither agree nor disagree

5. People like me have no influence in society.
   □ Strongly agree □ Disagree
   □ Agree □ Strongly disagree
   □ Neither agree nor disagree

Suppose that we arbitrarily assign response scores in the following way: Strongly agree = 4; Agree = 3; Neither = 2; Disagree = 1; and Strongly disagree = 0. A respondent who answers “Strongly agree” to all five statements will have a total score of 20, indicating a high degree of alienation; a respondent who answers “Strongly disagree” to all five statements will have a total score of 0, indicating that that person is not alienated at all. In reality, most respondents will obtain scores between these two extremes. The researcher will then work out a scale that classifies respondents according to their degree of alienation on the basis of their total scores—for example, respondents who score 0 to 6 are not alienated, respondents who score from 7 to 13 are somewhat alienated, and those who score between 14 and 20 are most alienated. This type of index is sometimes termed an arbitrary scale because nothing about the procedure guarantees that any one statement or item taps the same attitude as the other items. Do items 3 and 5 tap the same aspect of alienation? Does item 4 correspond to the aspects of alienation not covered by items 3 and 5? Will another researcher who uses the index obtain the same findings? That is, is the index reliable? We will address these important questions next, in our discussion of scaling methods.

**SCALING METHODS**

**Likert Scales**

Likert scaling is a method designed to measure attitudes. To construct a Likert scale, researchers usually follow six steps: (1) compile a list of possible scale items, (2) administer these items to a random sample of respondents, (3) compute a total score for each respondent, (4) determine the discriminative power of the items, (5) select the scale items, and (6) test the scale’s reliability.

**COMPILING POSSIBLE SCALE ITEMS.** In the first step, a series of items is compiled that expresses a wide range of attitudes, from extremely positive to extremely negative. The respondent is then requested to check one of five offered fixed-alternative expressions, such as “strongly agree,” “agree,” “neither agree nor disagree,” “disagree,” and “strongly disagree,” which comprise a continuum of responses. (Occasionally, three, four, six, or seven fixed-alternative expressions are used. Optional expressions include “almost always,” “frequently,” “occasionally,” “rarely,” and “almost never.”) In this five-point continuum, values of 1, 2, 3, 4, 5 or 5, 4, 3, 2, 1 are assigned. These values express the relative weights and direction of the responses, determined by the favorableness or unfavorableness of the item.

In a classic use of the Likert method, Wayne Kirchner developed a 24-item scale to measure attitudes toward employment of senior citizens. The following four items illustrate his scoring technique.

1. Most companies are unfair to older employees.
   □ Strongly agree □ Disagree
   □ Agree □ Strongly disagree
   □ Neither agree nor disagree

2. I think that older employees make better employees.
   □ Strongly agree □ Disagree
   □ Agree □ Strongly disagree
   □ Neither agree nor disagree

3. In a case where two people can do a job about equally well, I'd pick the older person for the job.
   □ Strongly agree □ Disagree
   □ Agree □ Strongly disagree
   □ Neither agree nor disagree

4. I think older employees have as much ability to learn new methods as other employees.
   □ Strongly agree □ Disagree
   □ Agree □ Strongly disagree
   □ Neither agree nor disagree

Kirchner scored this scale by assigning higher weights to responses to positive items (acceptance of the idea of hiring older persons) as follows: Strongly agree, 5; Agree, 4; Uncertain, 3; Disagree, 2; Strongly disagree, 1. If he had included negative items (items indicating rejection of the idea of hiring older persons) in the scale, their weights would have been reversed.

**ADMINISTERING ALL THE ITEMS.** In the second step, a large number of respondents, selected randomly, are asked to indicate their attitudes on the list containing all of the proposed items.

**COMPUTING A TOTAL SCORE.** In this step, the researcher calculates a total score for each respondent by summing the value of all items checked. Suppose that a respondent checked “Strongly agree” in item 1 (score 5), “Uncertain” in item 2 (score 3), “Agree” in item 3 (score 4), and “Disagree” in item 4 (score 2). This person’s total score would be $5 + 3 + 4 + 2 = 14$.

**DETERMINING THE DISCRIMINATIVE POWER.** In the fourth step, the researcher has to determine a basis for the selection of items for the final scale. Regardless of the scaling method employed, the objective is to find items that consistently distinguish respondents who are high on the attitude continuum from those who are low. This can be done by applying either the internal consistency method—that is, by correlating each item with the total score and retaining those with the highest correlations—or with item analysis. Both methods yield an internally consistent scale. With item analysis, the researcher subjects each item to a measurement of its ability to differentiate the highs

---

Table 18.5

Table for Computing the DP for One Item

<table>
<thead>
<tr>
<th>Group</th>
<th>Number in Group</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Weighted Total*</th>
<th>Weighted Mean†</th>
<th>DP (Q1 - Q3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High (top 25%)</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>35</td>
<td>3.89</td>
<td>2.00</td>
</tr>
<tr>
<td>Low (bottom 25%)</td>
<td>9</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>1.89</td>
<td></td>
</tr>
</tbody>
</table>

*Weighted total = score \times number who check that score
†Weighted mean = \frac{weighted total}{number in group}

(clearly positive attitudes) from the lows (clearly negative attitudes). This measure is called the discriminative power (DP) of the item. In calculating the DP, we add the scored items for each respondent and place the scores in an array, usually from lowest to highest. Next we compare the range above the upper quartile (Q3) with that below the lower quartile (Q1), and calculate the DP as the difference between the weighted means of the scores above Q1 and of those that fall below Q3, as illustrated in Table 18.5.

SELECTING THE SCALE ITEMS. The DP value is computed for each of the possible scale items; those items with the highest DP values are selected. These are the items that best discriminate among the individuals expressing differing attitudes toward the attitude being measured.

TESTING RELIABILITY. The reliability of the scale can be tested in much the same manner as we would test other measuring procedures. For example, we can select enough items for two scales (at least 100) and divide them into two sets, constituting two scales. We can then employ the split-half reliability test (see Chapter 7).

Other Composite Measures

Social scientists have developed various scaling procedures that incorporate a number of features taken from Likert scaling techniques. These procedures almost always include the steps described above, such as initial compilation of possible scale items, administration of items to a large number of respondents, and some methods for selecting the set of items to be included in the final scale. The most common format for the items is a rating scale on which respondents are asked to make a judgment in terms of sets of ordered categories.

Most statistical computer programs today include procedures and statistics that make it easier to select items for your own scales and to evaluate how well the various items measure the underlying phenomena.

One of the simplest statistics used to examine items is the bivariate correlation coefficient (Pearson's r), which indicates how closely linked each item is with other items or the entire scale. In general, items that are strongly associated with other items will show higher overall correlations with the total scale. Examining the bivariate correlation helps researchers to decide which items to include in the scale and which items to discard.

Items that correlate strongly with one another should be selected, and vice versa. Another helpful statistic is Cronbach's alpha, which estimates the average of all possible split-half reliability coefficients. (For a discussion of reliability, see Chapter 7.) The alpha measures the extent to which the items comprising the scale "hang together." A high alpha (.70 is an acceptable level) indicates that the items in the scale are "tightly connected."

Guttman Scaling

The Guttman scale, first developed by Louis Guttman in the early 1940s, was designed to empirically test the unidimensionality of a set of items when constructing scales. Guttman suggested that if the items comprising the scale tap the same dimension underlying an attitude, they can be arranged along a continuum that indicates the strength of that dimension. More explicitly, Guttman scales are unidimensional and cumulative. The cumulative characteristic implies that a researcher can order the items by degree of difficulty (or intensity or specificity) and that respondents who reply positively to a difficult item will also respond positively to less difficult items when presented in the appropriate order. Similarly, a respondent who (dis)agrees with a statement expressing a strong attitude will also (dis)agree with a statement expressing a milder version of that attitude. If we take an example from the physical world, we know that if an object is 4 feet long, it is longer than 1 foot and also longer than 2 or 3 feet. In the social world, we know that if an executive director of a corporation would approve of his daughter marrying an electrician, he would also approve of an electrician belonging to his country club. Similarly, if he did not object to an electrician being a member of his club, he would not mind having an electrician for a neighbor.

Table 18.6 illustrates the scale that would result from administering these three items to a group of respondents. This scale is unidimensional as well as cumulative—the items can be ranked on a single underlying dimension, social acceptability in this case; the scale is cumulative because none of the respondents gave totally mixed responses to the questions in the order in which they were presented to them. Thus information on

Table 18.6

A Hypothetical Perfect Guttman Scale

<table>
<thead>
<tr>
<th>Respondent</th>
<th>Item 1: Admit to close kinship by marriage</th>
<th>Item 2: Admit to the same social club</th>
<th>Item 3: Admit as a neighbor</th>
<th>Total Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>

+ indicates agreement with the statement; - indicates disagreement.

the position of any respondent’s last positive response allows the researcher to predict all of the responses to the scale items following the selected item.

In practice, a perfect Guttman scale is rarely obtained. In most cases, inconsistencies appear. Consequently, a criterion is set for the purpose of evaluating how unidimensional and cumulative the scale really is. The coefficient of reproducibility (CR), developed by Guttman, does just that: It measures the degree of conformity of the scale to what would be a perfectly unidimensional and cumulative scale. The greater the conformity of the scale being constructed to a perfectly arranged scale, the greater is the validity of the scale. The CR is discussed further later in the chapter.

SELECTING SCALE ITEMS. Raymond L. Gorden lists three conditions, in their order of importance, that must be met in the process of discovering and selecting items for a Guttman scale: 8

1. An attitude toward the object (class of objects, events, or ideas) must actually exist in the minds of the population to be sampled and tested.
2. Statements about the object can be constructed that have meaning to the members of the sample and elicit responses that are valid indicators of that attitude.
3. The items in the set of statements or questions must represent different degrees along a single dimension.

Attitude scale items can be selected by a variety of methods and from any available source: newspapers, books, scholarly articles, personal knowledge of the phenomenon. Experts in interviewing as well as a small group of respondents, chosen from the population to be studied, can also contribute good items. (It should be clear by now that the process of securing the items for scales is very similar to that involved in index construction.) After a large set of potential items is compiled, the researcher selects a preliminary set of items. These items should relate clearly to the attitude being measured and should cover the total continuum from strongly favorable to strongly unfavorable. Two to seven response categories may be constructed for each statement expressing the item. The most common formats are Likert-type items with five-point scales, as in the following example:

Please indicate how much you agree or disagree with the following statement:

Nowadays a person has to live pretty much for today and let tomorrow take care of itself.

☐ Strongly agree ☐ Disagree
☐ Agree ☐ Strongly disagree
☐ Neither agree nor disagree

The selected items are then included in a questionnaire (the pretest) that the researcher administers to a sample of the target population. Before the answers to the questionnaire are scored, items are arranged so that higher values will consistently stand for either the most positive or most negative feelings or beliefs. Items that do not correspond to this pattern should be rearranged.

CALCULATING THE COEFFICIENT OF REPRODUCIBILITY. The coefficient of reproducibility is defined as the extent to which the total response pattern on a set of items can be reproduced even if the total score alone is known. This reproducibility depends on the extent to which the pattern of responses attains to a perfectly scalable pattern, demonstrated in Table 18.6. When the obtained coefficient of reproducibility is below the required .90 criterion, the scale needs to be refined until the coefficient of reproducibility reaches the desired level. The CR is calculated as follows:

\[
CR = 1 - \frac{\sum e}{Nr}
\]  

(18.4)

where

CR = the coefficient of reproducibility

\( \sum e \) = the total number of inconsistencies

Nr = the total number of responses (number of cases \( \times \) number of items)

A CR of .90 is the minimum standard for accepting a scale as unidimensional and cumulative.

Guttman Scale Application: An Example

After the researcher has developed and refined a Guttman scale, the results can be used to describe the distribution of the variable measured as well as to relate the scale to other variables in the study. Jules J. Wanderer’s study on the severity of riots in American cities is a particularly interesting application of the Guttman scaling technique because it is based on behavioral indicators rather than on attitudes. 9 Wanderer analyzed 75 reports of riots and criminal disorders that took place during the summer of 1967. The information Wanderer used in the construction of the scale was provided by municipalities at the request of a U.S. Senate subcommittee. The scale includes the following items of riot severity: killing, calling up of the National Guard, calling up of the state police, sniping, looting, interference with firefighters, and vandalism. These items are ordered from most to least severe and by the frequency with which they are reported. The coefficient of reproducibility of this Guttman scale of riot severity is .92. Cities are then organized into eight scale types according to the degree of severity of the riots reported, with 8 indicating the least severe and 1 the most severe riot activity. Table 18.7 presents the scale and the distribution of the cities along the scale.

At the second stage of the analysis, Wanderer treated riot severity, as measured by the Guttman scale, as a dependent variable. He then examined a set of independent variables in terms of their relationship to riot severity. For example, he found a relationship between the percentage of nonwhites participating in a riot and the degree of riot severity as measured by the scale: Once a riot has begun, the higher the percentage of nonwhites participating, the greater is the severity of the riot.

The Guttman scale of riot severity developed in this study suggests that the events that constitute riots and civil criminal disorders are neither erratic nor randomly occurring. On the contrary, by employing a Guttman scale, the researcher could predict the sequence of events that will take place at each level of riot severity.

Factor Analysis

Factor analysis is a statistical technique for classifying a large number of interrelated variables into a limited number of dimensions or factors. It is a useful method for

Table 18.7
A Guttman Scale of Riot Severity

<table>
<thead>
<tr>
<th>Scale Type</th>
<th>Percentage of Cities</th>
<th>Items Reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>4</td>
<td>No scale items</td>
</tr>
<tr>
<td>7</td>
<td>19</td>
<td>&quot;Vandalism&quot;</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>All of the above and &quot;interference with firefighters&quot;</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>All of the above and &quot;looting&quot;</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>All of the above and &quot;sniping&quot;</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>All of the above and &quot;called state police&quot;</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>All of the above and &quot;called National Guard&quot;</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>All of the above and &quot;law officer or civilian killed&quot;</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>


Constructing multiple-item scales, where each scale represents a dimension of a highly abstract concept. Factor analysis, by helping to identify the most powerful indicators of a concept, contributes to increasing the efficiency as well as validity of the research. Consider, for example, community satisfaction. Many statements or items can be used to describe community satisfaction: satisfaction with public schools, shopping facilities, personal safety, cultural amenities, the friendliness of the neighborhood, and so forth. However, the measurement of community satisfaction can be simplified by identifying the underlying dimensions of community satisfaction.

Some studies dealing with community satisfaction have indeed adopted this approach. Researchers divided community or neighborhood satisfaction into the following subconcepts (factors): (1) satisfaction with service delivery, (2) satisfaction with community organization, (3) satisfaction with neighborhood quality, and (4) satisfaction with cultural amenities. The relationship between the subconcepts and community satisfaction could then be expressed as follows:

Community satisfaction = $S($service delivery$) + S($community organization$) + S($neighborhood quality$) + S($cultural amenities$)

where $S =$ satisfaction.

In this formulation, community satisfaction is a construct represented by four basic factors. In factor analysis, the factors are not observed directly; rather, they are defined by a group of variables or items that are components of the abstract factors. The research actually begins by selecting a large number of items (employed as indicators) that may define each of the factors. These items are presented to respondents in the form of questionnaire statements.

In the first stage of factor analysis, the bivariate correlations (Pearson's $r$) between all the items are computed. The researcher then arranges the correlations in a matrix. The correlation matrix becomes input to the factor analysis procedure. The construction of factors is based on the identification of strong relations between a set of items. The method assumes that variables or items representing a single factor will be highly correlated with that factor.

The correlation between an item and a factor is represented by a factor loading. A factor loading is similar to a correlation coefficient; its values range between 0 and 1.0, and it can be interpreted in the same way. Table 18.8 presents the factor loadings of 14 items expressing community satisfaction on their four underlying factors. The items with the highest loading on each factor are underlined; these items are taken as the best indicators for these factors. Among the 14 items, only 3 have a high loading on factor 1. These items all refer to satisfaction with services; thus we can identify factor 1 as representing the dimension of service delivery. Similarly, factor 2 represents satisfaction with community organization; factor 3, quality of life; and factor 4, cultural amenities. Loadings of .30 or below are generally considered too weak to represent a factor. Examination of the results reveals that although all items display loadings on each of the factors, most item loadings are too weak to be considered as good indicators of the dimensions investigated.

Let us consider another example, sexual harassment. If we want to understand just what makes such behavior so threatening, it may not be sufficient merely to scale the severity of different acts. In attempting to devise an instrument that could measure the influence of the components underlying harassment, Maureen Murdock and Paul McGovern conducted a factor analysis of the items on their Sexual Harassment Inventory. They calculated the factor loadings and alpha coefficients of each item with

Table 18.8
Factor Loadings of Community Satisfaction Items

<table>
<thead>
<tr>
<th>Item Description</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Neighborness</td>
<td>.12361</td>
<td>.03216</td>
<td>.76182</td>
<td>.32101</td>
</tr>
<tr>
<td>2. Parks and playgrounds</td>
<td>.62375</td>
<td>.33610</td>
<td>.32101</td>
<td>.01260</td>
</tr>
<tr>
<td>3. Public schools</td>
<td>.74519</td>
<td>.34510</td>
<td>.12102</td>
<td>.01320</td>
</tr>
<tr>
<td>4. Shopping facilities</td>
<td>.32100</td>
<td>.06121</td>
<td>.68123</td>
<td>.12356</td>
</tr>
<tr>
<td>5. Police protection</td>
<td>.90987</td>
<td>.12618</td>
<td>.21361</td>
<td>.01320</td>
</tr>
<tr>
<td>6. Local churches</td>
<td>.21032</td>
<td>.75847</td>
<td>.21362</td>
<td>.11620</td>
</tr>
<tr>
<td>7. Church groups and organizations in the community</td>
<td>.01362</td>
<td>.82310</td>
<td>.01231</td>
<td>.11632</td>
</tr>
<tr>
<td>8. Community entertainment and recreational opportunities</td>
<td>.25617</td>
<td>.01320</td>
<td>.12341</td>
<td>.75624</td>
</tr>
<tr>
<td>9. Cultural activities</td>
<td>.16320</td>
<td>.12310</td>
<td>.32134</td>
<td>.02316</td>
</tr>
<tr>
<td>10. Quality of air</td>
<td>.02313</td>
<td>.11621</td>
<td>.03612</td>
<td>.32131</td>
</tr>
<tr>
<td>11. Noise level</td>
<td>.26154</td>
<td>.21320</td>
<td>.78672</td>
<td>.21368</td>
</tr>
<tr>
<td>13. Racial problems</td>
<td>.08091</td>
<td>.11320</td>
<td>.82316</td>
<td>.16342</td>
</tr>
<tr>
<td>14. Neighborhood pride</td>
<td>.18642</td>
<td>.11218</td>
<td>.71321</td>
<td>.18321</td>
</tr>
<tr>
<td>Percentage of variance</td>
<td>18.2</td>
<td>5.6</td>
<td>40.1</td>
<td>2.4</td>
</tr>
</tbody>
</table>
Table 18.9
Standardized Factor Score Coefficients

<table>
<thead>
<tr>
<th>Item</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.6812</td>
</tr>
<tr>
<td>10</td>
<td>.7234</td>
</tr>
<tr>
<td>11</td>
<td>.6916</td>
</tr>
<tr>
<td>12</td>
<td>.8162</td>
</tr>
<tr>
<td>13</td>
<td>.8110</td>
</tr>
<tr>
<td>14</td>
<td>.6910</td>
</tr>
</tbody>
</table>

respect to three underlying variables: hostile environment, quid pro quo behavior, and criminal sexual misconduct (the last being the most serious cases of harassment, i.e., rape and attempted rape). Taken together, the three factors accounted for 57 percent of the variance in the model. The authors intend to use these findings to predict the influence of different types of sexual harassment on the health of the victims, as well as the relationship between hypothesized risk factors and harassment.10

The extent to which each factor is expressed by the item’s loadings is reflected in the percentage of explained variance. Generally, factors with the highest percentage of explained variance provide the clearest representation of the items. That is, this factor can be employed fairly exclusively to represent the dimension studied. Table 18.8, which indicates the results of the loadings received by each factor in the community satisfaction study, reveals that the factor most clearly representing community satisfaction is quality of life (40.1 percent); the most ambiguous factor is cultural amenities (2.4 percent). Practically speaking, this means that items related to cultural activities (e.g., number of movie houses, frequency of public concerts) can be eliminated from the questionnaire without significantly reducing the quality of the research.

In the final step of factor analysis, the researcher develops a composite scale for each factor. For each observation or item, a factor score (scale score) is calculated. A factor score is an item’s score on a factor. It is obtained by using yet another type of coefficient, a factor score coefficient. To construct an item factor score, we multiply the factor score coefficients for each item by the standardized values of the variable obtained for that item. For example, Table 18.9 represents the factor score coefficients for the items that load on factor 3. We may construct an item’s factor score \( f_3 \), a composite scale representing factor 3, as follows:

\[
f_3 = .6812Z_1 + .7234Z_{10} + .6916Z_{11} + .8162Z_{12} + .8110Z_{13} + .6910Z_{14}
\]

\(Z_1\) through \(Z_{14}\) represent the standardized values of items 1 through 14 for that case.

In conclusion, factor analysis is an efficient method for reorganizing the items a researcher is investigating into conceptually more precise groups of variables.


**SUMMARY**

1. An index is a composite measure of two or more variables or items. Four major issues are involved in constructing indexes: the purpose for which the index is being compiled, the sources of the data, the base of comparison, and methods of aggregation and weighting. The Consumer Price Index (CPI), which is a composite measure of changes in retail prices, is one of the best-known and most frequently used indexes.

2. Scaling is a method of measuring the amount of a property possessed by a class of objects or events. It is most often associated with the measurement of attitudes. Attitude scales consist of statements expressing different attitudes, with which the respondent is asked to agree or disagree. Scaling techniques are applied to order the statements along some continuum. That is, they transform qualitative variables into a series of quantitative variables. The scales discussed in this chapter are either unidimensional or can be tested for unidimensionality. This means that the items comprising the scale can be arranged on a continuum presumed to reflect one and only one concept.

3. One technique for scale construction is Likert scaling. Likert scaling requires the researcher to compile a list of possible scale items, administer them to a random sample of respondents, compute a total score for each respondent, determine the discriminative power of each item, and only then select the final scale items.

4. Another method of scaling is the Guttman scaling technique. This method was designed to incorporate an empirical test of the unidimensionality of a set of items within the scale construction process. Guttman scale items are unidimensional—ranked on a single underlying dimension—as well as cumulative, in that information on any respondent’s last positive response allows the researcher to predict all of that person’s responses to the other items in the series. To measure the degree of conformity of a proposed scale to a perfectly scalable pattern, Guttman developed the coefficient of reproducibility (CR). A coefficient of reproducibility of .90 is the criterion used to decide whether a scale is indeed unidimensional.

5. Factor analysis is a statistical technique for classifying a large number of interrelated variables into a smaller number of factors. These factors represent a more complex dimension of a more abstract phenomenon. Factor analysis is a useful method for constructing multiple-item scales, where each scale represents a specific factor.

**KEY TERMS FOR REVIEW**

- coefficient of reproducibility (CR) (p. 426)
- discriminative power (DP) (p. 424)
- factor analysis (p. 427)
- factor loading (p. 429)
- factor score coefficient (p. 430)
- Guttman scale (p. 425)
- index (p. 415)
- item (p. 414)
- Likert scale (p. 422)
- simple aggregate (p. 418)
- unidimensionality (p. 415)
- weighted aggregate (p. 419)

**STUDY QUESTIONS**

1. What is the difference between a scale and an index? Can you identify commonly used examples of each?
2. Why do social scientists use scales and indexes in their research?
3. Develop an index to measure "popularity" among college students. Interview at least 10 persons (fellow students, teachers) to obtain the data for your scale. Use a method of aggregation with items of the type "How many times ... ?" "How often ... ?" "How many ... ?" Incorporate weighting procedures when appropriate. Discuss the problems of validity and reliability as they apply to your index. On the basis of your results, submit a revised index of popularity.
4. Based on the results to Question 3, suggest how the index can be converted into a scale.
5. When would factor analysis be preferable to scaling? State your methodological as well as theoretical reasons for choosing between them.

**SPSS PROBLEMS**

1. Construct a scale measuring abortion attitudes. Include the following variables in your scale: "abany," "abdelil," "abhhil," "abnom," "abpoor," "absingle." Generate Frequencies of your scale and determine the percentage of respondents who have the highest score on your scale. Describe their overall attitude toward abortion. How about those with the lowest score? Next, use SPSS to evaluate the reliability of your scale. What is the alpha coefficient for your scale?
2. Examine the relationship between church attendance ("attend") and the abortion scale constructed in Problem 1. Note that your new scale is an interval variable. Use the Bivariate Correlations procedure to analyze the relationship. What do you conclude?
3. Construct a scale measuring tolerance. Include the following variables in your scale: "colmill," "colhorno," "colrac." What is the alpha coefficient for your scale? Can you improve the alpha coefficient for the scale? How?

**ADDITIONAL READINGS**