From H. Russell Bernard (2013). *Social Research Methods: Qualitative and Quantitative Approaches* (2nd ed.) Thousand Oaks, CA: Sage

11

Scales and Scaling

IN THIS CHAPTER

This chapter is about building and using **composite measures**. I'll cover four kinds of composite measures: (1) indexes, (2) Guttman scales, (3) Likert scales, and (4) semantic differential scales. At the end of the chapter, I'll cover a few other interesting scales. First, though, some basic concepts of scaling.

SIMPLE SCALES: SINGLE INDICATORS

A scale is a device for assigning units of analysis to categories of a variable. The assignment is usually done with numbers, and questions are used a lot as scaling devices. Here are three typical scaling questions:

1. How old are you?

You can use this question to assign individuals to categories of the variable "age." In other words, you can *scale* people by age. The number that this first question produces has ratio properties (someone who is 50 is twice as old as someone who is 25).

2. How satisfied are you with your classes this semester? Are you satisfied, neutral, or unsatisfied?

You can use *this* question to assign people to one of three categories of the variable "satisfied." That is, you can *scale* them according to how satisfied they are with their classes. Suppose we let satisfied = 3, neutral = 2, and unsatisfied = 1. Someone who is assigned the number 3 is *more* satisfied than someone who is assigned the number 1. We don't know if that means 3 times more satisfied, or 10 times, or just marginally more satisfied, so this scaling device produces numbers that have ordinal properties.

3. Do you consider yourself to be Protestant, Catholic, Jewish, Muslim, some other religion? Or do you consider yourself as having no religion?

This scaling device lets you assign individuals to—that is, *scale them* by—categories of the variable "religious affiliation." Let Protestant = 1, Catholic = 2, Jewish = 3, Muslim = 4, and no religion = 5. The numbers produced by *this* device have nominal properties. You can't add them up and find the average religion.

These three questions have different content (they tap different concepts), and they produce numbers with different properties, but they have two very important things in common. All three questions are devices for scaling people and in all three cases the respondent is the principal source of measurement error.

When you use your own judgment to assign units of analysis to categories of a scaling device, *you* are the major source of measurement error. In other words, if you assign individuals by your own observation to the category "male" or "female," then any mistakes you make in that assignment (in scaling people by sex) are *yours*.

The same is true no matter what the unit of analysis is. Suppose you have a list of 100 countries and your job is to assign each to a category of government (parliamentary republic, constitutional monarchy, dictatorial monarchy, military dictatorship, etc.). For each country, you have some literature—scholarly books and articles, stories from the *New York Times*, reports from the U.S. State Department, and so on. You read each of these, looking for clues about the nature of governance, and assign a number (1, 2, 3, etc.) to each country. Each country is a unit of analysis and is scaled on the nominal variable called "predominant type of government."

Later, in the analysis, you might ask a question like: "Are democracies less likely to go to war with one another than, say, dictatorships?" Any mistakes you make in assigning the countries to a category of government will affect the relations you find (or don't find) in your analysis.

COMPLEX SCALES: MULTIPLE INDICATORS

A single question on a questionnaire is technically a scale if it lets you assign the people you're studying to categories of a variable. A lot of really interesting variables in social science, however, are complex and can't easily be assessed with single indicators. What single question could you ask someone to measure the amount of stress they are experiencing? Their overall political orientation, from far left to far right? How much they value physical attractiveness compared to other characteristics in potential marriage partners? How prejudiced they are against Asian immigrants on the job?

We try to measure complex variables like these with complex instruments—that is, instruments that are made up of several indicators. These complex instruments are what people commonly call scales.

A classic social science concept is "socioeconomic status" or SES. It is often measured by combining measures of income, education, and occupational prestige. Each of these measures is an operationalization of the concept SES, but none of the measures, by itself, captures the complexity of the idea of socioeconomic status. Each indicator captures a piece of the concept, and together the indicators produce a single measurement of SES (**Further Reading:** measuring SES).

Some variables are best measured by single indicators and, by Ockham's razor, we would never use a complex scale to measure something when a simple scale will do. So: The function of **single-indicator scales** is to assign units of analysis to categories of a variable. The function of composite measures, or **complex scales**, is exactly the same, but they are used when single indicators won't do the job.

INDEXES

The most common composite measure is a **cumulative index**. Indexes are made up of several items, all of which count the same. **Indexes** are everywhere. The Dow-Jones Industrial Average is a weighted index of the prices of 30 stocks that are traded on the New York Stock Exchange. The U.S. Consumer Price Index is a measure of how much it costs to buy a fixed set of consumer items in the United States. We use indexes to measure people's health risks: the risk of contracting HIV, of getting lung cancer, of having a heart attack, of giving birth to an underweight baby, of becoming an alcoholic, of suffering from depression, and on and on.

And we use indexes with a vengeance to measure cognitive and physical functions. Children in industrial societies of the world begin taking intelligence tests, achievement tests, and tests of physical fitness from the first day they enter school—or even before that. Achievement indexes—like the SAT, ACT, and GRE—affect so many people in the United States that there's a thriving industry devoted to helping children and adolescents do well on them.

Indexes can be **criterion referenced** or **norm referenced**. If you've ever taken a test where the only way to get an "A" was to get at least 90%, you've had your knowledge of some subject assessed by a criterion-referenced index. If you've ever taken a test where getting an "A" required that you score in the top 10% of the class—even if the highest grade in the class were 70%—then you've had your knowledge of some subject assessed by a norm-referenced index.

Standardized tests (whether of achievement, or of performance, or of personality traits) are usually norm referenced: Your score is compared to the norms that have been established by thousands of people who took the test before you.

How Indexes Work

Multiple-choice exams are cumulative indexes. The idea is that asking just one question about the material in a course would not be a good indicator of students' knowledge of the material. Instead, students typically are asked a bunch of multiple-choice questions.

Taken together, the reasoning goes, all the questions measure how well a student has mastered a body of material. If you take a test that has 60 multiple-choice questions and you get 45 correct, you get 45 points, one for each correct answer. That number, 45 (or 75% of 60 questions), is a cumulative index of how well you did on the test.

Note that in a cumulative index, it makes no difference *which* items are assigned to you. In a test of just 10 questions, for example, there are obviously just 10 ways to get one right—but there are 45 ways to get two right, 120 ways to get three right. . . . Students can get the same score of 80% on a test of 100 questions and miss entirely different sets of 20 questions. This makes cumulative indexes **robust**—that is, they provide many ways to get at an underlying variable (in the case of an exam, the underlying variable is knowledge of the material).

On the other hand, stringing together a series of items to form an index doesn't guarantee that the composite measure will be useful—any more than stringing together a series of multiple-choice questions will fairly assess a student's knowledge of, say, sociology or political science.

We pretend that: (1) knowledge is a **unidimensional** variable; (2) a fair set of questions is chosen to represent knowledge of some subject; and therefore (3) a cumulative index is a fair test of the knowledge of that subject. We know that the system is imperfect, but we pretend to get on with life.

We don't have to pretend. When it comes to scaling units of analysis on complex constructs—like scaling countries on the construct of freedom or people on the construct of political conservatism—we can test the unidimensionality of an index with a technique called **Guttman scaling**.

GUTTMAN SCALES

In a Guttman scale, as compared to a cumulative index, the measurements for the items have a *particular pattern indicating that the items measure a unidimensional variable*. To understand the pattern we're looking for, consider the following three questions.

1. How much is 124 plus 14?
2. How much is 1/2 + 1/3 + 1/5 + 2/11?
3. If 3X = 133, then how much is X?

If you know the answer to question 3, you probably know the answer to questions 1 and 2. If you know the answer to question 2, but not to 3, it's still safe to assume that you know the answer to question 1. This means that, in general, *knowledge about basic math* is a unidimensional variable (Goodenough 1944; Guttman 1944).

Suppose you're studying worker alienation in a factory. After running a focus group and talking to some of the union leaders, you decide on three indicators of alienation, each indicating increasing alienation: (1) signing a petition against new work rules (an expression of alienation that's backed up by support from others—you're with a crowd and not out there, on your own against management); (2) calling in sick a lot (now you're making a statement on your own, but not quite challenging management directly); and (3) filing a grievance against management (now it's really you against them, nowhere to hide, winner takes all, and the loser goes home).

If your hypothesis is correct, then worker alienation—in this factory, at this moment—is a unidimensional variable. It starts with signing a petition, and as it gets stronger, it is expressed by calling in sick and finally by filing a grievance. To test this, set up a table like Table 11.1 and assign each worker one point for each of these three indicators.

Respondents 1, 2, and 3 scored positive on all three items. The next three respondents (4, 5, and 6) signed the petition and call in sick regularly, but haven't filed any grievances. Respondents 7, 8, and 9 signed the petition but did not call in sick and did not file a grievance. And respondents 10, 11, and 12 have no alienation points on this scale. They have not signed the petition, have not called in sick a lot, and have not filed a grievance. So far so good.

The next three (13, 14, 15) have filed grievances but have neither signed the petition nor called in sick a lot. Finally, respondent 16 signed the petition and filed a grievance, but does not call in sick a lot.

Table 11.1 An Index That Scales With a Guttman Coefficient of Reproducibility < 0.90

Respondent	Signed a petition	Called in sick a lot	Filed a grievance
1	+	+	+
2	+	+	+
3	+	+	+
4	+	+	–
5	+	+	–
6	+	+	–
7	+	–	–
8	+	–	–
9	+	–	–
10	–	–	–
11	–	–	–
12	–	–	–
13	–	–	+
14	–	–	+
15	–	–	+
16	+	–	+

If we had data from only the first 12 respondents, the data would form a perfect Guttman scale. For those first 12 respondents, in other words, the three behaviors (signing a petition, calling in sick, and filing a grievance) are indicators of a unidimensional variable, worker alienation.

The Coefficient of Reproducibility

Unfortunately, we've got those other four respondents to deal with. For whatever reasons, respondents 13–16 do not conform to the pattern seen in respondents 1–12. The data for respondents 13–16 are "errors" in the sense that their data diminish the extent to which the index of alienation forms a perfect scale. To test how closely any set of index data reproduces a perfect scale, apply Guttman's (1944) **coefficient of reproducibility**, or CR. The formula for Guttman's CR is:

$$1 - \frac{\text{number of errors}}{\text{number of entries}} \qquad \textbf{formula 11.1}$$

Given the pattern in Table 11.1, we don't expect to see those plus signs in column 3 for respondents 13, 14, and 15. If the data scaled according to our hypothesis, then anyone who filed a grievance—anyone who has a plus in column 3—should have all pluses and a score of 3 on alienation. If we give respondents 13, 14, and 15 a scale score of 3 (for having filed a grievance), then those three cases would be responsible for *six* errors—you'd have to stick two pluses in for each of the cases to make them come out according to the hypothesis. Yes, you could make it just three, not six errors, by sticking a minus sign in column 3. Some researchers use this scoring method, but I prefer the more conservative method of scoring more errors (Goodenough 1944). It keeps you on your toes.

Finally, we don't expect that minus sign in column 2 of respondent 16's data. That case creates just one error (you only need to put in one plus to make it come out right). All together, that makes 6 + 1 = 7 errors in the attempt to reproduce a perfect scale. For Table 11.1, the CR is:

$$1-\frac{7}{48}=.85$$

which is to say that the data come within 15% of scaling perfectly. By convention, a coefficient of reproducibility of 0.90 or greater is accepted as a significant approximation of a perfect scale. Guttman (1944) recommended a coefficient of 0.85 or better, and I'm willing to settle for that, especially with the conservative method for scoring errors.

Some Examples of a Guttman Scale

Christopher Mooney and Mei-Hsien Lee (1995) studied the history of abortion law reform. The 1973 *Roe v. Wade* decision by the U.S. Supreme Court made abortion on demand a woman's right in all 50 states. This decision didn't just happen all at once.

Before 1973, abortion was outlawed in all 50 states, but it was legal in some states when carrying the fetus to term was a threat to the woman's life. Beginning in the 1950s, various groups began working to get states to enact regulation reform and to make abortion legal under more and more circumstances. The first state in this era to enact regulation reform was Mississippi in 1966. By the time the *Roe v. Wade* decision came down, 17 other states had enacted some kind of legislation reforming the regulation of abortion.

Mooney and Lee (1995) found that the laws in these 18 states formed a perfect Guttman scale, based on four successively more liberal conditions. In addition to cases involving threats to the woman's life, abortion would be legal: (1) when the pregnancy resulted from rape or incest; (2) when the fetus was defective or there was a risk to the woman's physical health; (3) when there was a threat to the woman's mental health; and (4) whenever a woman decided she wanted one.

Table 11.2 shows the data. When you collect data on cases, you don't know what (if any) pattern will emerge, so you pretty much grab cases and code them for traits in random order. If you grabbed cases in chronological order, for example, you wouldn't see the perfect pattern of pluses and minuses in Table 11.2.

When you have the data in a table, the first thing to do is arrange the pluses and minuses in their "best" order—the order that conforms most to the perfect Guttman scale—and compute the CR. We look for the trait that occurs most frequently (the one with the most pluses across the row) and place that one at the bottom of the matrix. Then we look for the next-most-frequent trait, and put it on the next-to-the-bottom row of the matrix.

We keep doing this until we rearrange the data to take advantage of whatever underlying pattern is hiding in the matrix. Then we count up the "errors" in the matrix and compute Guttman's coefficient of reproducibility. For these 18 states and four traits, the coefficient is a perfect 1.0, and all 18 cases can be ranked on the degree of permissiveness regarding abortion. Obviously, if a state allows abortion on demand, it allows it in all specific cases, so it gets a scale score of 4. If a state allows abortion in cases where the woman's mental health is at risk, then it allows abortion in cases where the woman's physical health is at risk and in cases of rape or incest, so it gets a scale score of 3; and so on. (The actual work of arranging

Table 11.2 A Guttman Scale of Abortion Law During the 1960s and 1970s for 18 States in the U.S.

State	Year of Reform	Rape or Incest	Defect in Fetus or Threat to Woman's Physical Health	Threat to Woman's Mental Health	On Demand	Scale Score on Permissiveness
MS	1966	+	–	–	–	1
AR	1969	+	+	–	–	2
FL	1972	+	+	–	–	2
GA	1968	+	+	–	–	2
CA	1967	+	+	+	–	3
CO	1967	+	+	+	–	3
NC	1967	+	+	+	–	3
MD	1968	+	+	+	–	3
DE	1969	+	+	+	–	3
KS	1969	+	+	+	–	3
NM	1969	+	+	+	–	3
SC	1970	+	+	+	–	3
VA	1970	+	+	+	–	3
OR	1969	+	+	+	–	3
AK	1970	+	+	+	+	4
HA	1970	+	+	+	+	4
NY	1970	+	+	+	+	4
WA	1970	+	+	+	+	4

Source: Constructed from data in C. Z. Mooney and M-H. Lee, "Legislating Morality in the American States: The Case of Pre-Roe Abortion Regulation Reform." *American Journal of Political Science* 39:599–627. Copyright © 1995.

the data and counting the errors is done by computer. (See, for example, Anthropac, Appendix E.)

What this means is that when it came to abortion, permissiveness during the 1960s and early 1970s in the United States was a unidimensional variable. That's nice to know, but there's more. The scale scores in Table 11.2 have a Spearman's rank-order correlation of 0.44 with the year of reform and this correlation was statistically significant (see Chapter 21 for more on correlation). This is support for the theory of incremental policy reform in political science. We see this process of incremental reform at work, for example, in acceptance, over time, of same-sex marriage and of medical marijuana in the various states in the United States.

According to the theory, as pressure builds for some reform, one or two states lead the way with tentative steps. Other states hang back and watch the results. Then there is a rush of states that follow and the steps are less tentative. Finally, all the states that are going to take the steps have done so and the process goes back to a slow pace again, as the remaining states hang back and assess the situation some more.

The process is evident in Table 11.2. Mississippi led off with a small step in 1966. By 1970, 17 states had enacted reform, but it would take two more years before the 18th state, Florida, would join, and that state reversed the trend to more and more liberal reform by enacting less permissive legislation than the 10 states before it had done (Box 11.1).

Box 11.1 The Bogardus Social Distance Scale

An early example of a Guttman scale is the **Bogardus Social Distance Scale**, developed by Emory Bogardus in 1925. Since Guttman didn't describe his method for testing the unidimensionality of a scale until 1944, the Bogardus scale is not usually *called* a Guttman scale, but a Guttman scale it is, nevertheless.

Bogardus showed people names of ethnic groups and asked them, for each group, which of the following seven opinions they agreed with most: "I would be willing to accept members of this group: (1) as kin through marriage; (2) as personal friends; (3) as neighbors; (4) as co-workers; (5) as citizens of their country; (6) only as visitors to their country; (7) under no condition, not even as visitors to my country."

Some version of this scale has been used in dozens of studies over the years, so there is now a substantial literature on racial and ethnic distance (**Further Reading:** Bogardus Social Distance Scale).

Data Scale, Variables Don't

Remember, *only data scale, not variables*. That is, scales like these are **sample dependent**. If the items in a cumulative index form a strong Guttman scale, we can say that, *for the sample we've tested*, the concept measured by the index is unidimensional—that the items are a composite measure of one and only one underlying concept.

Billie DeWalt (1979) used Guttman scaling to test an index of material style of life in a Mexican farming community. He scored 54 people on whether they owned eight material items (a radio, a stove, a sewing machine, etc.) and achieved a CR of 0.95. My hunch is that DeWalt's material-style-of-life scale has its analog in nearly all societies. The particular list of items that DeWalt used in rural Mexico may not scale in an African American community in the American South (Dressler et al. 1985), but *some* list of material items *will* scale there. You just have to find them.

The way to do this is to code every household in your study for the presence or absence of a list of material items. The particular list could emerge from participant observation or from informal interviews. Then you'd use a program like Anthropac to sort out the matrix, drop some material items, and build the index until it has a CR of 0.90 or better (Box 11.2).

Box 11.2 Indexes that don't scale

Indexes that do not scale can still be useful in comparing populations. Dennis Werner (1985) studied psychosomatic stress among Brazilian farmers who were facing the uncertainty of having their lands flooded by a major dam. He used a 20-item stress index developed by Berry (1976).

Since the index did not constitute a unidimensional scale, Werner could not differentiate among his *informants* (in terms of the amount of stress they were under) as precisely as DeWalt could differentiate among *his* informants (in terms of their quality of life). But farmers in Werner's sample gave a stress response to an average of 9.13 questions on the 20-item test, while Berry had found that Canadian farmers gave stress responses on an average of 1.79 questions. It is very unlikely that a difference of such magnitude between two *populations* would occur by chance (**Further Reading:** Guttman scaling).

LIKERT SCALES

Perhaps the most commonly used form of scaling is attributed to Rensis Likert (1932). Likert introduced the ever-popular five-point scale that we talked about in Chapter 9 on questionnaire construction. Recall that a typical question might read as follows:

> Please consider the following statements carefully. After each statement, check the answer that most reflects your opinion. Would you say you agree a lot with the statement, agree a little, are neutral, disagree a little, or disagree a lot with each statement? Ok, here's the first statement:
>
> Congress is doing all it can to prevent another financial crisis.
>
> □ Agree a lot
>
> □ Agree
>
> □ Neutral
>
> □ Disagree a little
>
> □ Disagree a lot

The five-point scale might become three points or seven points, and the agree-disagree scale may become approve-disapprove, favor-oppose, or excellent-bad, but the principle is the same. These are all **Likert-type** scales.

I say "Likert-type scales" rather than just "Likert scales" because Likert did more than just introduce a format. He was interested in measuring internal states of people (attitudes, emotions, orientations) and he realized that most internal states are multidimensional. It's easy to label people as either conservatives or liberals, but the concept of political orientation is very complex. A person who is liberal on matters of domestic policy—favoring single-payer, government-run health care, for example—may be conservative on matters of foreign political policy—against involvement in any foreign military actions. Someone who is liberal on matters of foreign economic policy—favoring economic aid for all democracies that ask for it—may be conservative on matters of personal behavior—against same-sex marriage, for example.

The liberal-conservative dimension on matters of personal behavior is also complicated. There's no way to assign people to a category of this variable by asking one question. People can have live-and-let-live attitudes about sexual preference and extramarital sex and be against a woman's right to an abortion on demand.

Of course, there are packaging effects. People who are conservative on one dimension of political orientation are *likely* to be conservative on other dimensions, and people who are liberal on one kind of personal behavior are *likely* to be liberal on others. Still, no single question lets you scale people in general on a variable as complex as "attitude toward personal behavior," let alone "political orientation." That's why we need composite scales.

Steps in Building a Likert Scale

Likert's method was to take a long list of possible scaling items for a concept and find the subsets that measured the various dimensions. If the concept were unidimensional, then one subset would do. If it were multidimensional, then several subsets would be needed. Here are the steps in building and testing a Likert scale.

1. Identify and label the variable you want to measure. This is generally done by induction—that is, from your own experience (Spector 1992:13). After you work in some area of research for a while, you'll develop some ideas about the variables you want to measure. The people you talk to in focus groups, for example, may impress you with the idea that "People are afraid of crime around here," and you decide to scale people on the variable "fear of crime."

Or you may observe that some people love to poke around for hours in malls, while others prefer using the Internet to buy all their clothes, gifts, etc. Some people seem to have a black belt in shopping, while others would rather have root canal surgery than set foot in a mall. The task is then to scale (measure) people on a variable you might call "shopping orientation" with all its multidimensionality. You may need a subscale for "shopping while on vacation," another just for "car shopping," and another for "shopping for clothing that I really need." (The other way to identify variables is by deduction [see Box 1.3]. This generally involves analyzing similarity matrices, about which more in Chapters 15 and 16.)

2. Write a long list of indicator questions or statements. This is usually another exercise in induction. Ideas for the indicators can come from reading the literature on whatever research problem has captured you, from personal experience, from ethnography, from reading newspapers, from interviews with experts.

Free lists are a particularly good way to get at indicators for some variables. If you want to build a scaling device for the concept of "attitudes toward growing old," you could start by asking a large group of people to "list things that you associate with growing old" and then you could build the questions or statements in a Likert scale around the items in the list.

Be sure to use both negative and positive indicators. If you have a statement like "One of the great things about this university is the emphasis on consistently winning sports teams," then you need a negatively worded statement for balance, like "One of the bad things about this university is the emphasis they put on sports."

And don't make the indicator items extreme. Here's a badly worded item: "The emphasis on sports is the most terrible thing that has ever happened here." Let people tell *you* where they stand by giving them a range of response choices (strongly agree–strongly disagree). Don't bludgeon people with such strongly worded scale items that they feel forced to reduce the strength of their response.

In wording items, all the cautions from Chapter 9 on questionnaire design apply: Remember who your respondents are and use *their* language. Make the items as short and as uncomplicated as possible. No double negatives. No **double-barreled items**. Here is a terrible item:

> On a scale of 1–5, how much do you agree or disagree with the following statement:
>
> "People should speak English and give up any language they brought with them when they came to this country."

People can agree or disagree with both parts of this statement, or agree with one part and disagree with the other. When you get through, you should have four or five times the number of items as you think you'll need in your final scale. If you want a scale of, say, six items, use 25 or 30 items in the first test (DeVellis 2003:66).

3. Determine the type and number of response categories. Some popular response categories are agree-disagree, favor-oppose, helpful–not helpful, many-none, like me–not like me, true-untrue, suitable-unsuitable, always-never, and so on. Most Likert scale items have an odd number of response choices: three, five, or seven. The idea is to give people a range of choices that includes a midpoint. The midpoint usually carries the idea of neutrality—neither agree nor disagree, for example. An even number of response choices forces informants to "take a stand"; an odd number of choices lets informants "sit on the fence."

There is no best format. But if you ever want to combine responses into just two categories (yes-no, agree-disagree, like me–not like me), then it's better to have an even number of choices. Otherwise, you have to decide whether the neutral responses get collapsed with the positive answers or the negative answers—or thrown out as missing data.

4. Test your item pool on some respondents. Ideally, you need at least 100—or even 200—respondents to test an initial pool of items (Spector 1992:29). This will ensure that: (1) you capture the full variation in responses to all your items; and (2) the response variability represents the variability in the general population to which you eventually want to apply your scale.

5. Conduct an **item analysis**—coming right up—to find the items that form a unidimensional scale of the variable you're trying to measure.

6. Use your scale in your study and run the item analysis again to make sure that the scale is holding up. If it does, then look for relations between the scale scores and the scores of other variables for persons in your study.

ITEM ANALYSIS

This is the key to building scales. The idea is to find out which, among the many items you're testing, need to be kept and which should be thrown away. The set of items that you keep should tap a single social or psychological dimension. In other words, the scale should be unidimensional.

In the next few pages, I'm going to walk through the logic of building scales that are unidimensional. Read these pages very carefully. At the end of this section, I'll advocate using **factor analysis** to do the item analysis quickly, easily, and reliably. No fair, though, using factor analysis for scale construction until you understand the logic of scale construction itself.

There are three steps to doing an item analysis and finding a subset of items that constitute a unidimensional scale: (1) **scoring the items**; (2a) taking the **interitem correlation** and (2b) calculating **Cronbach's coefficient alpha**; and (3) taking the **item-total correlation**.

Scoring the Responses

The first thing to do is make sure that all the items are properly scored. Assume that we're trying to find items for a scale that measures the strength of support for lots of training in research methods among sociology students. Here are two potential scale items:

> Training in multivariate statistics should be required for all undergraduate students of social science.

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Social science undergraduates don't need training in multivariate statistics.

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

You can let the big and small numbers stand for any direction you want, but you must be consistent. Suppose we let the bigger numbers (4 and 5) represent support for training in multivariate statistics and let the smaller numbers (1 and 2) represent lack of support for that concept. Respondents who circle "strongly agree" on the first item get a 5 for that item. Those who circle "strongly agree" on the second item get scored as 1.

Taking the Interitem Correlation

Next, test to see which items contribute to measuring the construct you're trying to get at and which don't. This involves two calculations: the intercorrelation of the items and the correlation of the item scores with the total scores for each respondent. Table 11.3 shows the scores for three people on three items, where the items are scored from 1 to 5.

Table 11.3 The Scores for Three People on Three Likert Scale Items

Person	Item		
	1	2	3
1	1	3	5
2	5	2	2
3	4	1	3

To find the interitem correlation, we would look at all pairs of columns. These are shown in Table 11.4.

A simple measure of how much these pairs of numbers are alike or unalike involves, first, adding up their *actual differences*, Σ_d, and then dividing this by the total *possible differences*, $\max_d$.

In the first pair, the actual difference between 1 and 3 is 2; the difference between 5 and 2 is 3; the difference between 4 and 1 is 3. The sum of the differences is $\Sigma_d = 2 + 3 + 3 = 8$.

For each item, there could be as much as 4 points difference—in Pair 1, someone could have answered 1 to item 1 and 5 to item 2, for example. So for three items, the total possible difference, $\max_d$, would be $4 \times 3 = 12$. The

Table 11.4 The Data From the Three Pairs of Items in Table 11.3

Pair 1		Diff	Pair 2		Diff	Pair 3		Diff
1	3	2	1	5	4	3	5	2
5	2	3	5	2	3	2	2	0
4	1	3	4	3	1	1	3	2
$\sum_d$(Sum of the differences)		8			8			4
$\sum_d/\max_d$		0.67			0.67			0.33
$1-\{\sum_d/\max_d\}$		0.33			0.33			0.67

actual *difference* is 8 out of a possible 12 points, so items 1 and 2 are 8/12 = 0.67 *different*, which means that these two items are $1 - \Sigma_d / \max_d = 0.33$ *alike*. Items 1 and 3 are also 0.33 alike, and items 2 and 3 are 0.67 alike.

Items that measure the same underlying construct should be related to one another. If I answer "strongly agree" to the statement "Training in multivariate statistics should be required for all undergraduate students of sociology," then (if I'm consistent in my attitude and if the items that tap my attitude are properly worded) I should strongly disagree with the statement that "sociology undergraduates don't need training in multivariate statistics." If everyone who answers "strongly agree" to the first statement answers "strongly disagree" to the second, then the items are perfectly correlated.

Cronbach's Alpha

Cronbach's alpha is a statistical test of how well the items in a scale are correlated with one another. One of the methods for testing the unidimensionality of a scale is called **the split-half reliability test**. If a scale of, say, 10 items, were unidimensional, all the items would be measuring parts of the same underlying concept. In that case any five items should produce scores that are more or less like the scores of any other five items. This is shown in Table 11.5.

Split Halves and the Combinations Rule

There are many ways to split a group of items into halves and each split will give you a different set of totals. Here's the formula, known as **the combinations rule**, for selecting n elements from a set of N elements, paying no attention to the ordering of the elements:

$$\frac{N!}{n!(N-n)!} \qquad \textbf{formula 11.2}$$

Table 11.5 The Schematic for the Split-Half Reliability Test

Person	Split A: Score on items 1-5	Split B: Score on items 6-10
1	X_1	Y_1
2	X_2	Y_2
3	X_3	Y_3
.	.	.
.	.	.
N	X_n	Y_n
	Total for A	Total for B

If you have 10 respondents, then there are 10!/5![(10 – 5)!] = 252 ways to split them into halves of five each. For 20 items, there are 184,756 possible splits of 10 each. Cronbach's coefficient alpha provides a way to get the average of all these split-half calculations directly. The formula for Cronbach's alpha is:

$$\alpha = \frac{N_\rho}{1 + \rho(N-1)} \qquad \textbf{formula 11.3}$$

where ρ (the Greek letter *rho*) is the average interitem correlation—that is, the average correlation among all pairs of items being tested.

By convention, a good set of scale items should have a Cronbach's alpha of 0.80 or higher. Be warned, though, that if you have a long list of scale items, the chances are good of getting a high alpha coefficient. An interitem correlation of just 0.29 produces an alpha of 0.80 in a set of 10 items (DeVellis 2003:98).

Eventually, you want an alpha coefficient of 0.80 or higher for a *short* list of items, all of which hang together and measure the same thing. Cronbach's alpha will tell you if your scale hangs together, but it won't tell you which items to throw away and which to keep.

To do that, you need to identify the items that do not discriminate between people who score high and people who score low on the total set of items.

Finding the Item-Total Correlation

First, find the total score for each person. Add up each respondent's scores across all the items. Table 11.6 shows what it would look like if you tested 50 items on 200 people (each x is a score for one person on one item).

For 50 items, scored from 1 to 5, each person could get a score as low as 50 (by getting a score of 1 on each item) or as high as 250 (by getting a score of 5 on each item). In practice, each person in a survey will get a total score somewhere in between.

A rough and ready way to find the items that discriminate well among respondents is to divide the respondents into two groups, the 25% with the highest total scores and the 25% with the lowest total scores. Look for the items that the two groups have in common. Those items are *not discriminating* among informants with regard to the concept being tested. Items that fail, for example, to discriminate between people who strongly favor training in methods (the top 25%) and people who don't (the bottom 25%) are not good items for scaling people in this construct. Throw those items out.

There is a more formal way to find the items that discriminate well among respondents and the items that don't. This is the item-total correlation. Table 11.7 shows the data you need for this:

Table 11.6 Finding the Item-Total Correlation

Person	Item 1	Item 2	Item 3	.	.	Item 50
1	x	x	x	.		x
2	x	x	x	.	.	x
3	x	x	x	.	.	x
.	.	.	.	.	.	.
.	.	.	.	.	.	.
200	x	x				x

Table 11.7 The Data for the Interitem Correlation

Person	Total Score	Item 1	Item 2	Item 3	.	.	50
1	x	x	x	x	.	.	x
2	x	x	x	x	.	.	x
3	x	x	x	x	.	.	x
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
N	x	x	x	x	.	.	x

With 50 items, the total score gives you an idea of where each person stands on the concept you're trying to measure. If the interitem correlation were perfect, then every item would be contributing equally to our understanding of where each respondent stands. Some items do better than others. The ones that don't contribute a lot will correlate poorly with the total score for each person. Keep the items that have the highest correlation with the total scores.

You can use any statistical analysis package to find the interitem correlations, Cronbach's alpha, and the item-total correlations for a set of preliminary scale items. Your goal is to get rid of items that detract from a high interitem correlation and to keep the alpha coefficient above 0.80. (For an excellent step-by-step explanation of item analysis, see Spector 1992:43–46).

TESTING FOR UNIDIMENSIONALITY WITH FACTOR ANALYSIS

Factor analysis is a technique for data reduction. If you have 30 items in a pool of potential scale items and responses from a sample of people to those pool items, factor analysis lets you reduce the 30 items to a smaller set—say, five or six. Each item is given a score, called its **factor loading**. This tells you how much each item "belongs" to each of the underlying factors.

If a scale is unidimensional, there will be a single factor that underlies all the variables (items) and all the items will "load high" on that single factor. If a scale is multidimensional, then there will be a series of factors that underlie sets of variables. Scale developers get a large pool of potential scale items (at least 40) and ask a lot of people (at least 200) to respond to the items. Then they run the factor analysis and select those items that load high—typically, 0.35–0.60—on the factor or factors (the underlying concept or concepts) they are trying to understand. They also test their results—their scale—on a new sample and refine their scale questions over time. Here's an example.

Morokoff's Sexual Assertiveness Scale for Women

Patricia Morokoff and her colleagues at the University of Rhode Island used factor analysis to develop a scale of sexual assertiveness in women (Morokoff et al. 1997). Morokoff et al. hypothesized three dimensions to this variable: (1) women varied in their ability to initiate wanted sex; (2) women varied in their ability to refuse unwanted sex; and (3) women varied in their ability to protect themselves from pregnancy and sexually transmitted disease (by demanding that the man use a condom).

Morokoff et al. made up several questionnaire items for each of nine sexual behaviors in women: kissing, touching of breasts, touching by partner of genitals, touching of partner's genitals, receiving oral sex, performing oral sex, vaginal intercourse, anal intercourse, and protecting themselves against pregnancy or disease by asking a partner to use a condom. For each behavior, the questionnaire items covered the three dimensions and the presence or absence of pressure.

For example, for kissing, when the woman was *not* under a lot of external pressure to give in, Morokoff et al. had items like: "I feel comfortable refusing to kiss a partner when I don't want to" and "If a partner wants to kiss and I don't want to, we do it anyway." For kissing when the woman *is* under a lot of external pressure to give in, they had items like: "If a partner pressures me to kiss him after I have refused, I continue to refuse" and "If I refused to kiss a partner and he continued to pressure me, I would give in."

There were similar items for touching of genitals, vaginal intercourse, demanding the use of a condom, and so on. All in all, they had 112 items about self-reported sexual behavior. The items were rated by respondents on a five-point scale: never, sometimes (about 25% of the time), about 50% of the time, usually (about 75% of the time), and always (100% of the time). Morokoff et al. also had 24 items about attitudes, and these, too, were rated on a five-point scale: disagree strongly, disagree, mixed, agree somewhat, agree strongly.

The full 136-item test was given to 260 women. The results (a matrix of 260 women by 136 responses) was factor analyzed. The analysis isolated 42 items that loaded 0.45 or higher on each of the three factors: 17 items for the *initiation* factor, 14 items for the *refusal* factor, and 11 items for the *pregnancy-STD prevention* factor (the condom factor).

Morokoff et al. next gave the 42-item questionnaire to an entirely different sample of 136 women. They used factor analysis and item-total correlation on the results of the second sample to winnow the 42 items down to just 18, with six items for each of the three subscales.

Morokoff et al. went on to test their scale on 752 more women to improve the language of the questions (they substituted the words "begin sex" for "initiate sex," for example) and on 354 women from their original sample during a one-year follow-up to see how the subscales held up. Consistently, women reported being less assertive in refusing unwanted sex and in demanding the use of a condom when they anticipated a negative reaction from their partners (Morokoff et al. 1997:802). Morokoff et al.'s final scale is shown in Table 11.8.

All of that work gives Morokoff et al.'s sexual assertiveness scale credibility and, indeed, the scale has been used by other researchers for whom this variable is important. Jacobs and Thomlinson (2009), for example, used the scale in their study of how self-silencing increased the risk of sexually acquired HIV in women over 50. You may not develop major scales for others to use but you *should* test the unidimensionality of any composite measure you develop for your

Table 11.8 Sexual Assertiveness in Women Scale

Items that are reverse-coded are indicated by (R). Notice that in all three subscales, half the items are reverse-coded. In other words, the same concepts are tested with items that are worded positively and with items that are worded negatively. The factor loadings for each of these 18 items were all above .55 in two separate studies.

Initiation

1. I begin sex with my partner if I want to.
2. I let my partner know if I want my partner to touch my genitals.
3. I wait for my partner to touch my genitals instead of letting my partner know that's what I want. (R)
4. I wait for my partner to touch my breasts instead of letting my partner know that's what I want. (R)
5. I let my partner know if I want to have my genitals kissed.
6. Women should wait for men to start things like breast touching. (R)

Refusal

7. I give in and kiss if my partner pressures me, even if I already said no. (R)
8. I put my mouth on my partner's genitals if my partner wants me to, even if I don't want to. (R)
9. I refuse to let my partner touch my breasts if I don't want that, even if my partner insists.
10. I have sex if my partner wants me to, even if I don't want to. (R)
11. If I said no, I won't let my partner touch my genitals even if my partner pressures me.
12. I refuse to have sex if I don't want to, even if my partner insists.

Protection against pregnancy-STD

13. I have sex without a condom or latex barrier if my partner doesn't like them, even if I want to use one. (R)
14. I have sex without using a condom or latex barrier if my partner insists, even if I don't want to. (R)
15. I make sure my partner and I use a condom or latex barrier when we have sex.
16. I have sex without using a condom or latex barrier if my partner wants. (R)
17. I insist on using a condom or latex barrier if I want to, even if my partner doesn't like them.
18. I refuse to have sex if my partner refuses to use a condom or latex barrier.

Source: P. J. Morokoff et al., "Sexual Assertiveness Scale (SAS) for Women: Development and Validation." *Journal of Personality and Social Psychology* 73:790–804. Copyright © 1997 by the American Psychological Association.

own data, using factor analysis—once you understand the principles of scale development that I've laid out here. I'll show you how to do that in Chapter 22, when we get to factor analysis (**Further Reading:** Likert scaling) (Box 11.3).

Box 11.3 Scales get simpler and more widely useful over time

Notice how the scale that Morokoff and her colleagues developed got simpler as it went through a couple of tests. They started with 136 items. These were reduced to 61 (the ones that scored over 0.45 on the first factor analysis). They removed 19 of those 61 items that were redundant (they were more or less rewordings of the same thing). The remaining 42 items were reduced to 18 in the next phase of the research when they gave the test to the next sample. In building scales, researchers err on the side of using too many questions rather than too few—no sense in leaving out some items that *may* be important until you know that you can do without them. But as scales are tested and retested, researchers often find that some items are redundant and the scales get shorter.

The original **Michigan Alcoholism Screening Test (MAST)**, for example, has 25 items (Selzer 1971). Pokorny et al. (1972) showed that their 10-item **Brief-MAST** instrument was about as effective as the longer original. Shields et al. (2007) found 454 published studies that used some version of the MAST through 2005. There's a 24-question version that's just for the elderly. It's called the **MAST-G**, where G stands for "geriatric version" (Blow et al. 1992). It turns out that "Yes" answers to *just two* of the 24 questions ("When talking with others, do you ever underestimate how much you actually drink?" and "Are you drinking more now than in the past?") predict hazardous levels of drinking in old people as well as the full, 24-question test (Johnson-Green et al. 2009).

The original, 1970 version of the **Attitudes Toward Women Scale (AWS)** had 55 items (Spence and Helmreich 1972). It was down to 25 items a year later (Spence et al. 1973) and down to 15 items five years after that (Spence and Helmreich 1978). That 15-item AWS is still measuring a unidimensional variable—what people think women's rights should be—and, while attitudes are becoming more liberal/feminist all around, women are still more supportive than men are of full equality (Twenge 1997; Whatley 2008).

(Continued)

(Continued)

Besides getting shorter, scales also get validated on new populations over time. Henderson et al. (1980) developed a 50-item test of social support (called the ISSI, or **Interview Schedule for Social Interaction**) on respondents in Canberra, Australia. Undén and Orth-Gomér (1989) tested and validated a much-reduced version of the scale on Swedish men who were at risk for heart disease, and the reduced ISSI in Swedish was validated on sample of men and women in Sweden who had been diagnosed with mental illness (Eklund et al. 2007).

Revalidations and reformulations of scales are published, so it pays to do a thorough search for existing scales before launching out on your own to build new scales from scratch.

VISUAL PROPS AS SCALES

Several scales have been developed over the years with visual props. Four of them are the semantic differential, the ladder of life, the happiness stick, and the faces scale.

The Semantic Differential

I've always liked the semantic differential scaling method. It was developed in the 1950s by Charles Osgood and his associates at the University of Illinois (Osgood et al. 1957; Snider and Osgood 1969) and since then has been used by thousands of researchers across the social sciences. With good reason: The semantic differential test is easy to construct and easy to administer.

Osgood was interested in how people interpret things—inanimate things (like artifacts or monuments), animate things (like persons, or the self), behaviors (like incest, or buying a new car or shooting a deer), and intangible concepts (like gun control or literacy). This is exactly what Likert scales are designed to test, but instead of asking people to rate questionnaire items about things, Osgood tested people's feelings differently: He gave them a **target item** and a list of **paired adjectives** about the target. The adjective pairs could come from reading of the literature or from focus groups or from ethnographic interviews. Target items can be ideas (land reform, socialism, aggression), behaviors (smoking, running, collecting stamps), objects (the mall, a courtroom, horses), environmental conditions (rain, drought, jungle) . . . almost anything.

Figure 11.1 is an example of a semantic differential test. The target is the concept of "abortion on demand." If you were taking this test right now, you'd be asked to place a check on each line, depending on your reaction to each pair of adjectives.

With a Likert scale, you ask people a series of questions that get at the target concept. In a semantic differential scale, you name the target concept and ask people to rate their feelings toward it on a series of variables. The semantic differential is usually a seven-point scale, as I've indicated in Figure 11.1. Your score on this test would be the sum of all your answers to the 13 adjective pairs.

Osgood and his associates did hundreds of replications of this test, using hundreds of adjective pairs, in 26 different cultures. Their analyses showed that in every culture just three major kinds of adjectives account for most of the variation in people's responses: **adjectives of evaluation** (good-bad, difficult-easy), followed by **adjectives of potency** (strong-weak, dominant-submissive, etc.), and **adjectives of activity** (fast-slow, active-inactive, sedentary-mobile, etc.).

Figure 11.1 Semantic Differential Test for the Concept of a Woman's Right to Abortion on Demand. The Dimensions on This Scale Are Useful for Measuring How People Feel About Many Different Things

Hard	___	___	___	___	___	___	___	Easy
	1	2	3	4	5	6	7	
Active	___	___	___	___	___	___	___	Passive
Difficult	___	___	___	___	___	___	___	Easy
Permanent	___	___	___	___	___	___	___	Impermanent
Warm	___	___	___	___	___	___	___	Cold
Beautiful	___	___	___	___	___	___	___	Ugly
Strong	___	___	___	___	___	___	___	Weak
Reassuring	___	___	___	___	___	___	___	Unsettling
Important	___	___	___	___	___	___	___	Trivial
Fast	___	___	___	___	___	___	___	Slow
Clean	___	___	___	___	___	___	___	Dirty
Exciting	___	___	___	___	___	___	___	Boring
Useful	___	___	___	___	___	___	___	Useless

As the target for a semantic differential scale changes, you have to make sure that the adjective pairs make sense. The adjective pair ethical-corrupt works for some targets, but you probably wouldn't use it for having a cold. Indoor-outdoor works for lots of targets—kinds of music, hobbies, even famous persons—but it's not appropriate for targets like patio furniture, preservation of wilderness, or hang gliding, which are, by definition, outdoor things.

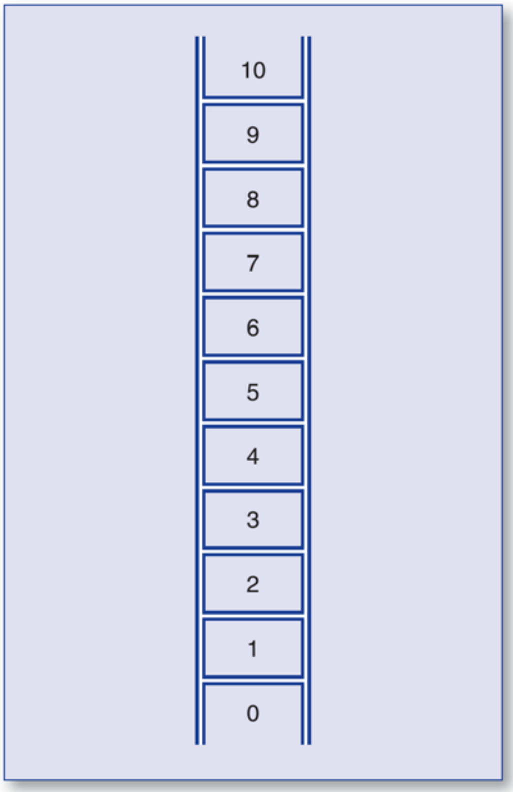
Vincke et al. (2001) used the semantic differential scale to explore the meaning of 25 sex acts among gay men in Flanders, Belgium. Their informants scaled each act (anal insertive sex, anal receptive sex, insertive fellatio, receptive fellatio, interfemoral sex, and so on) on six paired dimensions: unsatisfying/satisfying, stimulating/dull, interesting/boring, emotional/unemotional, healthy/unhealthy, and safety/danger. Vincke et al. then compared results on the semantic differential for men who practiced safe sex (with one partner or with a condom) and men who practiced unsafe sex (multiple partners and without a condom) to see which sex acts were more gratifying for high-risk-taking and low-risk-taking men (**Further Reading:** semantic differential).

Cantril's Ladder of Life

Figure 11.2 shows Hadley **Cantril's ladder of life** (1965). People are asked to list their concerns in life (financial success, healthy children, freedom from war, and so on). Then they are shown the ladder and are told that the bottom rung, 0, represents the worst-possible life and the top rung, 10, represents the best-possible life. For each of their concerns they are asked to point out where they are on the ladder right now, where they were five years ago, and where they think they'll be five years from now.

Note that the ladder of life is a **self-anchoring scale**. Respondents are asked to explain, in their own terms, what the top and bottom rungs of the ladder mean to them.

Figure 11.2 Cantril's Ladder of Life



Source: H. Cantril. *The Pattern of Human Concerns*. Copyright © 1965 by Rutgers, The State University. Reprinted by permission of Rutgers University Press.

Visual props like this can be used for interviewing nonliterate as well as literate respondents. Keith et al. (1994) used the ladder of life in a study of aging in seven cultures. In five of the sites (two in the United States, one in Hong Kong, and two in Ireland) where most informants were literate, they used a six-rung ladder. (In Hong Kong, where people were comfortable placing themselves *between* but not *on* rungs, the team redesigned the ladder into a flight of stairs.) Among the Herero and Kung of Botswana, where many people were not literate, they replaced the ladder with the five fingers of the interviewer's hand (Keith et al. 1994:xxx, 113).

Hansen and McSpadden (1993) used the ladder-of-life technique in their studies of Zambian and Ethiopian refugees in Zambia and the United States. In Zambia, Hansen actually constructed a small wooden ladder and found that the method worked well. McSpadden used several methods to explore how Ethiopian refugees adjusted to life in the United States. Even when other methods failed, McSpadden found that the ladder-of-life method got people to talk about their experiences, fears, and hopes.

Be careful to tell people exactly what you want when you use any kind of visual prop. M. Jones and Nies (1996) used Cantril's ladder to measure the importance of exercise to elderly African American women. At least Jones and Nies *thought* that's what they were measuring. The mean for the ladder rating was about 9 on a scale of 1–10. Respondents thought they were being asked *how important exercise is*, not how important exercise is *to them, personally*. The researchers failed to explain properly to their respondents what the ladder was supposed to measure, and even devout couch potatoes are going to tell you that exercise is important if you ask them the general question (**Further Reading:** ladder of life).

The Faces Scale

Another interesting device is the **faces scale** shown in Figure 11.3. It's a seven-point (or five-point, or nine-point) scale with stylized faces that change from joy to gloom.

This technique was developed by Kunin in 1955 to measure job satisfaction and has been used widely for this ever since. Andrews and Withey (1976) adapted the faces scale to study well being and the device is now widely used for that. Physicians, nurses, and psychologists use this scale when they ask patients to describe pain. It's particularly good when working with children (Gulur et al. 2009; Wong and Baker 1988), but it's effective with adults as well (A. Harrison 1993) and, like the ladder of life and the semantic differential, has been used in many populations, in one form or another.

You can use the faces scale to capture people's feelings about health care, personal safety—even consumer items (brands of beer, titles of current movies, etc.). People are told: "Here are some faces expressing various feelings. Which face comes closest to how you feel about xxx?" Try using this scale with names of well-known political figures or music artists just to get a feel for how interesting it is.

If you use the faces scale, check for differences in how men and women interpret the neutral face—the one in the middle with the straight-line mouth. Elfering and Grebner (2010) found that 57% of men in their study interpreted the neutral face as sad, compared to 80% for women (**Further Reading:** faces scale).

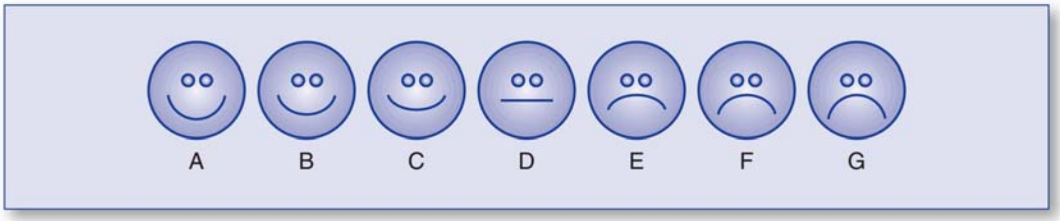
MAGNITUDE SCALING

Most scales in the social sciences are **category scales**. The semantic differential is usually a seven-point scale. Likert-like scales are often five-point scales. These are ordinal measures, but people often have more finely graded opinions than a 1–5 scale captures. For some time, researchers have been experimenting with methods for measuring the actual magnitude of people's impressions, feelings, and attitudes.

These methods, known as **magnitude scaling** are based on the **power law** in psychophysics (Stevens 1957). The power law looks like this:

$$\psi = R = kS^{b} \qquad \textbf{formula 11.4}$$

Figure 11.3 The Faces Scale



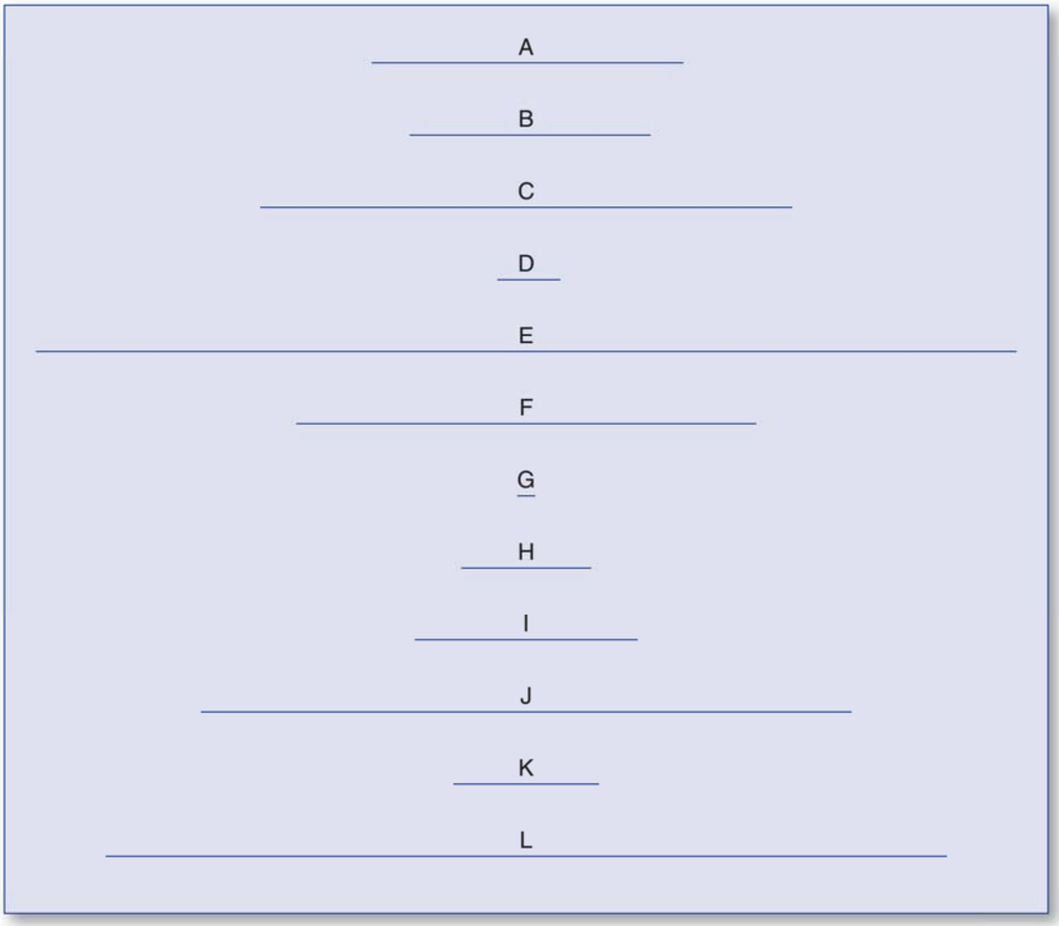
Source: F. M. Andrews and S. B.Withey, *Social Indicators of Well-Being: Americans' Perceptions of Life Quality, Appendix A,* p. 13. Copyright © 1976. Plenum.

where the Greek letter ψ (psi, pronounced "sigh" in English) is perception of the magnitude of some physical stimulus (like a light or a tone); R is what people *say* is the magnitude of the stimulus; and S is the physical stimulus itself (for example, the intensity of the tone in decibels). The exponent b is the power to which you have to raise people's responses to make R and S identical, and k is some constant (Lodge 1981:13).

Suppose you tell people that the brightness value of some light is 50 points and then show them a light that's twice as bright. *If the exponent, b, in formula 11.4 were exactly 1.00*, then, averaging over a lot of people, R would be 100. Over the years, researchers have discovered the exponent—the deviation from 1.00—for various visual, auditory, and other sense stimuli, including line length.

Figure 11.4 is a visual stimulus, from Milton Lodge's book on magnitude scaling (1981:9). If you ask people to rate the lines in Figure 11.4 on a scale of 1–5 (short,

Figure 11.4 Line Lengths as a Visual Stimulus for Direct Magnitude Scaling



Source: M. Lodge, *Magnitude Scaling: Quantitative Measurement of Opinions*, p. 9. Copyright © 1981 by Sage Publications.

medium short, medium, medium long, long), they'll give lines D and G a 1 (short) and they'll give lines E and L (and perhaps C) a 5 (long). Instead, Lodge told 375 people that line A has a length of 50 (not 50 inches or millimeters or miles . . . just 50) and asked them to say how long they thought each of the other lines were. The correlation between the average of their guesses about the length of each line and the actual length of each line was 0.988.

Now, line A in Lodge's test was is actually about 50mm. Hardly anyone can look at line A and tell you it's 50mm, but if you tell people that line A has a value of 50, most of the time they'll tell you that line E has a value of 100. Line E was actually 98mm, or 1.96 times as long as line A, but people round off and get the proportion almost dead on (the *almost* part is why the proportionality exponent is only 0.988 and not 1.00). Line G was 2 mm long, so it was 1/49th the length of line E. Most people give line G a 1 if you tell them that line E is 50. In other words, with a little error, people mostly get the proportions right. (You can test this yourself.)

Magnitude Scaling of Constructs

Does the power law for physical stimuli, like line lengths, translate into better measurement for subjective things, like attitudes? In 1977, the possibility of magnitude scaling of opinions was tested on the National Crime Victimization Survey. The 54,000 respondents saw subsets of 25 from a list of 204 crimes. Here are the instructions to the respondents:

> I would like to ask your opinion about how serious YOU think certain crimes are. The first situation is "A person steals a bicycle parked on the street." This has been given a score of 10 to show its seriousness. Use this first situation to judge all others. For example, if you think a situation is 20 TIMES MORE serious than the bicycle theft, the number you tell me should be around 200, or if you think it is HALF AS SERIOUS, the number you tell me should be around 5, and so on. There is no upper limit. Use ANY number so long as it shows how serious YOU think the situation is. If YOU think something is not a crime, give it a zero.

The respondents then saw a list of 25 crimes . . . things like: (1) A person using force, robs a victim of $10. The victim struggles and is shot to death. (2) A person steals property worth $10,000 from outside a building. (3) A person disturbs the neighborhood with loud, noisy behavior.

And so on. As it turns out, many respondents find magnitude scaling easy to do, and it appears that, for some stimuli, subjective responses do obey some version of the power law. For example, across repeated national studies in the United States, on average, people think a crime of theft is twice as serious as another crime of theft, if the dollar amount stolen in one crime is about 13 times greater than the dollar amount stolen in another crime (Lodge 1981:22).

Magnitude Scaling of Countries' Hostility to the United States

In the 1990s, after the collapse of the Soviet Union, two political scientists, Valerie Sulfaro and Mark Crislip (1997), hypothesized that Americans would have to realign their ideas about who the enemies are out there. Sulfar and Crislip asked 145 undergraduates to rate 19 countries (including one fictitious country, the United Arab Republic) on a seven-point scale. The end points of the scale were labeled "most hostile" and "least hostile" to the United States. Respondents practiced direct magnitude estimation by doing that line-length exercise in Figure 11.4. The correlation between the students' estimation of the line lengths and the actual line lengths was more than 0.99.

With this practice session behind them, the students moved on to estimating, by direct magnitude scaling, the amount of hostility they thought various countries had toward the United States.

Sulfaro and Crislip used France as their reference point for this exercise. They showed respondents a line whose length represented the amount of hostility that France has toward the United States. Respondents then drew lines representing how much hostility they thought the other 18 countries (Britain, Iraq, Panama, etc., etc.) had toward the United States. Each country wound up with two average scores, one for the categorical estimate (on a scale from 1–7) of hostility, and one for the line-drawing exercise. Sulfaro and Crislip converted these average scores into standard scores. The results are shown in Table 11.9.

The entries in Table 11.9 are in standard deviations above and below the mean. Positive standard scores tell you how friendly each country is perceived to be toward the United States, relative to the average for all countries; negative scores tell you how hostile each country is perceived to be, relative to the average for all countries. (I'll show you how to compute standard scores in Chapter 20, when we get to quantitative data analysis.) Canada, Australia, and Britain are more than one standard deviation higher on friendliness (the opposite of hostility). Cuba and Iraq are more than one standard deviation below the mean.

The correlation between these two measures of perceived hostility to the United States is a whopping 0.96, but notice the difference in the score for France on the line lengths and categorical estimates. When France is evaluated categorically, it is not directly compared to any other country. In

Table 11.9 Standardized Hostility/Friendliness Scores for 19 Countries

Country	Line Lengths	Categorical Estimates
Canada	1.24	1.36
Australia	1.24	1.32
Britain	1.15	1.17
Mexico	.78	.65
India	.51	.76
Japan	.23	.35
Saudi Arabia	.52	.21
Israel	.38	.29
Germany	.35	.30
France	.19	.85
Panama	–.03	.03
United Arab Republic	–.29	–.41
PRC (China)	–.32	–.23
Bosnia	–.35	–.84
Russia	–.40	–.29
Nicaragua	–.52	–.72
Serbia	–.42	–.81
Cuba	–1.11	–1.29
Iraq	–3.15	–2.67

Source: V. A. Sulfaro and M. S. Crislip, "How Americans Perceive Foreign Policy Threat: A Magnitude Scaling Analysis." *Political Psychology* 18. Copyright © 1997.

this condition, France gets a very low hostility-toward-the-U.S. score: 0.85 is nearly a full standard deviation above the mean, almost the same as Britain. When France's hostility toward the United States is evaluated *relative to that of other countries*, then France scores far, far below Britain—the same as Japan (Box 11.4).

Box 11.4 Why magnitude scaling is not used more

Magnitude scaling produces some excellent results, but it is complicated to administer (respondents don't always understand what they're supposed to do) and the data are a bit harder to analyze than are categorical data. For one thing, you need to calculate geometric means rather than arithmetic means of the measure of subjective stimuli. This involves taking the natural logarithm of each measure, taking the average of the logs, and then exponentiating the result to get back to where you started. Not exactly straightforward.

With easier-to-use computer programs available for data analysis these days, I expect magnitude scaling to come into its own. It's got a lot of appeal (**Further Reading:** magnitude scaling).

AND FINALLY . . .

There are thousands of published scales. Whatever you're interested in, the chances are good that someone has developed and tested a scale to measure it. Scales are not automatically portable—a scale that measures stress among Vietnamese American women may not measure stress among Hispanic American men—but it makes sense to seek out any published scales on variables you're studying. You may be able to adapt the scales to your needs, or you may get ideas for building and testing an alternative scale.

The Handbook of Research Design and Social Measurement (D. C. Miller and Salkind 2002) always seems hopelessly out of date, yet it remains the best place to start looking for published scales. It's a treasure house full of useful information (**Further Reading:** scales and scaling).

Key Concepts in This Chapter

composite measures
single-indicator scales
multiple indicators
cumulative index
indexes
criterion referenced
norm referenced
robust
unidimensional
Guttman scaling
coefficient of reproducibility
Bogardus Social Distance Scale
sample dependent
Likert scales
Likert-type scales
double-barreled items
item analysis
factor analysis
scoring the items
interitem correlation

- Cronbach's coefficient alpha
- item-total correlation
- split-half reliability test
- the combinations rule
- factor loading
- Michigan Alcoholism Screening Test (MAST)
- Brief MAST
- MAST-G
- Attitudes Toward Women Scale (AWS)
- Interview Schedule for Social Interaction
- the semantic differential
- target item
- paired adjectives
- adjectives of evaluation
- adjectives of potency
- adjectives of activity
- Cantril's ladder of life
- self-anchoring scale
- faces scale
- category scales
- magnitude scaling
- power law

Summary

- A scale is a device for assigning units of analysis to categories of a variable. The assignment is usually done with numbers, and questions are used a lot as scaling devices.
    - A single question on a questionnaire is technically a scale if it lets you assign the people you're studying to categories of a variable. A lot of really interesting variables in social science, however, are complex and can't easily be assessed with single indicators.
- The most common composite measure is a cumulative index. These are made up of several items, all of which count the same.
    - A test in which the only way to get an "A" is to get at least 90% is a criterion-referenced index. A test in which getting an "A" requires that you score in the top 10% of the class is a norm-referenced index.
- A Guttman scale is a unidimensional index. That is, the items are a composite measure of one, and only one underlying concept.
    - The unidimensionality of an index is sample dependent. If the items in a cumulative index form a Guttman scale, then the concept measured by the index is unidimensional, but only for the sample tested.
- Likert scales are the best-known and most widely used scales. A true Likert scale is more than just a format for asking questions. It is a series of items that have been shown to be indicators of an underlying, unidimensional concept.
    - Multidimensional concepts are often measured with complex Likert scales that have several subscales, each of which comprises indicators of a unidimensional concept.
    - Likert scales are tested through a procedure called item analysis. This involves taking the interitem correlation and calculating Cronbach's alpha as a measure of scale reliability. Computers make it easy to use factor analysis to test the unidimensionality of scales.
- With a Likert scale, you ask people a series of questions that get at the target concept. In a semantic differential scale, you name the target concept and ask respondents to rate their feelings toward it on a series of variables.

- Most scales use ordinal categories, but people have more finely graded opinions than, say, a 1–5 scale captures. Magnitude scaling adapts methods from psychophysics to measure attitudes more directly and at a higher level of measurement.
    - Other scales include Cantril's ladder of life and the faces scale.
    - Magnitude scaling is based on the power law in psychophysics.

Exercises

1. This exercise is on the history of scale development. Many scales have changed over the years. It's very instructive to pick an old scale and follow it through its evolution to a modern form. The Wilson-Patterson Conservatism Scale, for example, was developed in the 1960s in New Zealand (G. D. Wilson and Patterson 1968). It was modified for use in the United States by Bahr and Chadwick (1974). Later, Collins and Hayes tested a short version of scale (1993).

   Document the history of a widely used scale. Here are some you might choose from: the Locus of Control Scale (Rotter 1966), the Social Readjustment Scale (Holmes and Rahe 1967), the Authoritarian Personality Scale (Adorno et al. 1950), and the Bem Sex Role Inventory (Bem 1974).

2. Try building a Likert scale to test how serious college students are about their education. This may sound like an easy thing to do, but it isn't. Get together with a small group of students and work on this together. What questions would you ask students if you wanted to scale them on how serious they were about their education? Would the amount of time they claim to spend in the library, or the amount of time they claim to spend partying be useful data? How about what they want to do with their lives after they get their bachelor's degree?

   Once you have a list of questions about attitudes and behaviors, follow the rest of the steps outlined in this chapter, including the item analysis, to get your scale down to a small number of items. You can substitute any value or orientation you like for this exercise. If measuring how serious students are about their education is too close for comfort, then try that shopping orientation variable I mentioned earlier.

3. Get together with a group of other students and decide on a set of target items for a semantic differential test. The items can be types of jobs (forest ranger, family physician, insurance salesperson . . .), names of colors (blue, red, yellow, pink . . .), names of countries (France, Venezuela, Zambia . . .), kinds of music (reggae, jazz, country, classical . . .). Make copies of Figure 11.1 on your word processor and print out a series of semantic differential tests, one for each target item. Choose adjective pairs that make sense for the target items you are studying. For each target item, calculate the mean, across the respondents, of each adjective pair.

Further Reading

Measuring SES. Cirino et al. (2002), Ensminger and Fothergill (2003), Oakes and Rossi (2003).

Bogardus Social Distance Scale. McAllister and Moore (1991), Owen et al. (1981), Parillo and Donoghue (2005).

Guttman scaling. Goodenough (1963), Graves et al. (1969), Liao and Tu (2006), Maitra and Schensul (2002), Wutich and Ragsdale (2008).

Likert scaling. DeVellis (2003).

Semantic differential. Adams-Webber (1997), Arnold-Cathalifaud et al. (2008), Cooker and White (1993), Leunes et al. (1996), Montiel and Boehnke (2000), Ohanian (1990), Turnage (2008).

Ladder of life. Gallicchio et al. (2009), Suhail and Cochrane (1997).

Faces scale. Pasero (1997), Suhail and Chaudhry (2004).

Magnitude scaling. Bard et al. (1996), Goyder (2003), Ogata et al. (2004), Orth and Wegener (1983).

Scales and scaling. Beere (1990), Coombs (1964), Dunn-Rankin (2004), D. C. Miller and Salkind (2002), Netemeyer et al. (2003), Nunnally (1978), Nunnally and Bernstein (1994), Torgerson (1958).